NSBM GREEN UNIVERSITY TOWN

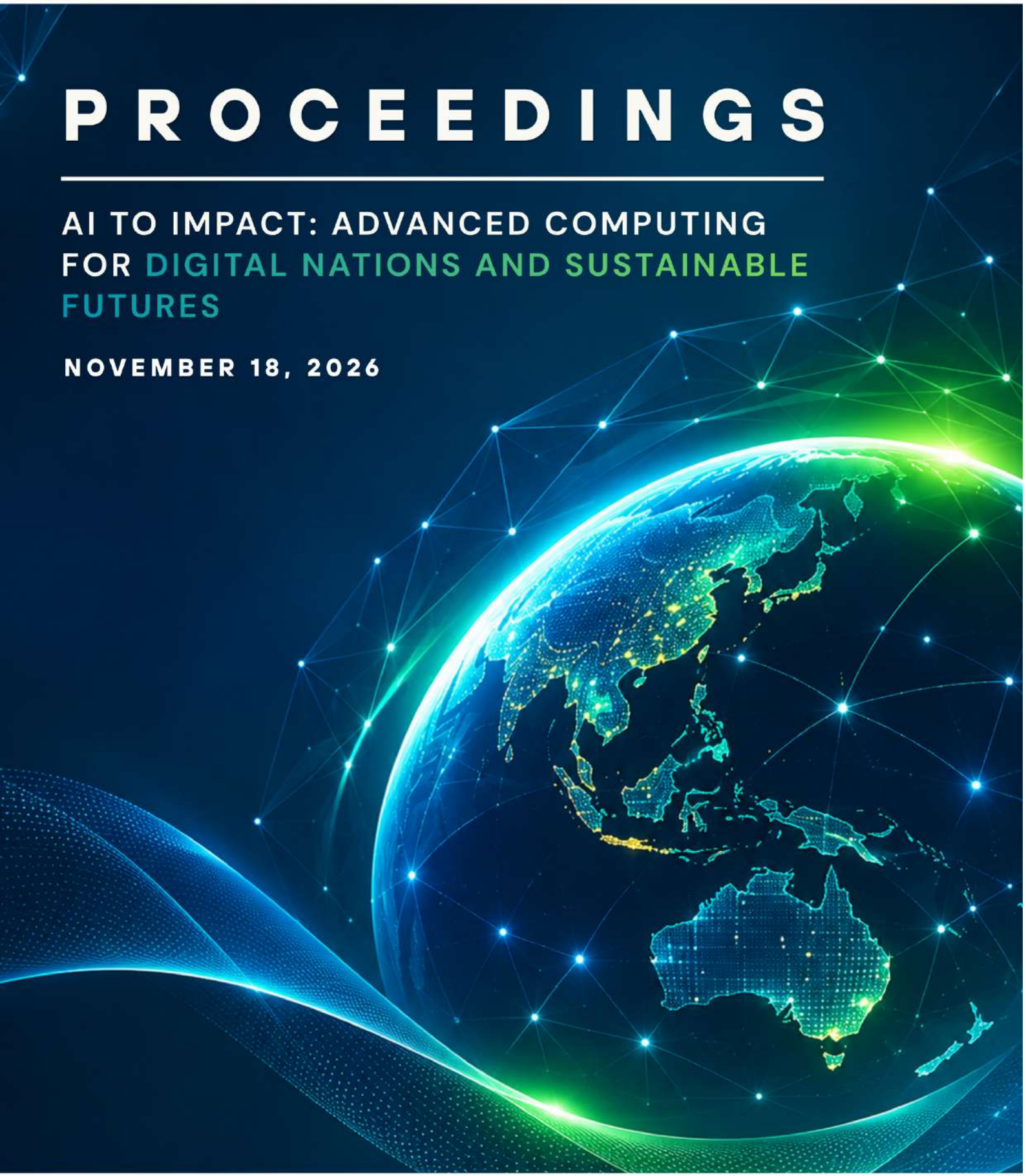
ICACT
INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING TECHNOLOGIES | 2026

RESEARCH CONFERENCE 2026

# PROCEEDINGS

## AI TO IMPACT: ADVANCED COMPUTING FOR DIGITAL NATIONS AND SUSTAINABLE FUTURES

**NOVEMBER 18, 2026**



# INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING TECHNOLOGIES

## FACULTY OF COMPUTING

## NSBM GREEN UNIVERSITY, SRI LANKA

## (ICACT 2026)

# 3rd INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING TECHNOLOGIES

*"AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures"*

## Focusing

**T1 - Intelligent, Secure & Scalable Computing Systems**

This track explores advancements in intelligent computing, artificial intelligence, and secure digital systems. Topics include machine learning, generative AI, cybersecurity, computer vision, and intelligent automation, emphasizing innovative approaches that enhance computational efficiency, system reliability, scalability, and technological innovation across diverse domains.

**T2 - Smart Engineering, IoT & Cyber-Physical Systems**

This track focuses on the integration of intelligent engineering, Internet of Things (IoT), and cyber-physical technologies. Topics include smart sensor networks, intelligent transportation, advanced communication systems, and autonomous technologies, aiming to enhance connectivity, operational efficiency, reliability, and sustainability in modern engineering applications.

**T3 - Digital Health & Computational Life Sciences**

This track explores the application of advanced computational technologies in healthcare and life sciences. Topics include medical imaging, disease prediction, physiological signal analysis, wearable health monitoring, and intelligent clinical decision support, aiming to improve diagnostic accuracy, healthcare accessibility, patient care, and medical innovation.

**T4 - Digital Business, Management & Innovation Systems**

This track examines the role of digital technologies in transforming business and management practices. Topics include digital transformation, business analytics, intelligent decision support, financial technologies, and enterprise systems, emphasizing innovative approaches that enhance organizational efficiency, strategic decision-making, and sustainable business development.

**T5 - Responsible Digital Societies: Governance, Ethics & Emerging Technologies**

This track explores the ethical, social, and governance dimensions of emerging digital technologies. Topics include responsible artificial intelligence, data privacy, digital governance, digital literacy, and sustainable communities, aiming to promote transparency, social inclusion, equitable access, and responsible technological advancement in modern digital societies.

**Organized by**
**Faculty of Computing**
**NSBM Green University**
**Homagama, Sri Lanka**

# DISCLAIMER

The views and opinions expressed in the papers published in these proceedings are solely those of the respective authors. The statements, analyses, and conclusions presented either in the papers or during the conference sessions do not necessarily reflect the official views or policies of NSBM Green University, Sri Lanka.

All papers included in the International Conference on Advanced Computing Technologies (ICACT 2026) proceedings have been formatted according to the IEEE referencing style and reviewed in accordance with the conference's academic and ethical standards.

Authors presenting work that has been previously published or accepted elsewhere are required to inform the Conference Chair in writing before submission. Authors may, however, present only a part of such studies, provided that the content does not duplicate the previously published material in its entirety. Restrictions may apply to the reuse or redistribution of previously published content, subject to publication ethics and copyright regulations.

*ISSN (Print): 3051-5106*
*ISSN (Online): 3093-6519*
*ISBN: 978-624-6035-12-9*

All rights reserved
Copyright ©2026 by NSBM Green University

# CONFERENCE AGENDA

| | |
|---|---|
| 8:30 AM | Registration |
| 9:15 AM | Arrival of Guests & Welcome |
| 9:25 AM | Inaugural and Lighting of the Traditional Oil Lamp |
| 9:30 AM | Welcome Dance |
| 9:35 AM | Welcome Address<br>***Dr. Rasika Ranaweera***<br>***Conference Chair - ICACT 2026*** |
| 9:40 AM | ICACT Event Video |
| 9:45 AM | Keynote Speech<br>***Mr. Shevan Goonetilleke - Chief Executive Officer (CEO), GovTech Sri Lanka*** |
| 10:00 AM | Keynote Speech (Academic)<br>***Professor Kasun De Zoysa - University of Colombo School of Computing (UCSC), Sri Lanka*** |
| 10:20 AM | Message from the Collaborative Partner<br>***Represented by: Dr. Ayman Ibaida***<br>***Victoria University, Australia*** |
| 10:30 AM | Message from the Partner University<br>***Prof. Bogdan Ghita***<br>***Plymouth University, United Kingdom*** |
| 10:35 AM | Launch of Conference Proceedings<br>In collaboration with Taylor and Francis Publishers |
| 10:40 AM | Concluding Remarks<br>***Mr. Saravanapavan Nasiketha***<br>***Conference Co-Chair – ICACT 2026*** |
| 10:45 AM | National Anthem |
| 11:15 AM | Oral Presentations |
| 12:45 PM | Lunch |
| 1:30 PM | Oral Presentations |

# PREFACE

On behalf of the Organizing Committee of the 3rd International Conference on Advanced Computing Technologies (ICACT 2026), it is our great pleasure to extend a warm welcome to all distinguished keynote speakers, authors, reviewers, researchers, industry professionals, and delegates participating in this prestigious event, hosted by the Faculty of Computing, NSBM Green University, Sri Lanka. We are delighted to present the ICACT 2026 Conference Proceedings, which bring together valuable research contributions, innovative ideas, and emerging perspectives from the rapidly evolving field of advanced computing.

Now in its third edition, ICACT continues to strengthen its role as an international platform for promoting research excellence, encouraging interdisciplinary collaboration, and fostering meaningful engagement between academia and industry. Building upon the achievements of the previous conferences, ICACT 2026 reflects NSBM Green University's continued commitment to advancing knowledge, supporting innovative research, and contributing to technological developments that address contemporary societal and industrial challenges.

The papers included in these proceedings underwent a rigorous double-blind review process, ensuring that the contributions reflect the highest standards of scholarship and innovation.

The theme of ICACT 2026, "*AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures*," highlights the transformative potential of artificial intelligence and advanced computing technologies in shaping intelligent, inclusive, and sustainable societies. This theme emphasizes the importance of moving beyond technological innovation towards practical applications that create meaningful benefits for individuals, industries, and communities. Through this conference, we aim to encourage researchers and professionals to explore how emerging technologies can support digital transformation, strengthen responsible innovation, and contribute to sustainable development. We wish all participants a stimulating and rewarding conference experience.

*The Conference Organizing Committee*
*ICACT 2026*

# MESSAGE FROM VICE-CHANCELLOR

PROF. E.A. WEERASINGHE
VICE-CHANCELLOR
NSBM GREEN UNIVERSITY, HOMAGAMA, SRI LANKA

It is with great honor and pleasure that I extend my warmest greetings to all distinguished speakers, researchers, authors, reviewers, industry professionals, and delegates of the 3rd International Conference on Advanced Computing Technologies (ICACT 2026), organized by the Faculty of Computing of NSBM Green University, Sri Lanka. I am delighted to welcome you to this important gathering of minds dedicated to advancing computing research and innovation.

The theme of ICACT 2026, "AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures," is both timely and inspiring. As artificial intelligence and emerging digital technologies continue to transform industries, institutions, and everyday life, we have a shared responsibility to ensure that innovation is secure, ethical, inclusive, and sustainable. This conference offers a valuable platform to explore how research can move beyond technological possibilities to deliver meaningful benefits for society. Now in its third edition, ICACT continues to bring together scholars and practitioners to exchange knowledge, present original research, and build lasting partnerships. The breadth of topics represented from artificial intelligence, machine learning, and data science to cybersecurity, the Internet of Things, and advanced software systems demonstrates the influential role of interdisciplinary collaboration in addressing the challenges of a rapidly evolving digital world.

At NSBM Green University, we remain firmly committed to excellence in education, research, and innovation. ICACT 2026 reflects our ambition to cultivate a dynamic research culture, strengthen connections between academia and industry, and encourage solutions that contribute to national development and global progress.

I extend my sincere appreciation to the organizing committee, the Faculty of Computing, and all those who have worked diligently to make this conference possible. I also thank the authors, reviewers, keynote speakers, and delegates for their invaluable contributions. May the ideas shared through these proceedings inspire further discovery, meaningful collaboration, and responsible technological advancement for a better and more sustainable future.

# MESSAGE FROM SENIOR DEPUTY VICE CHANCELLOR

PROF. CHAMINDA RATHNAYAKE
SENIOR DEPUTY VICE-CHANCELLOR
NSBM GREEN UNIVERSITY, HOMAGAMA, SRI LANKA

I take great pride in extending heartfelt greetings to all distinguished speakers, researchers, authors, reviewers, and delegates joining the 3rd International Conference on Advanced Computing Technologies (ICACT 2026), proudly organized by NSBM Green University's Faculty of Computing. Building on the success of its previous editions, ICACT continues to serve as a dynamic hub for intellectual dialogue, global collaboration, and transformative computing research.

Centered on the theme "AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures," ICACT 2026 highlights a fundamental truth: innovation only matters when it creates tangible value for humanity. As artificial intelligence and next-generation technologies reshape our world, we must ensure their adoption remains ethical, secure, inclusive, and socially responsible.

NSBM Green University takes great pride in driving research that translates academic excellence into practical, industry-ready solutions. ICACT embodies this mission by bringing together scholars, practitioners, and emerging thinkers to collaborate and address the pressing challenges of our digital future.

I extend my sincere appreciation to the organizing committee and the Faculty of Computing for their dedication in making this conference possible. I also thank the authors, reviewers, keynote speakers, and participants for enriching ICACT 2026 through their knowledge, commitment, and scholarly contributions.

May this conference inspire meaningful dialogue, lasting collaborations, and impactful research that contributes to a more intelligent, secure, and sustainable digital future.

# MESSAGE FROM DEPUTY VICE CHANCELLOR - ACADEMIC

PROF. THILINI DE SILVA
DEPUTY VICE-CHANCELLOR – ACADEMIC
NSBM GREEN UNIVERSITY, HOMAGAMA, SRI LANKA

It is a distinct privilege to welcome our esteemed speakers, researchers, authors, reviewers, and delegates to the 3rd International Conference on Advanced Computing Technologies (ICACT 2026), organized by the Faculty of Computing at NSBM Green University. Now in its third year, ICACT has established itself as a premier forum for knowledge exchange, collaborative networking, and pioneering computing research.

This year's theme, "AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures," underscores that genuine technological advancement is measured by how meaningfully it uplifts society. While artificial intelligence and emerging digital technologies present unprecedented possibilities, realizing their full potential requires deploying them responsibly, securely, ethically, and inclusively.

At NSBM Green University, we take pride in nurturing a research culture that connects academic excellence with practical innovation. ICACT brings together academics, industry professionals, and emerging researchers to exchange insights, explore new ideas, and develop solutions to the challenges of an increasingly digital world.

I deeply appreciate the tireless efforts of the organizing committee and the Faculty of Computing in making this gathering a reality. I also wish to thank our authors, reviewers, keynote speakers, and attendees for enriching ICACT 2026 with their invaluable insights and scholarly contributions. I am confident that this conference will spark meaningful conversations, build lasting partnerships, and produce research that paves the way for a smarter, more secure, and sustainable digital world.

# MESSAGE FROM THE CONFERENCE CHAIR

DR. RASIKA RANAWEERA
CONFERENCE CHAIR – ICACT 2026
NSBM GREEN UNIVERSITY, HOMAGAMA, SRI LANKA

It is my great pleasure to welcome all participants, researchers, and industry professionals to the International Conference on Advanced Computing Technologies (ICACT 2026), organized by the Faculty of Computing of NSBM Green University. As we convene for the third edition of this esteemed conference, we continue our commitment to advancing computing research that drives intelligent, secure, and transformative innovation. We are excited to showcase a diverse array of articles submitted by both local and international researchers. Following a rigorous double-blind review process, we look forward to presenting the best works, which will be published in Taylor and Francis. This process highlights the hard work of our national and international reviewers, ensuring the high standard of the conference proceedings.

The theme of ICACT 2026, *"AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures,"* symbolizes a new frontier in technology one where artificial intelligence and digital systems are reshaping every facet of human life. In this rapidly evolving era, the fusion of innovation and security is essential to ensure that technological progress remains ethical, inclusive, and beneficial to society.

ICACT 2026 serves as a dynamic platform for academics, practitioners, and innovators to share knowledge, explore pioneering research, and discuss emerging technologies that define the future of computing. From intelligent systems and data-driven decision-making to cybersecurity and human-centered AI, the contributions presented here reflect the vast potential of modern computing to redefine life itself.

At NSBM Green University, we take pride in fostering a vibrant research culture grounded in excellence, collaboration, and real-world impact. This conference stands as a testament to our vision of connecting academia and industry to build a smarter and more sustainable digital world. I extend my heartfelt appreciation to the organizing committee, speakers, and participants for their invaluable efforts and contributions. May ICACT 2026 ignite meaningful discussions, inspire secure innovations, and strengthen collaborations that shape the intelligent technologies of tomorrow.

# MESSAGE FROM THE EDITOR-IN-CHIEF

MR. J. GAYAN PERERA
EDITOR-IN-CHIEF – ICACT 2026
NSBM GREEN UNIVERSITY, HOMAGAMA, SRI LANKA

It is a great privilege and pleasure to present the proceedings of the International Conference on Advanced Computing Technologies (ICACT 2026), organized by the Faculty of Computing, NSBM Green University, Sri Lanka. Under the theme "AI to Impact: Advanced Computing for Digital Nations and Sustainable Futures," the conference brings together researchers, academics, and industry professionals to share their research findings and explore how advancements in artificial intelligence and computing technologies can address real-world challenges, drive digital transformation, and contribute to sustainable development.

ICACT 2026 received 160 research paper submissions across five conference tracks, reflecting a diverse range of research interests in advanced computing technologies. All submissions underwent a rigorous double-blind peer-review process, conducted exclusively by 172 external reviewers from Sri Lanka and abroad. This process ensured a fair and independent evaluation of submissions, contributing to the selection of high-quality research papers for the conference proceedings.

As Editor-in-Chief, I extend my sincere gratitude to the Vice-Chancellor, Senior Deputy Vice-Chancellor, Deputy Vice-Chancellor (Academic), Head of the Research Council, and Dean of the Faculty of Computing for their invaluable guidance and support. My appreciation also extends to the Conference Chairs, Conference Secretary, Editorial Committee, and Organizing Committee for their dedication and collective efforts, as well as to our national and international reviewers for their commitment to upholding the academic standards of ICACT 2026.

I hope these proceedings will inspire future research, foster meaningful collaborations, and encourage the transformation of innovative ideas into practical solutions that advance digital nations and sustainable futures.

# ORGANIZING COMMITTEE

## ADVISORY BOARD

Prof. E.A. Weerasinghe – Vice Chancellor
Prof. Chaminda Rathnayake – Senior Deputy Vice-Chancellor
Prof. Thilini De Silva – Deputy Vice-Chancellor Academic
Assoc. Prof. Chaminda Wijesinghe - Head, Department of Computer & Data Science
Dr. Udara Arachchige - Head, Research Council
Dr. Ravinder Singh - Senior Lecturer, Victoria University, Australia
Dr. Bo Li - Senior Lecturer, Victoria University, Australia

## CONFERENCE COMMITTEE

Dr. Rasika Ranaweera – Conference Chair
Mr. Saravanapavan Nasiketha – Conference Co-Chair
Mr. Chaminda Attanayake - Conference Co-Chair
Dr. Ayman Ibaida - Conference Co-Chair
Ms. Yasanthika Mathotaarachchi - Conference Secretary
Mr. Gayan Perera – Conference Editor-in-Chief
Ms. Dharani Rajasinghe – Conference Associate Editor
Ms. Pavithra Subhashini
Mr. Madusanka Mithrananda
Mr. Isuru Sri Bandara
Ms. Hirushi Dilpriya
Ms. Dulanjali Wijesekara
Ms. Thilini Bakmeedeniya
Ms. Hiruni Weerasinghe
Ms. Kavishka Rajapaksha
Mr. Krishna Kishore
Ms. Nipuni Vithana
Ms. Chathurika Ranaweera
Ms. Sandunika Upavindi
Mr. Janod Bandara
Ms. Kaveesha Wijesinghe
Ms. Chamathka Dahami

## EDITORIAL COMMITTEE

Ms. Nimesha Hewawasam
Mr. Anton Jayakody
Mr. Diluka Wijesinghe
Mr. Krishantha Ranaweera
Ms. Thisarani Wickaramasinghe

## TECHNICAL PROGRAMME COMMITTEE

Dr. Chandana Perera - *TPC Chair*

Dr. Pabudi Abeyrathne - *TPC Co-Chair*

Prof. Nathan Clarke – *Professor in Cyber Security and Digital Forensics, School of Engineering, Computing and Mathematics, Faculty of Science and Engineering*

Prof. S. Vasanthapriyan – *Professor, Faculty of Computing, Sabaragamuwa University*

Prof. Prasad M. Jayaweera – *Dean / Professor of Information Systems Engineering and Informatics, Faculty of Computing, University of Sri Jayewardenepura*

Prof. P. Ravindra S. De Silva – *Professor in Computer Science, University of Sri Jayewardenepura, Department of Computer Science*

Prof. (Mrs.) A.T.P. Silva – *Professor in AI, University of Moratuwa*

Prof. Jayalath Ekanayake – *Professor, Uva Wellassa University, Sri Lanka*

Dr. Hai-Van Dang – *Lecturer in Cyber Security, School of Engineering, Computing and Mathematics, Faculty of Science and Engineering*

Dr. Taimur Bakhshi – *Plymouth University, United Kingdom*

Dr. Pulasthi Gunawardhana – *Senior Lecturer (Grade II), University of Sri Jayewardenepura*

Dr. Udaya Wijenayake – *Senior Lecturer, Department of Computer Engineering, University of Sri Jayewardenepura*

Dr. S.C. Premaratne – *Senior Lecturer, University of Moratuwa, Sri Lanka*

Dr. (Ms.) I. Thilini S. Piyatilake – *Head, Department of Computational Mathematics, University of Moratuwa*

Dr. (Ms.) Suresha Nilmini Perera – *Lecturer, The Open University of Sri Lanka*

Dr. Dilrukshi Gamage – *Senior Lecturer, University of Moratuwa, Sri Lanka*

Dr. Ruwan Wickramarachchi – *Senior Lecturer, University of Kelaniya, Sri Lanka*

Dr. Alexander Rukshan – *Senior Lecturer, University of Vavuniya, Sri Lanka*

Dr. MM Mohamed Mufassirin – *Senior Lecturer, South Eastern University of Sri Lanka*

Dr. (Mrs.) Nirosha Wedasinghe – *Senior Lecturer, KDU, Sri Lanka*

Dr. (Mrs.) T.A. Weerasinghe – *Senior Lecturer, University of Colombo School of Computing, Sri Lanka*

Dr. (Ms.) KHELW Hettiarachchi – *Senior Lecturer, University of Colombo School of Computing, Sri Lanka*

Dr. KAKT Karunanayaka – *University of Colombo School of Computing, Sri Lanka*

Mr. V. Senthooran – *Senior Lecturer, Vavuniya Campus, Sri Lanka*

## DESIGN TEAM

Ms. Semini Siriwardana

Ms. Savindri Weerakoon

Mr. Savindu Dhamsara

Mr. Supun Gajendrasinghe

# REVIEW BOARD

## Meta-Reviewers

**Prof. Jayantha Lanel,** *University of Sri Jayewardenepura, Sri Lanka*
**Dr. E. Hettiarachchi,** *University of Colombo School of Computing, Sri Lanka*
**Dr. Pulasthi Gunawardhana,** *University of Sri Jayewardenepura, Sri Lanka*
**Prof. Ruwan Wickramarachchi,** *University of Kelaniya, Sri Lanka*
**Dr. Saminda Premaratne,** *University of Moratuwa, Sri Lanka*
**Dr. T. A. Weerasinghe,** *University of Colombo School of Computing, Sri Lanka*
**Dr. Mohamed Mufassirin Mohamed Muzammil,** *South Eastern University of Sri Lanka, Sri Lanka*
**Mr. Dhammika Marasinghe,** *99x Technology AS, Sri Lanka*

## Reviewers

**Prof. T. S. L. W. Gunawardana,** *University of Ruhuna, Sri Lanka*
**Prof. James Chung-Wai Cheung**, *The Hong Kong Polytechnic University, Hong Kong SAR*
**Prof. Kokila Konasinghe**, *University of Colombo, Sri Lanka*
**Prof. Kushanthi Harasgama**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Prof. Nilushi Nugara**, *University of Sri Jayewardenepura, Sri Lanka*
**Prof. P. I. N. Fernando,** Uva Wellassa University of Sri Lanka, Sri Lanka
**Prof. R. G. N. Meegama**, *University of Sri Jayewardenepura, Sri Lanka*
**Prof. Salinda Premadasa**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Prof. Samantha Rajapaksha**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Prof. Sanath Wijesinghe**, *The Open University of Sri Lanka, Sri Lanka*
**Prof. S. Kasthuriarachchi**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Prof. Theekshana Suraweera**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Prof. Wasantha Rajapakshe**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Prof. W. Abeygunawardhana,** *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Dr. Anusha Jayasiri**, *University of Kelaniya, Sri Lanka*
**Ts. Dr. Azlin Abd Jamil,** *Universiti Teknologi Malaysia, Malaysia*
**Dr. Barathy Mayurathan**, *University of Jaffna, Sri Lanka*
**Dr. Bhagya Senadheera**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. Chamara Liyanage**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. D. S. Deegalla,** *University of Peradeniya, Sri Lanka*
**Dr. Deepesh Kumar**, *Indian Institute of Technology (BHU) Varanasi, India*
**Dr. Deepika Seneviratne**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. Dulharie Wijeratne**, *The Open University of Sri Lanka, Sri Lanka*
**Dr. Gajanayake Mudalige Kumara**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. Geetha P**, *Anna University, India*
**Dr. Geetha V**, *National Institute of Technology Karnataka, Surathkal, India*
**Eng. Dr. Gehan Anthonys**, *The Open University of Sri Lanka, Sri Lanka*
**Dr. Guruge Dilrukshi**, *The Open University of Sri Lanka, Sri Lanka*
**Dr. Gyanendra Verma**, *National Institute of Technology Raipur, India*
**Dr. Ishara Lakmali**, *The Open University of Sri Lanka, Sri Lanka*
**Dr. Jayalath Ekanayake**, *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Dr. Jayani Hapugoda**, *The Open University of Sri Lanka, Sri Lanka*
**Dr. Kasun Hemachandra**, *University of Moratuwa, Sri Lanka*
**Dr. Kasun Piyumal**, *University of Kelaniya, Sri Lanka*
**Dr. Kaushalya Dissanayake**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Dr. Kokul Thanikasalam**, *University of Jaffna, Sri Lanka*
**Dr. Laalitha Liyanage**, *University of Kelaniya, Sri Lanka*
**Dr. Lakmini Malasinghe**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*

**Dr. Liew Shiung**, *Universiti Malaya, Malaysia*
**Dr. Lojenaa Navanesan**, *University of Vavuniya, Sri Lanka*
**Dr. Madushani Upamalika**, *University of Ruhuna, Sri Lanka*
**Dr. Mansoor Cassim**, *South Eastern University of Sri Lanka, Sri Lanka*
**Dr. Mohamad Haider Abu Yazid**, *Universiti Teknologi Malaysia, Malaysia*
**Dr. Mohammad Khan**, *Multimedia University, Malaysia*
**Dr. Nadia Shaira Shafii**, *Universiti Teknologi Malaysia, Malaysia*
**Dr. Namalee Napagoda**, *Wayamba University of Sri Lanka, Sri Lanka*
**Dr. Nazmona Mat Ali**, *Universiti Teknologi Malaysia, Malaysia*
**Dr. Pavithra Basnayake**, *University of Peradeniya, Sri Lanka*
**Dr. Piumi Ishanka**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Dr. Pradeep Kalansooriya**, *General Sir John Kotelawala Defence University, Sri Lanka*
**Dr. Praneetha Sree R**, *National Institute of Technology Raipur, India*
**Dr. Pubudu Jayasena**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Dr. RM Maheshi Ruwanthika**, *University of Moratuwa, Sri Lanka*
**Dr. Rukshani Liyanaarachchi**, *University of Moratuwa, Sri Lanka*
**Dr. Ruvini Abhayapala**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. S. Wickramasinghe,** *Institute of Chartered Accountants of Sri Lanka, Sri Lanka*
**Dr. Shajini Majuran**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. Shamila Wickramasuriya**, *University of Peradeniya, Sri Lanka*
**Dr. Suneth Karunarathna**, *University of Peradeniya, Sri Lanka*
**Dr. Suresha Perera**, *The Open University of Sri Lanka, Sri Lanka*
**Dr. Swarnalatha Radhakrishnan**, *University of Peradeniya, Sri Lanka*
**Dr. Udaya Wijenayake**, *University of Sri Jayewardenepura, Sri Lanka*
**Dr. Zati Hakim Binti Azizul Hasan**, *Universiti Malaya, Malaysia*
**Mr. Akalanka Thilakarathna**, *University of Colombo, Sri Lanka*
**Mr. Ameer Mohamed Aslam Sujah**, *South Eastern University of Sri Lanka, Sri Lanka*
**Ms. Amrithaa Thuvakaran**, *University of Vavuniya, Sri Lanka*
**Ms. Anupama Perera**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Mr. Anuradha Herath**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Ms. Aruni Premarathne**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Ashan Ratnayake,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Mrs. B. B. D. S. Abeykoon,** *University of Ruhuna, Sri Lanka*
**Ms. Binuri Raigamkorale**, *University of Ruhuna, Sri Lanka*
**Mr. Biswajith Dissanayake**, *University of Peradeniya, Sri Lanka*
**Ms. Chamali Gamage**, *University of Sri Jayewardenepura, Sri Lanka*
**Mr. Chamila Karunathilake**, *University of Sri Jayewardenepura, Sri Lanka*
**Mr. Daminda Herath**, *General Sir John Kotelawala Defence University, Sri Lanka*
**Mr. Deepal Samarakoon,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Ms. Dilanka de Silva**, *SLTC Research University, Sri Lanka*
**Mr. Dimuth Pathirana,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Ms. Dinesha Samararatne**, *University of Colombo, Sri Lanka*
**Ms. Dinithi Pandithage**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Dulitha Hewadikaram**, *General Sir John Kotelawala Defense University, Sri Lanka*
**Ms. Dushanthi Kuruppu**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Eng. Gaston Ravin Dias,** Healthcare Engineering (Pvt) Ltd, Sri Lanka
**Mr. Gihan Seneviratne**, *University of Colombo School of Computing, Sri Lanka*
**Ms. Hansamali Paul**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. Hansika Mahadikara**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Hansika Mannapperuma**, *University of Ruhuna, Sri Lanka*
**Ms. Harshani Perera**, *CINEC Campus, Sri Lanka*
**Ms. Himashi De Mel**, *Asia Pacific Institute of Information Technology (APIIT), Sri Lanka*
**Ms. Ishani Fonseka**, *University of Colombo School of Computing, Sri Lanka*

**Mr. Jagath Kulathilake,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Mr. Janotheepan Mariyathas,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Mrs. Jayamuthu Edirisinghe**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. K.B.A.B. Chathurika**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Kanishka Yapa**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Kasun Wikramarathna**, *SLTC Research University, Sri Lanka*
**Ms. Kaveesha Silva**, *University of Sri Jayewardenepura, Sri Lanka*
**Mr. Khedika Krishnaraj**, *Trincomalee Campus, Eastern University, Sri Lanka*
**Ms. Lavanka Harshani**, *University of Moratuwa, Sri Lanka*
**Ms. Lokesha Weerasinghe**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mrs. M.S Faathima Fayaza**, *South Eastern University of Sri Lanka, Sri Lanka*
**Ms. Maheshika Dissanayake**, *University of Kelaniya, Sri Lanka*
**Ms. Malithi Nawarathne**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Manori Gamage**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Maryam Khalid**, *Universiti Malaya, Malaysia*
**Mr. Muhammad Rifthy Kalideen**, *South Eastern University of Sri Lanka, Sri Lanka*
**Mr. N.H.P. Swarnakantha**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Natasha Wijeyesekera**, *Asia Pacific Institute of Information Technology (APIIT), Sri Lanka*
**Mr. Chathuranga Amarasena**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Nethmini Weerawarna**, *University of Colombo, Sri Lanka*
**Mr. Nimantha Madhushan**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. Nirasha Kulasooriya**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. Nirubikaa Ravikumar**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Mr. Nusky Ahamed**, *South Eastern University of Sri Lanka, Sri Lanka*
**Ms. Pavithra Madushanka**, *General Sir John Kotelawala Defense University, Sri Lanka*
**Mrs. Piumika Karunanayake**, *General Sir John Kotelawala Defense University, Sri Lanka*
**Mr. Piyumal Ranawaka**, *University of Moratuwa, Sri Lanka*
**Mr. PPG Dinesh Asanka**, *University of Kelaniya, Sri Lanka*
**Ms. Priyanga Munidasa**, *The Open University of Sri Lanka, Sri Lanka*
**Ms. Puvanendran Rukshani**, *University of Vavuniya, Sri Lanka*
**Ms. Rangi Dahanayake**, *Asia Pacific Institute of Information Technology (APIIT), Sri Lanka*
**Ms. Rangi Liyanage**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Ratnasingam Sakuntharaj**, *University of Vavuniya, Sri Lanka*
**Ms. S.D.H.S. Wickramarathne,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Mr. Sajeewa Kumarasingha**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Samadhi Rathnayake**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Sandya Hewameealla**, *The Open University of Sri Lanka, Sri Lanka*
**Ms. Sanjeevi Chandrasiri**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Sankani Heenkenda**, *University of Sri Jayewardenepura, Sri Lanka*
**Mrs. Saranga Somaweera**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Ms. Selvaraj Puwanitha**, *University of Colombo, Sri Lanka*
**Mr. Shantha Puhulwella**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. Shashika Lokuliyana**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Sittampalam Sotheeswaran**, *Eastern University, Sri Lanka*
**Mr. Sivanenthira Sivapragasam**, *University of Vavuniya, Sri Lanka*
**Ms. Suriyaa Kumari**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Ms. Sweshthika Sureshkumar**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Mr. T Arudchelvam**, *Wayamba University of Sri Lanka, Sri Lanka*
**Mr. Thajudeen Ahamed**, *University of Vavuniya, Sri Lanka*
**Ms. Thakshila Samarakkody**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mrs. Thanuja Dananjali**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Ms. Thanushika Jathunga**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mrs. Thanushya Thanujan**, *Eastern University, Sri Lanka*

**Ms. Tharindi Kamalasena**, *University of Ruhuna, Sri Lanka*
**Ms. Tharmini Kiriharan**, *Eastern University, Sri Lanka*
**Mr. Thilina Gunarathne**, *University of Peradeniya, Sri Lanka*
**Mr. Thillainayagam Dushyenthan**, *University of Vavuniya, Sri Lanka*
**Ms. Thisara Shyamalee**, *Sri Lanka Institute of Information Technology (SLIIT), Sri Lanka*
**Mr. Umesh Ranasooriya,** *Uva Wellassa University of Sri Lanka, Sri Lanka*
**Ms. Upeka Premaratne**, *University of Moratuwa, Sri Lanka*
**Ms. Upeksha Hansani**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. Upeksha Kudagamage**, *Sabaragamuwa University of Sri Lanka, Sri Lanka*
**Ms. Upeksha Samarasinghe**, *University of Sri Jayewardenepura, Sri Lanka*
**Ms. Varaniya Varatharasa**, *University of Vavuniya, Sri Lanka*
**Mr. Veerapathirapillai Vinoharan**, *University of Vavuniya, Sri Lanka*
**Mr. Vijayanathan Senthooran**, *University of Vavuniya, Sri Lanka*
**Ms. Vithuckshiha Puvanasingam**, *University of Vavuniya, Sri Lanka*
**Ms. Vithusha Balachandiran**, *University of Vavuniya, Sri Lanka*
**Mr. Vithusia Prashanth**, *Trincomalee Campus, Eastern University, Sri Lanka*
**Ms. Waruni Rajapaksha**, *Asia Pacific Institute of Information Technology (APIIT), Sri Lanka*
**Mrs. Wasana Panditharatne**, *The Open University of Sri Lanka, Sri Lanka*
**Ms. Wasana Uduwela**, *The Open University of Sri Lanka, Sri Lanka*
**Mrs. Wasantha Seneviratne,** University of Colombo, Sri Lanka

# MESSAGE FROM THE KEYNOTE SPEAKER (ACADEMIC)

**Professor Kasun De Zoysa**
University of Colombo School of Computing (UCSC)
Sri Lanka

## Building Trust in the Age of Intelligent Systems

The rapid advancement of computing technologies is transforming how we live, work, and interact with the digital world. Artificial Intelligence, intelligent systems, and emerging digital innovations present unprecedented opportunities for progress. However, these advancements also introduce new challenges concerning security, trust, and responsible technological development. As intelligent systems become increasingly integrated into everyday activities, cybersecurity must extend beyond protecting technological infrastructure to understanding the interactions between humans and AI. Social engineering, once primarily directed at individuals, is evolving to exploit intelligent agents capable of accessing information, making decisions, and performing actions on our behalf. Addressing these emerging threats requires not only technological innovation but also greater awareness, ethical responsibility, and interdisciplinary collaboration.

The *3rd International Conference on Advanced Computing Technologies (ICACT 2026)* provides an important platform for researchers, academics, and industry professionals to exchange knowledge, challenge existing perspectives, and explore solutions to contemporary technological problems. Such academic gatherings remind us that meaningful research must extend beyond publications and contribute to practical solutions that benefit society.

I commend NSBM Green University and the organizing committee for bringing together diverse research perspectives through ICACT 2026. May this conference inspire meaningful discussions, strengthen research collaborations, and encourage innovations that contribute to a more secure, trustworthy, and sustainable digital future.

# MESSAGE FROM THE KEYNOTE SPEAKER

**Mr. Shevan Goonetilleke**
Chief Executive Officer (CEO)
GovTech Sri Lanka

## Sri Lanka's Journey to a Digital Nation: From Blueprint to Reality

What does it take to become a truly digital nation? Keynote outlines the foundations: a trusted digital identity, clean national registries, connected government systems, and secure, privacy-respecting data sharing. Together they make possible paperless, cashless and presenceless services where citizens provide information once and never queue again. It examines where Sri Lanka stands today against its national digital blueprint, the policy, legal and data foundations that must come first, and where AI can accelerate the journey. It closes with what citizens can expect in the coming years, the role academia and researchers must play, and why Sri Lanka's momentum towards a digital nation has never been stronger.

A truly digital nation rests on trusted identity, connected systems and services citizens can use without a single queue. This keynote examines where Sri Lanka stands on that journey, what must come first, how AI and research can accelerate it, and why the momentum has never been stronger.

The *3rd International Conference on Advanced Computing Technologies (ICACT 2026)* represents an opportunity to strengthen this connection. I hope the research presented and conversations initiated here will challenge conventional thinking, inspire new partnerships, and encourage solutions that respond to the evolving needs of our society.

Ultimately, our success should not be measured by how advanced our technologies become, but by how effectively they improve lives, create opportunities, and shape a better future for SriLanka.

# MESSAGE FROM THE PARTNER UNIVERSITY

**Professor Bogdan Ghita**
Professor of Network Security, Performance and Trust

Associate Pro Vice Chancellor (Transnational Education)

University of Plymouth, United Kingdom

## From AI to Impact: Building Digital Systems We Can Trust

The theme of ICACT 2026, AI to Impact, focuses on the move from artificial intelligence conceptual novelty to real-world tangible benefits; this, in practice, is the hardest part of the journey: algorithms, models and architectures are now developing faster than our ability to show that they are dependable, affordable and useful to the people they are meant to serve. The challenge presented to current researchers is how to close the gap between scientific advances of AI and their adoption by society. As part of this journey, we are presented with three questions that mirror the human acceptance process, relating to beneficiaries, trustworthiness, and context. The success of AI and advanced computing depends on three fundamental questions: who benefits, whether the technology can be trusted, and whether its outcomes are sustainable and inclusive. Improvements in accuracy and efficiency are meaningful only when they address real-world challenges and deliver measurable benefits across areas such as healthcare, business, and public services. As intelligent systems become increasingly adaptive and autonomous, security, resilience, transparency, and continuous evaluation must be embedded into their design to ensure reliability and trustworthiness. Furthermore, the long-term implications of AI on energy consumption, infrastructure, workforce development, and accessibility must be carefully considered to ensure that technological advancements benefit diverse communities.

These questions extend across all five tracks of the conference, highlighting the importance of interdisciplinary research and collaboration. The research presented throughout the event offers innovative perspectives and solutions to these emerging challenges, creating opportunities for meaningful discussion and knowledge exchange. Conference participants are encouraged to explore how advanced computing can move beyond technical excellence to deliver responsible and lasting societal impact.

# Contents

1. Explainable Machine Learning for Academic Procrastination Prediction: A Comparative Evaluation of Conventional and Retrieval-Based Models ..... 1
*Daminda Herath, Thanuja Ambegoda, Kaneeka Vidanage*

2. FedQHO-Med: Energy-Aware Quantum-Inspired Spectrum Allocation with Federated Gradient-Aligned Learning for 6G Cognitive Radio Medical-IoT Networks ..... 7
*W. D. J. I. Senarathna, T. A. H. Dilpriya*

3. An LLM-Based Multi-Agent Framework for Adaptive Digital Marketing Campaign Design and Optimization ..... 14
*I.K.D. Vilochana, T. H. Wickramage, M.T.A. Wickramasinghe*

4. Enhanced YOLOv10 Framework for Real-Time Road Detection in Foggy and Complex Driving Environments ..... 21
*Sachin Chandanshive, Niranjan Kulkarni*

5. Coordinating Verification and Generation in Multi-Agent Systems: A Real-Time, Evidence- Grounded Lecture Note Assistant ..... 27
*Mihidukulasuriya Nethmi Ruasha Fernando, Thisarani Wickramasinghe*

6. AI Student Guide Hub: An LLM-Orchestrated Context-Aware Framework for Personalized Career Recommendation and Learning Path Generation ..... 34
*Banoth Shiva Kumar, Talla Prashanthi, Budidha Adarsh Goud, Ganesh B Regulwar, Neelakantam Shiva, Babitha Lincy R*

7. A Hybrid Machine Learning Framework for Hydropower Generation Forecasting: A Case Study of the Samanalawewa Power Station, Sri Lanka ..... 42
*N. A. D. Zoysa, T. A. H. Dilpriya*

8. Assessing Associations Between Wayground AI Features and Intrinsic Motivation among Middle School Students in Online Education: A Quantitative Survey Study ..... 48
*V. L. Karunarathne, D. G. S. N. Samarakoon*

9. AI-Powered Ceiling Material Estimation from Plan Area Images ..... 53
*Kumuditha Ranasinghe, G. Perera*

10. Security-by-Design for AI-Enabled Digital Twins in Critical Infrastructure: An Integrated Frame- work for Cyber, AI- Model and Twin-Synchronization Risks ..... 58
*Senesh Fitzroy, Pawani Bandara*

11. A Unified On-Device Classifier for NSFW and Malicious Web Pages Using Screenshot ..... 65
*Adhishtanaka Kulasooriya, R. Ranaweera*

12. MBR Integrity Protection and Automated Recovery System Against Boot-Sector Ransomware ..... 71
*Sanoj Aminda, Chamara Disanayake, Kaveesha Gimhani, Madusanka Mithrananda*

13. AI-Assisted Agent-Based Repayment Guidance for Credit Card Interest Minimization: A Design Study with Simulation- Based Evaluation
for the Sri Lankan Market ..... 77
*Sadeesha V. Jayaweera*

14. MAGHR-RAG: Meta-Adaptive Graph-Hybrid Reflective Retrieval-Augmented Generation for Hallucination Mitigation in LLM Chat Assistants ..... 83
*L. A. D. Kavindu Shehan Deshappriya, Thilini Bakmeedeniya*

15. AI-Based Restoration and Digital Preservation of Sri Lankan Ola Leaf Manuscripts Using Deep Learning and Computer Vision ..... 91
*T. H. Wickramage, S. H. G. G. Sewjith, I. K. D. Vilochana*

16. An Explainable Hybrid CNN-BiLSTM Framework with Mish Activation and Random Over-Sampling for Multi-Class ECG Arrhythmia Classification ..... 98
*Kavindu Piumal, Hiruni Weerasinghe*

17. Gated but Not Explained: A Faithfulness Audit of Intrinsic Explanations in Multi-Agent Reinforcement Learning ..... 104
*S. H. G. G. Sewjith, T. H. Wickramage, I. K. D. Vilochana*

18. A Secure AI-Assisted Publishing Architecture for Institutional Websites that Require Controlled Content Governance and Low-Frequency Updates ..... 110
*K. K. L. B. Adikaram, P. A. Jayantha*

19. Intelligent Travel Companion: An AI-Powered Personalised Itinerary Planner for Sri Lanka ..... 115
*W. L. C. Wickramasinghe, K. K. P. Subashini*

20. SignaSpeech: A Real-Time Mobile-Based Sign Language Recognition and Speech Conversion System Using YOLOv8 ..... 121
*Prabuddhi Navodya Mahayahage, Thisarani Wickremesinghe*

21. SafeCircle: A Real-Time Mobile Platform for Location Streaming, Augmented Reality Guidance, and Anti-Theft Recovery ..... 127
*Nuwangi Kavindya Premawansha, Pabudi Abeyrathne*

22. Development of a Unified Web-Based Quality Evaluation Platform for Mobile Applications through Automated Functional and Non-Functional Test Result Aggregation and Visualization ..... 132
*Jasmi Alankarage, Pavithra Subashini*

23. A Machine Learning Approach for Recommending Suitable Technology Stacks Based on Project Characteristics ..... 139
*S. Kanujan, H. M. K. T. Gunawardhana, R. M. T. Sandaruwan*

24. Ship Route and Weather-Aware Anomaly Detection System with Port-Level Berthing Delay Notifications ..... 145
*R A Wijethunga, TAH Dilpriya, D K Wijethunga*

25. An Adaptive Data Prioritization Framework Inspired by Neural Attention Mechanisms for Handwritten Image-Based Evaluation: A Case Study in Forensic Handwritten Text Recognition ..... 149
*Imalsha sandali Gedara, Dilpriya TAH*

26. AI-Based Real-Time Workout Posture Correction System ..... 155
*L.S.A. De Silva, G. Perera*

27. Perception Should Not Interpret: An Empirical Case for Separating Visual Description from Emotion Classification in Multimodal Conversational AI ..... 161
*W.K. Lakshan Dulhara, Hiruni Weerasinghe*

28. Child-Aware Expected-Harm Manoeuvre Selection for Autonomous Vehicles: A Pre-Registered Paired Simulation Study ..... 166
*T. R. D. T. Dulshan, Anton Jayakody*

29. Explainable Adaptive Conversational AI for Improving Spoken Technical Interview Responses ..... 172
*Weruni S Sathsarani*

30. Smart Urban Waste Collection System ..... 179
*Hansika Subhashini Keeragala, G. Perera*

31. Few Shot Gemstone Classification: Comparing Prototypical Networks Against Standard CNNs on Scarce Mineral Image Data ..... 185
*S. I. M. S. Wimalasiri, YKA Rathnasiri, Thilini Bakmeedeniya*

32. Learning Collector Orchestration with Proximal Policy Optimization in a Multi-Agent Requirements Engineering Pipeline ..... 191
*Sujair Ibrahim, Vadivel Abishethvarman, B. T. G. S. Kumara*

33. SolarisPredict-SL: A Weather-Driven Machine Learning Framework for Retrospective Net-Load Proxy Modelling and Indicative Dispatch Support in the Sri Lankan Power Grid ..... 196
*Geemal Muthugala*

34. Explainable AI for Academic Article Classification To Support Research and Learning ..... 201
*S. A. G. K. Gunasekara, P. K. D. K. Kaushalya*

35. Adversarial Threats and Robust Defenses in Satellite-Based Deep Learning for Multi-Hazard Early Warning Systems: A Survey of Flood and Landslide Prediction ..... 207
*Satheesha Fernando, Madusanka Mithrananda*

36. PACT: Post-hoc Adaptive Conditional Transfer for Risk-ControlledLLMAbstentioninLow-Resource-Languages ..... 212
*Senesh Fitzroy, Isidara Gallage, Sanuji Walgampaya*

37. AI-Powered Smart Passenger Travel Assistant: A Full Design, Implementation and Evaluation Study of Machine Learning Flight Delay Prediction, Route Guidance, and Sleep-Onset Boarding Alerts ..... 219
*H. A. Javindi Nethnika Heshani, Hirushi Dilpriya Thilakarathna*

38. Coherence, Not Volume: Sub-Threshold Dilution Attacks on Provenance-Based Agent Memory and a Spectral Defence ..... 226
*Senesh Fitzroy, Isidara Gallage, Sanuji Walgampaya, Thyagi Luvishewa*

39. Policy As Code Based System for Detection and Risk Prioritization of AWS IAM, S3 and Security Group Misconfigurations ..... 232
*Uvindu Perera, Madusanka Mithrananda*

40. A Comparative Evaluation of Regression and Classification Approaches for Student Performance Prediction Using Machine Learning ..... 238
*R.K.A.I.C. Gunarathna, R.W.K.T. Rajapaksha*

41. The Location-Identity Problem in Automated Land Valuation: Satellite Flood Hazard Variables and GN-Division Spatial Validation in Sri Lanka ..... 242
*P. A. K. Dilshan, Hirushi Dilpriya*

42. Short-Term Residential Load Forecasting: A MachineLearningEvaluationAgainstaPersistence Baseline Using Fused Smart-Meter, Weather, and Socio-Economic Data in Sri Lanka ..... 248
*Y.D. K. Soysa, R.W.K.T. Rajapaksha*

43. Leakage-Aware Multi-Horizon PM2.5 Forecasting at a Single New Delhi Station: Model Compari- son, Conformal Uncertainty, and Actual-Model Explanations ..... 255
*AHK Thiwanka, RWKT Rajapaksha*

44. Development and Evaluation of a Context-Aware Multimodal Machine Learning Framework for Real-Time Stress Inference Using Smartphone Camera PPG-Derived HRV, Keystroke Dynamics and Environmental Sensing ..... 261
*W. G. G. A. Sahan Kaushalya Dias, Nimesha Hewawasam*

45. Adaptive Confidence-Based Fusion of IoT Sensor Networks and Transformer-Based Visual Detec- tion for Real-Time Fire Identification ..... 267
*Bommera Sharath Chandra, Rampalli Shiva Pallavi, G. Rajashekhar*

46. CALM-Hallu: Calibration-Aware Claim-Level Hallucination Detection under Model, Language, and Decoding Shifts ..... 272
*Rahul Kumar, Tapsi Nagpal, Jatin Singh Rajawat*

47. Smart EV Routing and Charging Optimization for Sustainable Urban Mobility in Sri Lanka ..... 278
*Irantha Thisal Dayaratne, Pavithra Subhashani*

48. CardioSentinel AI: An Explainable Edge Artificial Intelligence Framework for Electrocardiogram Synthesis and Electrophysiological Arrhythmia Classification ..... 284
*B.S.S.T. Cooray*

49. Introspective Perception: Predicting Visual-Odometry Failure Under Controlled Degradation ..... 289
*S. H. G. G. Sewjith, M.T.A. Wickramasinghe, I. K. D. Vilochana, T. H. Wickramage*

50. Uneven Ground: A Human-Validated Sidewalk Equity Assessment in Two Colombo Wards and a Benchmark of Pretrained Street-View Models ..... 295
*W. G. M. M. Ajmal Mansoor, Hiruni Weerasinghe*

51. Store-and-Forward GPS/GSM Train-Location Tracking Under Intermittent Connectivity: A Sri Lankan Railway Prototype ..... 302
*Pawani Maheshika Bandara, Janith Disanayake, Niranga Dharmaratna*

52. Machine Learning and Explainable Artificial Intelligence for Thalassemia Screening: A Review of CBC-Based Approaches,
Clinical Interpretability, and Research Gaps ..... 308
*Izna Idris, G. Perera*

53. Melanoma Skin Cancer Detection and Prevention Mobile Application Using Machine Learning ..... 314
*K L D P P Kakulandala*

54. An Instrumentation-Centered Multimodal AI Framework for Continuous Wearable Neurological Monitoring: A Conceptual Framework for Migraine Prediction ..... 318
*Lathushan RK, Damayanthi Dahanayake, Supun Gajendrasinghe*

55. Dual-Domain Deep Learning for MRI Brain Tumor Segmentation Using Spatial and FFT-Derived Frequency Representations ..... 326
*W A C K Wickramasinghe, T A H Dilpriya*

56. Characterizing Electrophysiological Responses of Solanum Lycopersicum To Environmental Stress Using Machine Learning ..... 333
*R N I Basnayake, G M S C Gajendrasinghe, C Disanayake*

57. AMMFF: Adaptive Multi-Modal Fusion for Chest Radiograph Diagnosis under Missing Clinical Data ..... 339
*H P G D Dilina Mewan, Rasika Ranaweera*

58. An ICU Capacity-Aware Emergency Transfer Decision Support System: Multi-Objective Routing and Ambulance Dispatch for the Colombo Hospital Network ..... 345
*Gunawardane Oneli Wickramaratne, Hirushi Dilpriya*

59. An AI Powered Triage and Dynamic Queue Management System for Outpatient Departments in Sri Lankan Government Hospitals ..... 351
*W. M. H. S. Karunarathna, K. K. P. Subashini*

60. Multimodal Deep Learning for NSCLC Recurrence Prediction: Gated Fusion of Histopathology and Clinical Data with Clean Label Derivation and Explainable AI.....357
*Piyara Morawakaarachchi, Thisarani Wickramasinghe*

61. MediHealth Predict: An Explainable Machine Learning Web Platform for Early Multi-Disease Risk Prediction.....363
*G.M.A.V. Senevirathna, Thilini Bakmeedeniya*

62. Systematic Method to Manage Patient Flow and Queue for Echocardiography Units In Sri Lankan Public Hospitals .....372
*R A Manusha Prabath Ranaweera, Hirushi Dilpriya*

63. A Dual-Method Framework Integrating AHP-TOPSIS and Temporally Validated Machine Learning for Export Market Recommendation: Evidence from Sri Lanka's Apparel Sector .....378
*UGD Nijai, RWKT Rajapaksha*

64. Factors Affecting Citizens' Acceptance of the Inland Revenue Department's Online Tax Services in Sri Lanka: An Application of the Technology Acceptance Model (TAM).....384
*S.G.N.S. Rupasinghe, Thisarani Wickramasinghe*

65. UbanSync: A Centralized Urban Management System for Sri Lanka.....390
*D. N. Pevinya, G. Perera*

66. Transparency as a Substitute for Fairness? Organizational Justice and System Trust in Algorithmic Human Resource Management.....396
*L.S.V. Samarappulige, L.S. Lekamge, A.S.A. Perera*

67. Mathematical Modelling and Dynamic Graph Analysis of Real-Time Financial Market Interdependencies for Systemic Risk Prediction .....403
*W. D. J. I. Senarathna, T. A. H. Dilpriya*

68. AI-Driven Digital Platform Ecosystems: Adaptive Business Model Innovation, Intelligent Ecosystem Orchestration, and Sustainable Value Creation .....410
*G. C. M. Sewwandi, T. A. H Dilpriya*

69. Predicting Bitcoin Price Movements from Multi-Source News Sentiment Using an Exponential De- cay Market Memory Framework.....417
*L. A. D. K. S. Deshappriya, Samsudeen Ashad, T. A. H. Dilpriya*

70. A Web Based Decision Support Platform for Assessing Digital Readiness among Sri Lankan SMEs 423
*Uvini Kawya, Thisarani Wickramasinghe*

71. Assessing Enterprise ERP Cloud Migration Strategies in Emerging markets:A Critical Success Factor Model .....430
*M.B.S .Sasanka*

72. LLM-Assisted Navigation of Sri Lankan Procurement Regulations .....434
*B.A.N.N. Wimalarathne, R.G.N. Meegama*

73. Ethical Concerns of Artificial Intelligence Usage in Higher Education: A Study Among Undergraduate Students in Sri Lanka.....440
*Sanduni Wickramanayake, Dharani Rajasinghe, G.D.S. Silva, I.G.S. Minodi*

74. A Household-Level Profiling Framework for Residential Electricity Consumption in Sri Lanka .....446
*Sanuri A. Divyanjalee, Thilini Bakmeedeniya*

75. The Impact of Digital Literacy on the Effective Use of E-Learning Platforms Among Undergraduate Students in Sri Lanka. .....453
*Vimansa Gamage, Dharani Rajasinghe, Dewmini Silva, Rasheena Sewwandi, Imalsha Sithumi*

76. Urban Green Space Proximity and Population Coverage Using GIS and Urban Computing for Sustainable Planning in Colombo District Sri Lanka ..... 459
*W.I.Y. Rodrigo, Hiruni Weerasinghe*

77. A Role-Sensitive Framework for Assessing Organizational Readiness Under Sri Lanka's Personal Data Protection Act ..... 465
*Hishara Samaranayake, T A H Dilpriya, Madusanka Mithrananda*

78. Data Privacy Awareness Among Sri Lankan Higher Education Students: Comparing Student Perception with the Documented Data Governance
Practices of Generative AI Platforms ..... 471
*Jayani Adikari, Madusanka Mithrananda*

79. AI-Enabled Secure Digital Support Platform for Women's Safety, Wellbeing, and Empowerment in Sri Lanka ..... 478
*R. G. Risini Senara Wijewickrama, Hirushi Dilpriya*

# Explainable Machine Learning for Academic Procrastination Prediction: A Comparative Evaluation of Conventional and Retrieval-Based Models

Daminda Herath
*Department of Computer Science*
*General Sir John Kotelawala Defence University*
*Dehiwala-Mount Lavinia, Sri Lanka*
herath.hmds@kdu.ac.lk

Thanuja Ambegoda
*Department of Computer Science and Engineering*
*University of Moratuwa*
*Moratuwa, Sri Lanka*
thanuja@cse.mrt.ac.lk

Kaneeka Vidanage
*Department of Computer Science*
*General Sir John Kotelawala Defence University*
*Dehiwala-Mount Lavinia, Sri Lanka*
kaneeka@kdu.ac.lk

***Abstract*—Academic procrastination is a common issue among university students and is influenced by a range of psychological and behavioral factors. This study examines whether machine learning and retrieval-based approaches can identify patterns associated with academic procrastination. Using the Academic Procrastination Data of Students in Makassar, Indonesia, the study compares conventional machine learning, ensemble learning, deep learning, and Tabular Retrieval-Augmented Learning (TabR). Model performance is evaluated using predictive metrics, calibration analysis, statistical testing, and SHapley Additive exPlanations (SHAP)-based explanations. Among the evaluated models, Full TabR achieved the highest holdout accuracy (0.8475) and F1-score (0.8388). However, five-fold cross-validation showed slightly higher mean accuracy and weighted F1-score for the Multilayer Perceptron (MLP), with no statistically significant differences between the models ($p > 0.05$). MLP also produced better-calibrated probability estimates than TabR. SHAP analysis identified Emotional, Personality, Cognitive, and Social factors as important contributors to the predictions, while TabR additionally incorporated information from similar student representations. These findings suggest that retrieval can provide contextual information, although its contribution depends on the characteristics of the available data. Overall, the study highlights the importance of considering predictive performance, reliability, and explainability together when developing Artificial Intelligence (AI)-based approaches to understanding academic procrastination.**

***Keywords—Academic procrastination, machine learning, retrieval-augmented learning, explainable artificial intelligence, educational data mining***

## I. Introduction

Academic procrastination remains a major challenge in higher education. This phenomenon has been associated with reduced academic achievement, increased stress levels, and diminished psychological health among university students [1], [2], [3]. Procrastination is no longer viewed solely as a matter of inadequate time management. Previous studies have demonstrated that procrastination is affected by various interacting factors, including motivation, emotional processes, self-regulation, and cognitive control [4], [5]. The impact of these factors is able to vary considerably among students, impeding the identification of procrastination based on a single behavioral characteristic.

The increasing availability of educational and psychological data has created new opportunities to AI and machine learning (ML) techniques to analyze student behavior [6]. Furthermore, machine learning and deep learning (DL) methods have been increasingly utilized for educational prediction tasks, such as predicting academic performance, student engagement, and identifying students who may be at risk [7]. However, most predictive systems consider each student as an independent observation and focus primarily on individual student characteristics. Consequently, potentially valuable information regarding similarities between students may not be fully utilized.

This limitation matters for academic procrastination because students with similar psychological and behavioral profiles may show similar procrastination patterns. Retrieval-augmented learning yields a possible way to address this issue by combining learned representations with information obtained from similar instances [8], [9]. Rather than predicting solely from the target student's feature vector, a retrieval-based model can also use information from similar learners. Despite this potential, retrieval-augmented learning for academic procrastination prediction has received relatively little attention.

To investigate this approach, this study evaluates TabR for academic procrastination prediction. TabR combines tabular representation learning with nearest-neighbor retrieval, enabling information from similar learner representations to be considered during prediction. Its performance is compared with conventional machine learning, ensemble learning, and deep learning models. Rather than assuming that retrieval will automatically improve predictive performance, this study measures its contribution through model comparison, ablation analysis, calibration analysis, and statistical testing.

Predictive accuracy alone, however, does not provide a complete picture of a model's usefulness in an educational setting. Understanding the factors that contribute to a prediction is also important, particularly when the results may be used to support students or inform educational decisions. Therefore, SHAP-based explainability is used to examine the contributions of psychological and behavioral features to the model predictions. The analysis further compares the explanations generated by MLP and TabR to determine whether retrieval-based learning changes the way important features contribute to individual predictions.

This study deals with the following three research questions: RQ1: How do different machine learning, ensemble, deep learning, and retrieval-based models perform in predicting academic procrastination? RQ2: Does retrieval augmentation result in a statistically significant enhancement in predictive performance as compared to a non-retrieval deep learning model? RQ3: Which psychological and behavioral features most influence academic procrastination prediction, and how do their contributions differ between MLP and TabR models?

This study makes three main contributions to the area of academic procrastination prediction. Firstly, it provides a head-to-head comparison of conventional ML, ensemble ML, deep learning and retrieval augmented tabular models using a range of psychological and behavioral student characteristics. Secondly, it looks at the contribution of the nearest-neighbor retrieval via ablation and statistical comparison to determine whether information from similar students adds any value to the prediction. Thirdly it uses SHAP based explainability and probability calibration to look at the differences in feature importance and prediction reliability between the conventional models and the retrieval augmented models. These are the key contributions of the study in terms of providing empirical evidence on the potential benefits and drawbacks of retrieval-based learning for academic procrastination prediction.

The remainder of this paper is organized as follows. Section II reviews related work, Section III describes the methodology, Section IV presents the experimental results, Section V discusses the findings and limitations, and Section VI concludes the paper.

## II. Related Work

This section reviews previous research on machine learning and deep learning for academic procrastination and educational prediction, followed by retrieval-augmented tabular learning and explainable AI. It then identifies the research gap that this study addresses.

### *A. Machine Learning for Academic Procrastination Prediction*

Machine learning has increasingly been used to analyze student behavior and predict academic outcomes [7], [10]. In academic procrastination research, previous studies have identified psychological and behavioral factors such as self-regulation, fear of failure, motivation, and emotional characteristics as important predictors [3], [4]. This evidence indicates that multiple factors influence procrastination and that data-driven approaches can be used to investigate it.

### *B. Deep Learning for Educational Prediction*

Researchers have applied deep learning to educational prediction because it can learn nonlinear relationships among multiple student characteristics [9], [11]. Multilayer Perceptron (MLPs) are suitable for structured tabular data, whereas Long Short-Term Memory (LSTM) networks are primarily designed for sequential information. Since the dataset used in this study mainly contains psychological and behavioral characteristics rather than temporal sequences, both MLP and LSTM are evaluated to examine their suitability for academic procrastination prediction.

### *C. Retrieval-Augmented Tabular Learning*

Retrieval-augmented learning extends tabular prediction through incorporating information from similar observations [8]. TabR combines learned feature embedding with nearest-neighbor retrieval, allowing predictions to use contextual information from similar instances [12]. This approach is potentially relevant to academic procrastination because students with similar psychological and behavioral characteristics may exhibit similar procrastination patterns.

### *D. Explainable Artificial Intelligence*

Explainable artificial intelligence (XAI) is important in educational applications because predictive accuracy alone does not explain why a model produces a particular prediction [13]. Understanding the contribution of individual features can provide useful information for explaining student predictions. In this study, SHAP is used to examine the feature contributions of MLP and TabR at both global and individual levels.

### *E. Research Gap*

Previous studies have applied machine learning to academic procrastination prediction. For example, Song et al. [1] used a decision tree model to identify factors associated with procrastination among university students. Other studies have examined psychological and behavioral factors influencing procrastination [4], [5]. Although these studies demonstrate the potential of predictive modeling, the application of retrieval-augmented tabular learning to academic procrastination remains insufficiently explored. TabR, introduced by Gorishniy et al. [8], incorporates information from similar instances to support tabular prediction. However, its potential contribution to academic procrastination prediction has not been adequately investigated. Furthermore, the combined examination of retrieval-based prediction, model calibration, and SHAP-based explainability remains an important area for investigation. This study addresses these gaps through comparative model evaluation, retrieval ablation, statistical testing, calibration analysis, and comparative SHAP explanations.

## III. Methodology

This study suggests an approach for predicting academic procrastination using students' psychological and behavioral characteristics. The methodology embraces data collection and preparation, exploratory analysis, target label construction, model training, result evaluation, explainability, and risk prediction. Fig. 1 illustrates the complete process.

### *A. Data Collection and Preprocessing*

This study uses the Academic Procrastination Data of Students in Makassar, Indonesia, obtained from Mendeley Data [14]. The dataset contains 586 university student records, including 18 questionnaire items measuring academic procrastination and 26 items examining its associated factors. Data preprocessing included converting questionnaire responses into numerical values and removing records with all questionnaire responses missing. Missing questionnaire values were imputed using the median, while categorical variables were imputed using the mode and encoded numerically. Numerical features were standardized for subsequent analysis and model development.

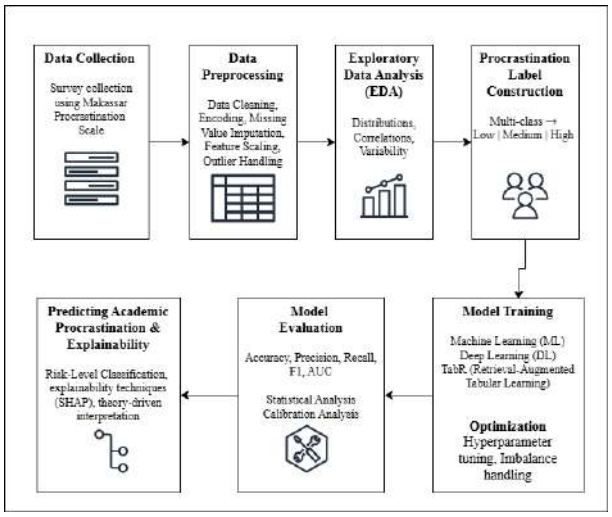



Fig. 1. Proposed framework for academic procrastination prediction.

### *B. Exploratory Analysis and Label Construction*

Exploratory data analysis (EDA) examined feature distributions and relationships [15]. The 18 questionnaire items measuring academic procrastination were aggregated to calculate a direct procrastination score. K-Means clustering (k=3) was subsequently applied to this score to derive three procrastination categories, labeled Low, Moderate, and High according to their respective mean scores. The six associated-factor dimensions, comprising Cognitive, Emotional, Motivational, Behavioral, Social, and Personality scores, were retained as independent predictor variables. This separation reduces direct dependency between target construction and predictor representation.

### *C. Model Training*

Four groups of models were evaluated: conventional machine learning (Decision Tree, Naive Bayes, and SVM) [16], ensemble learning (XGBoost, Random Forest, and AdaBoost) [17], and deep learning such as MLP, Artificial Neural Network (ANN), and LSTM. MLP served as the primary tabular-data baseline, while LSTM was included despite the absence of explicit temporal sequences.

The retrieval-based experiments evaluated TabR using an MLP encoder to generate 16-dimensional embeddings [8]. L2-normalized embeddings were indexed using Facebook AI Similarity Search (FAISS) IndexFlatIP for exact cosine-similarity search [18]. The five nearest training instances (k=5) were retrieved using similarity-weighted voting. Three configurations were evaluated: Retrieval Only, MLP + Retrieval (equal probability fusion), and Full TabR (weighted fusion, alpha=0.65). The encoder comprised three hidden layers (128, 64, and 32 neurons), dropout rates of 0.30 and 0.20, and a 16-dimensional embedding layer. It used Adam optimization, a batch size of 16, and a maximum of 100 epochs. The retrieval index contained training instances only.

### *D. Model Evaluation*

Model performance was assessed using accuracy, precision, recall and F1-score [19]. A stratified 80:20 training–test split was used for the primary model evaluation. Five-fold stratified cross-validation was additionally conducted to examine the consistency of MLP and TabR performance across different data partitions. Preprocessing parameters were estimated using only the training data within each fold. Model calibration was assessed using Log Loss, the Brier Score, and Expected Calibration Error (ECE). The performance differences between MLP and TabR were examined using the Wilcoxon signed-rank test and paired t-test based on fold-level results.

### *E. Explainability and Risk Prediction*

SHAP-based analysis examines global and instance-level feature contributions and compares explanations from MLP and TabR [20]. Finally, the trained models classify students into Low, Moderate, or High procrastination risk, providing an interpretable representation of the predicted procrastination level.

## IV. RESULTS

The experimental results obtained using the Mendeley academic procrastination dataset are presented in this section. The evaluated models are compared using accuracy, precision, recall, and F1-score. Further analyzes examine the contribution of retrieval via ablation experiments, the reliability of model probabilities through calibration analysis, the statistical significance of performance differences, and model behavior through SHAP-based explainability. The results are presented according to the main stages of the proposed experimental framework.

### *A. Baseline Model Performance*

Table I shows how the three baseline classifiers performed. SVM achieved the highest accuracy (0.8220) and F1-score (0.8184), followed by Naive Bayes. Decision Tree performed much worse, which suggests that models able to learn more flexible classification boundaries represent the dataset better.

TABLE I. BASELINE MODEL PERFORMANCE

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| SVM | 0.8220 | 0.8210 | 0.8220 | 0.8184 |
| Naive Bayes | 0.7966 | 0.7924 | 0.7966 | 0.7934 |
| Decision Tree | 0.6356 | 0.6414 | 0.6356 | 0.6300 |

### *B. Ensemble Model Performance*

Table II shows the results for the ensemble models. XGBoost performed best, with an accuracy of 0.8051 and an F1-score of 0.8000. Random Forest came next with an accuracy of 0.7881, and AdaBoost had the lowest performance in this group.

TABLE II. ENSEMBLE MODEL PERFORMANCE

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| XGBoost | 0.8051 | 0.8029 | 0.8051 | 0.8000 |
| Random Forest | 0.7881 | 0.7871 | 0.7881 | 0.7832 |
| AdaBoost | 0.7627 | 0.7606 | 0.7627 | 0.7610 |

### *C. Deep Learning Model Performance*

Table III presents the results for the deep learning models. MLP achieved the highest performance, with an accuracy of 0.8390 and an F1-score of 0.8347. ANN and LSTM performed considerably lower. In particular, the lower performance of LSTM is consistent with the structure of the dataset, which does not provide an explicit temporal sequence for learning sequential dependencies.

TABLE III. DEEP LEARNING MODEL PERFORMANCE

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| MLP | 0.8390 | 0.8341 | 0.8390 | 0.8347 |
| ANN | 0.6102 | 0.5814 | 0.6102 | 0.5712 |
| LSTM | 0.5763 | 0.4009 | 0.5763 | 0.4693 |

These results identify MLP as the strongest neural baseline and provide the reference model for the subsequent retrieval-based experiments.

### *D. Retrieval-Augmented Learning and Ablation Analysis*

The retrieval experiments investigate whether information from similar student representations provides additional predictive value. FAISS is used to retrieve nearest representations from the training partition. Table IV compares the retrieval configurations.

TABLE IV. RETRIEVAL-AUGMENTED LEARNING AND ABLATION RESULTS

| Model Configuration | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Retrieval Only | 0.8390 | 0.8341 | 0.8390 | 0.8347 |
| MLP + Retrieval | 0.8390 | 0.8355 | 0.8390 | 0.8309 |
| **Full TabR (FAISS)** | **0.8475** | **0.8464** | **0.8475** | **0.8388** |

Full TabR achieved the highest accuracy (0.8475) and F1-score (0.8388). Retrieval Only achieved the same evaluation metrics as the MLP baseline (accuracy: 0.8390; F1-score: 0.8347), although the two configurations employ different prediction mechanisms. Retrieval Only uses similarity-weighted voting based on retrieved neighbors, whereas MLP generates predictions through its

classification layer. MLP + Retrieval achieved the same accuracy but a slightly lower F1-score (0.8309). These results suggest that retrieved information did not consistently improve predictive performance. One possible explanation is that the original psychological and behavioral features already capture substantial predictive information, limiting the additional contribution of retrieved representations.

### *E. Calibration Analysis*

Beyond classification performance, the reliability of predicted probabilities was examined using Log Loss, Brier Score, and Expected Calibration Error (ECE). Lower values indicate better calibration. Table V shows calibration performance.

TABLE V. CALIBRATION PERFORMANCE

| Model | Log Loss ↓ | Brier Score ↓ | ECE ↓ |
|---|---|---|---|
| **MLP** | **0.4565** | **0.1314** | **0.0734** |
| TabR | 0.5364 | 0.1439 | 0.1002 |

MLP obtained lower values for all three calibration measures. Thus, although TabR remained competitive in predictive performance, MLP produced better-calibrated probability estimates. This indicates that retrieval augmentation did not improve probability reliability in the current dataset.

### *F. Statistical Comparison*

A five-fold cross-validation experiment was conducted to compare the predictive performance of MLP and TabR. MLP achieved a higher mean accuracy ($0.8123 \pm 0.0177$) than TabR ($0.8038 \pm 0.0185$). Similarly, MLP obtained a higher mean weighted F1-score ($0.8073 \pm 0.0202$) than TabR ($0.7991 \pm 0.0200$). However, the differences between the models were small. A statistical comparison was conducted between MLP and TabR using the Wilcoxon signed-rank test and paired t-test. As shown in Table VI, neither accuracy nor F1-score produced a statistically significant difference at the 0.05 significance level. Therefore, the results do not provide sufficient evidence to conclude that retrieval augmentation consistently increases predictive performance over MLP.

TABLE VI. STATISTICAL COMPARISON OF MLP AND TABR

| Metric | Wilcoxon p-value | Paired t-test p-value |
|---|---|---|
| Accuracy | 0.3750 | 0.2309 |
| F1-score | 0.3125 | 0.2535 |

### *G. Explainability Analysis*

SHAP analysis identified Emotional Score as the most influential feature in both models. In the MLP model, Emotional Score had the highest mean absolute SHAP value (0.1774), followed by Cognitive Score (0.0916), Personality Score (0.0859), and Social Score (0.0752), as shown in Fig. 2.

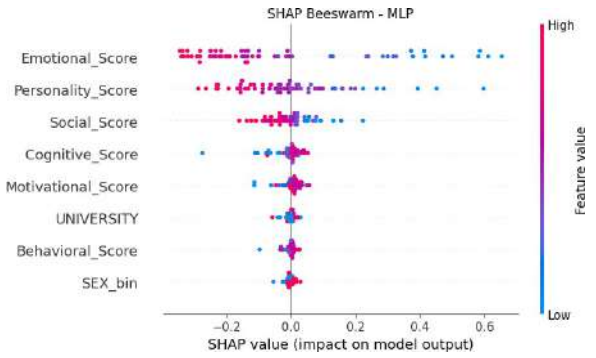



Fig. 2. Global SHAP feature importance of the MLP model based on mean absolute SHAP values.

The TabR model also identified Emotional Score as the most influential original feature (0.0472). Personality Score (0.0372) and RET Personality Score (0.0372) followed closely, as shown in Fig. 3.

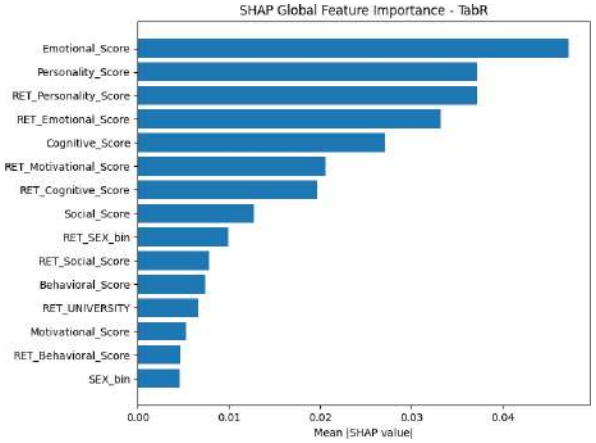



Fig. 3. Global SHAP feature importance of the TabR model based on mean absolute SHAP values.

Overall, the SHAP analysis highlights the importance of emotional, cognitive, personality, and social characteristics in the predictions of both models. It also demonstrates that TabR incorporates information from retrieved student representations, providing a different pattern of feature contributions compared with MLP.

## V. Discussion

The results demonstrate that model performance varied across the evaluated approaches. For RQ1, in baseline models, the best was SVM, while among the ensemble approaches, the best was XGBoost. Among the deep learning models, MLP performed the best overall accuracy. In the retrieval-based experiments, Full TabR achieved the highest holdout accuracy of 0.8475. However, five-fold cross-validation showed slightly higher mean accuracy for MLP than for TabR. These findings highlight the importance of evaluating models across different validation settings rather than relying on a single test result. The findings indicate that the MLP and Full TabR models are well suited to capturing relationships among the psychological and behavioral characteristics in the dataset. In contrast, the relatively poor performance of the LSTM model suggests that a sequential architecture may offer limited advantages when the available data do not contain strong temporal patterns [21]. The models also highlight the importance of psychological characteristics in understanding student procrastination, which is consistent with previous studies associating procrastination with self-regulation, motivation, fear of failure, and emotional factors [2], [4].

For RQ2, retrieval augmentation did not lead to a statistically meaningful improvement over MLP. Full TabR achieved the highest holdout accuracy compared with MLP, the cross-validation results showed slightly higher mean accuracy for MLP. Both the Wilcoxon and paired t-tests found no significant difference between MLP and TabR ($p > 0.05$). The comparable performance of MLP and TabR suggests that the original feature representation already captures important predictive information. The relatively small dataset may also limit the benefits of retrieving similar student representations.

For RQ3, the SHAP analysis found that emotional, personality, cognitive, and social scores were some of the most important factors in the model's predictions. TabR also gave more weight to retrieved features, which suggests that information from students with similar profiles may have influenced its predictions. These results match earlier research that points to the importance of psychological and behavioral factors in academic procrastination [4], [5]. Still, the SHAP results show associations at the model level and do not prove causal relationships.

### *A. Implications*

The results show that predictive accuracy, calibration, and explainability all need to be considered when building educational AI systems. MLP gave more accurate calibration than TabR, but TabR provided better context by using retrieved features [13]. Retrieval augmentation can help with contextual representation, but it does not always improve predictive performance for every dataset.

### *B. Limitations and Future Work*

This study has several limitations. First, it uses a single dataset of 586 university students from Makassar, Indonesia, limiting the generalizability of the findings. Second, the procrastination categories were generated using K-Means clustering. The use of associated-factor scores in both target construction and prediction may introduce dependency and influence model performance. Future studies should examine independently validated labels.

Furthermore, the dataset contains questionnaire-based information without longitudinal learning activity records. Integrating LMS data could provide a better understanding of changes in student behavior over time. Finally, external validation using larger datasets from different institutions is needed to assess model reliability and determine whether retrieval-based approaches provide consistent benefits.

## VI. Conclusion

This study investigated the application of machine learning and retrieval-based methods for predicting academic procrastination categories from questionnaire data on students. The MLP performed best of the standard deep learning

approaches and Full TabR achieved the highest holdout accuracy (0.8475). Five-fold cross-validation however showed a slightly better mean accuracy for MLP than TabR with no significant difference between the models. Thus, retrieval augmentation did not appear to provide a consistent benefit over the original feature representation.

The SHAP analysis showed that the Emotional Score was by far the most important feature in both models. Cognitive, Personality and Social Scores all made a contribution to the MLP model predictions, while TabR used both the original and the retrieved features. This underlines the potential of explainable machine learning to identify the features that are most influential in relation to the procrastination categories that have been constructed.

Despite the findings, the study is constrained by its use of a single dataset and clustering derived target labels. Future work should make use of larger and more diverse student datasets, incorporate longitudinal LMS data and validate the proposed approaches using independently derived labels. Further work is also required to investigate the utility of the various retrieval and explainability methods in an educational context.

Overall, this study is an empirical comparison of conventional machine learning, deep learning and retrieval-based approaches and it does a good job of highlighting both the potential and the limitations of explainable machine learning in the context of academic procrastination.

## REFERENCES


[1] P. Song, X. Liu, X. Cai, M. Zhong, Q. Wang, and X. Zhu, “Predictive analysis of college students’ academic procrastination behavior based on a decision tree model,” *Humanities and Social Sciences Communications*, vol. 11, no. 1, Jul. 2024, doi: 10.1057/s41599-024-03300-1

[2] Y. Fentaw, B. T. Moges, and S. M. Ismail, “Academic procrastination behavior among public university students,” *Education Research International*, vol. 2022, pp. 1–8, Jul. 2022, doi: 10.1155/2022/1277866

[3] Rozental, D. Forsstro¨m, A. Hussoon, and K. B. Klingsieck, “Procrastination among university students: differentiating severe cases in need of support,” *Frontiers in Psychology*, vol. 13, p. 783570, Mar. 2022, doi: 10.3389/fpsyg.2022.783570.

[4] S. A. Zarrin and E. Gracia, “Prediction of academic procrastina-tion by fear of failure and self-regulation,” *Educational Sciences: Theory & Practice*, vol. 20, no. 3, pp. 34–43, Jul. 2020, doi: 10.12738/jestp.2020.3.003

[5] Y. Ma and M. Chen, “AI-empowered applications effects on EFL learners’ engagement and academic procrastination,” *BMC Psychology*, vol. 12, no. 1, p. 739, Dec. 2024, doi: 10.1186/s40359-024-02248-w.

[6] M. A. Hashmani, M. M. Memon, M. Ebrahim, K. Raza, and A. A. Arain, “Student procrastination analysis in virtual learning environ-ments,” *AIP Conf. Proc.*, vol. 2433, Art. no. 020005, Jan. 2022, doi: 10.1063/5.0072423.

[7] Dawar, S. Negi, S. Lamba, and A. Kumar, ”Enhancing student academic performance forecasting: A comparative analysis of machine learning algorithms,” *SN Computer Science*, vol. 5, no. 6, Art. no. 758, 2024, doi: 10.1007/s42979-024-03118-3.

[8] Y. Gorishniy, I. Rubachev, and A. Babenko, ”TabR: Tabular deep learning meets nearest-neighbors,” *arXiv preprint* arXiv:2307.14338, 2023.

[9] R. Levin, V. Cherepanova, A. Schwarzschild, A. Bansal, C. B. Bruss,

[10] T. Goldstein, A. G. Wilson, and M. Goldblum, ”Transfer learning with deep tabular models,” arXiv preprint arXiv:2206.15306, Jun. 2022, doi: 10.48550/arXiv.2206.15306.

[11] D. Herath, C. Dinuwan, C. Ihalagedara, and T. Ambegoda, “Enhanc-ing educational outcomes through AI powered learning strategy recommendation system,” *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 10, pp. 739–748, 2024, doi: 10.14569/IJACSA.2024.0151075.

[12] S. B. Dias, S. J. Hadjileontiadou, J. Diniz, and L. J. Hadjileontiadis, “DeepLMS: A deep learning predictive model for supporting online learning in the Covid-19 era,” *Scientific Reports*, vol. 10, no. 1, Art. no. 19888, 2020, doi: 10.1038/s41598-020-76740-9

[13] D. Herath and U. Thayasivam, “Improving Student Performance Prediction with Retrieval-Augmented Deep Learning: A TabR-Based Approach in Educational Data Mining”, *International Conference on Artificial Intelligence and Data Science (ICAIDS 2025)*, Nov. 2025. doi: 10.5281/zenodo.23166906.

[14] M. Afzaal, A. Zia, J. Nouri, and U. Fors, “Informative feedback and explainable AI-based recommendations to support students’ self-regulation,” *Technol. Knowl. Learn.*, vol. 29, no. 1, pp. 331–354, 2024, doi: 10.1007/s10758-023-09650-0

[15] M. Rusdi, *Academic procrastination data of students in Makassar, Indonesia*, Mendeley Data, Nov. 2020, doi: 10.17632/8z9532ps65.10.

[16] A. Magray and G. K. Sodhi, “Predicting student placement in college using machine learning,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 2, pp. 1384–1395, Feb. 2024, doi: 10.22214/ijraset.2024.58548.

[17] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, Fujita, and N. A. Md. Ghani, “Multiclass prediction model for student grade prediction using machine learning,” IEEE Access, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[18] M.-R. Borna, H. Saadat, A. T. Hojjati, and E. Akbari, “Analyzing click data with AI: implications for student performance prediction and learning assessment,” *Frontiers in Education*, vol. 9, Dec. 2024, doi: 10.3389/feduc.2024.1421479.

[19] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, “Retrieval-augmented generation for educational application: A systematic survey,” Computers and Education: Artificial Intelligence, vol. 8, p. 100417, May 2025.

[20] M. P. R. I. R. Silva, R. A. H. M. Rupasingha, and B. T. G. S. Kumara, “Identifying complex causal patterns in students’ performance using machine learning,” Technology, Pedagogy and Education, vol. 33, no. 1, pp. 103–119, Dec. 2023, doi: 10.1080/1475939X.2023.2288015.

[21] M. Qiang, Z. Liu, and R. Zhang, “Explainable AI in education: integrating educational domain knowledge into the deep learning model for improved student performance prediction,” Scientific Reports, vol. 16, no. 1, Feb. 2026, doi: 10.1038/s41598-026-40538-y.

[22] G. Theophilus, C. I. Eke, O. J. Agushaka, and T. Moses, “Shallow and deep learning approaches for predicting learners’ engagement in electronic learning systems: feature selection, extraction and engineer-ing approaches,” *Neural Computing and Applications*, vol. 38, no. 9, Apr. 2026, doi: 10.1007/s00521-026-11987-3.

# FedQHO-Med: Energy-Aware Quantum-Inspired Spectrum Allocation with Federated Gradient-Aligned Learning for 6G Cognitive Radio Medical-IoT Networks

W. D. J. I. Senarathna
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
djisenarathna@students.nsbm.ac.lk

T. A. H. Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—The sixth generation (6G) of wireless networks is expected to interconnect an unprecedented density of devices, among which battery-constrained medical Internet-of-Things (IoT) nodes impose simultaneously stringent demands on spectral efficiency, energy frugality, and data privacy. Cognitive radio networks alleviate spectrum scarcity through opportunistic access, yet dynamic spectrum allocation in such settings remains an NP-hard combinatorial problem that classical and centralized intelligent methods address only partially. This paper proposes FedQHO-Med, a two-layer framework that unifies energy-aware quantum-inspired spectrum allocation with privacy-preserving federated learning for 6G cognitive radio medical-IoT networks. In the first layer, a quantum-inspired quantum-behaved particle swarm optimizer encodes channel and power assignments as quantum angles and embeds a simultaneous wireless information and power transfer term directly within its fitness function, so that throughput, interference, and harvested energy are jointly optimized under the color-sensitive graph-coloring model. In the second layer, distributed secondary users collaboratively train a cooperative spectrum-sensing model through federated averaging, with local updates produced by H. Dilpriya's Momentum (HDM), a convergence-guaranteed gradient-aligned optimizer that explicitly counters the gradient divergence induced by non-independent and identically distributed client data. The framework is evaluated through extensive simulations comprising thirty independent allocation trials and eight federated training seeds, with statistical significance assessed by the paired Wilcoxon signed-rank test. The proposed allocator attains the highest network throughput and spectrum utilization, significantly outperforming particle swarm optimization, grey wolf optimization, the snake optimizer, and the quantum genetic algorithm, whereas HDM achieves the highest within-client gradient alignment and the most stable convergence among the compared methods at competitive accuracy. Ablation and scalability analyses delineate the precise sources of these gains and the operating regimes in which the framework is most effective, providing an honest and reproducible account of both its strengths and its limitations for sustainable 6G spectrum management.***

***Keywords*—Cognitive radio networks, Dynamic spectrum allocation, Quantum-inspired optimization, Federated learning, Energy harvesting, 6G**

## I. Introduction

The sixth generation (6G) of wireless networks envisions pervasive, intelligent connectivity at extreme device densities under stringent latency and sustainability constraints [1], with mission-critical and human-centric services placing security, privacy, and energy efficiency at the center of future architectures [2]. Because the radio spectrum is finite and unevenly utilized, its efficient management is a prerequisite for this vision. Cognitive radio addresses spectrum scarcity by sensing the environment and accessing licensed bands opportunistically without harming primary users [3], a capability formalized as brain-empowered communication [4] that underpins the dynamic spectrum access paradigm, regarded as the most promising route to higher utilization in dense deployments [5]. The pressure is acute in medical Internet-of-Things (IoT) settings, where clinical intelligence such as electroencephalography (EEG)-based seizure prediction [6] relies on continuous physiological streams and 6G cognitive radio IoT networks serve vast device populations under strict security guarantees [7] and increasingly rely on artificial intelligence for sensing and allocation [8], yet wearable sensors are battery-limited, must coexist without interference, and carry physiological data that cannot be exposed during processing. The allocation task reduces to an NP-hard color-sensitive graph-coloring problem [9], motivating population-based metaheuristics such as particle swarm optimization [10], grey wolf optimization [11], and the snake optimizer [12], all risking premature convergence and disregarding energy and privacy constraints. Quantum-inspired computation mitigates this through superposition and rotation-gate search on classical hardware, as in quantum-behaved particle swarm optimization with global-search properties [13] and quantum-inspired cognitive-radio allocators [14], whose 6G efficiency a recent survey confirms [15], while federated learning trains shared models without exchanging raw data [16] and is foundational for privacy-sensitive intelligence [17].

These advances leave four gaps: quantum-inspired allocators remain centralized and rarely benchmarked against modern metaheuristics under identical conditions; energy harvesting is treated as an external constraint rather than an

objective; cooperative-sensing privacy is seldom addressed jointly with allocation; and the local optimizer that mitigates client drift is unexamined in the spectrum context. This work therefore contributes an energy-aware quantum-inspired allocator benchmarked head-to-head against four classical and modern metaheuristics on the color-sensitive graph-coloring model; FedQHO-Med, a two-layer framework coupling it with privacy-preserving federated sensing so that physiological observations never leave the originating node; the first use of H. Dilpriya's Momentum (HDM) [18], a convergence-guaranteed gradient-aligned optimizer, for federated spectrum sensing under non-independent and identically distributed (non-IID) data; and a statistically rigorous evaluation with ablation and scalability studies. Section II reviews related work, Sections III–V present the model, framework, and setup, Sections VI–VIII report results, discussion, and future directions, and Section IX concludes.

## II. Related Work

Intelligent spectrum allocation has long relied on population-based search, and learning-based optimization now also schedules energy-constrained resources, as in renewable dispatch with digital-twin battery models [19]; reliable operation, however, depends on accurate spectrum sensing, whose energy-detection and cooperative techniques degrade sharply at low signal-to-noise ratios (SNRs) [20]. The color-sensitive graph-coloring formulation [9] remains the canonical benchmark, and a substantial body of work compares particle swarm optimization [10] with modern solvers such as grey wolf optimization [11] and the snake optimizer [12]; collectively these illustrate a trend toward increasingly elaborate metaphors whose benefit over simpler solvers is not always demonstrated under controlled conditions. Quantum-inspired evolutionary computation offers a theoretically grounded alternative: the quantum-inspired evolutionary algorithm replaced crossover with a qubit representation and a rotation-gate operator, resisting premature convergence with small populations [21]; quantum-behaved particle swarm optimization added a stochastic contraction–expansion search with global-convergence properties [13]; and quantum-inspired particle swarm methods have been applied directly to cognitive radio allocation [14], a strand synthesized by the contemporary 6G survey [15]. These allocators, however, are predominantly centralized, are seldom benchmarked against the modern metaheuristics above, and rarely treat energy as a first-class objective, as energy-aware computing frameworks now do [22].

The federated dimension draws on maturing literature. Federated averaging established communication-efficient decentralized training [16], and subsequent surveys articulated the challenges of training over heterogeneous device populations [17]. Within cognitive radio, federated learning has been applied to spectrum allocation in open radio access networks [23] and to secure cooperative spectrum sensing where privacy and robustness are paramount [24], while a contemporaneous study combines reinforcement learning with the snake optimizer to enhance spectrum utilization [25]; that work, however, is neither federated nor quantum-inspired and considers a single optimizer in isolation. The present framework is therefore distinguished by its simultaneous integration of quantum-inspired allocation, energy-aware optimization, and privacy-preserving federated training, alongside a controlled benchmark against both classical and modern metaheuristics that prior work does not provide.

## III. System Model and Problem Formulation

A heterogeneous cognitive radio medical-IoT network is considered in which $N$ battery-constrained secondary users, representing wearable or in-ward medical sensors, opportunistically access $M$ licensed bands owned by primary users. Following the color-sensitive graph-coloring abstraction [9], availability is described by a binary matrix $L \in \{0,1\}^{N\times M}$with $L_{n,\,m} = 1$when band $m$ is locally idle for user $n$; interference by a tensor $C \in \{0,1\}^{N\times N\times M}$ with $C_{n,\,k,\,m} = 1$ when users $n, k$ would interfere on band $m$; and benefit by a reward matrix $B \in \mathbb{R}^{N\times M}$ combining band bandwidth with a per-link channel gain, so that $B_{n,m}$ is the throughput gained by granting band $m$ to user $n$. A solution is a binary assignment $A \in \{0,1\}^{N\times M}$. The topology and its constraints appear in Fig. 1.

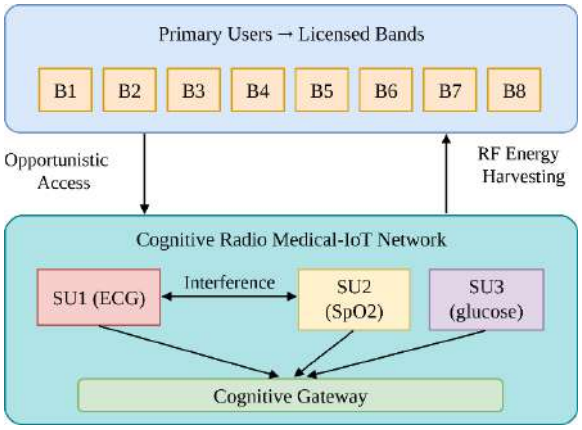



Fig. 1. System model: secondary users (SUs), e.g., electrocardiogram (ECG), blood-oxygen (SpO2), and glucose sensors, access idle primary-user bands under interference constraints and harvest radio-frequency (RF) energy.

The network throughput (utility) is the max-sum-reward

$$U(A) = \sum_{n=1}^{N}\sum_{m=1}^{M} A_{n,m}\, B_{n,m} \tag{1}$$

maximized subject to feasibility: an assignment never selects an unavailable band, never gives a band to two interfering users, and gives each user at most one band, that is

$$A_{n,m} \le L_{n,m}, \qquad \sum_{m=1}^{M} A_{n,m} \le 1, \qquad A_{n,m}A_{k,m} = 0 \;\; \forall (n,k,m)\colon C_{n,k,m} = 1 \tag{2}$$

Interference toward primary users is bounded by an interference-temperature limit and each admitted medical stream by a minimum signal-to-interference-plus-noise ratio. Because the users are energy-limited, a simultaneous wireless information and power transfer mechanism is incorporated in the spirit of the foundational capacity–energy analysis [26]: each occupied link draws transmit power $P_{\text{tx}}$ over a unit-duration slot, of which a fraction $\eta$ is recovered by radio-frequency (RF) harvesting, so the net energy per slot is

$$E(A) = (1-\eta)P_{\text{tx}} \sum_{n=1}^{N} \sum_{m=1}^{M} A_{n,m} \quad (3)$$

Rather than imposing energy externally, the formulation folds it into the objective as the energy-aware fitness

$$F(A) = U(A) - w_e E(A) \quad (4)$$

where $w_e$ trades off throughput against sustainability. The allocator maximizes $F(A)$; the reported metrics are the throughput $U(A)$, the spectrum utilization (the percentage of users granted a band), and the energy efficiency EE($A$) = $U(A)/E(A)$, the throughput delivered per unit of net energy. Since $B$ weights only throughput, utilization can rise without a throughput gain when a band goes to a weak-channel user. The resulting mixed-integer non-linear program motivates the quantum-inspired metaheuristic below. A complementary learning problem is also defined, in which each user classifies, from energy-detection features, whether a primary user is present, the goal being to train a shared classifier collaboratively without exchanging raw, potentially sensitive data.

## IV. Proposed FedQHO-Med Framework

FedQHO-Med comprises two cooperating layers, shown in Fig. 2. The allocation layer produces conflict-free, energy-aware assignments for the current network state, and the federated layer trains the cooperative spectrum-sensing model that informs availability estimates without centralizing any node's observations. The layers solve different problems and accordingly use different optimizers: a derivative-free quantum-inspired metaheuristic for the combinatorial allocation, and a gradient-based, convergence-guaranteed optimizer (HDM) for the differentiable learning task.

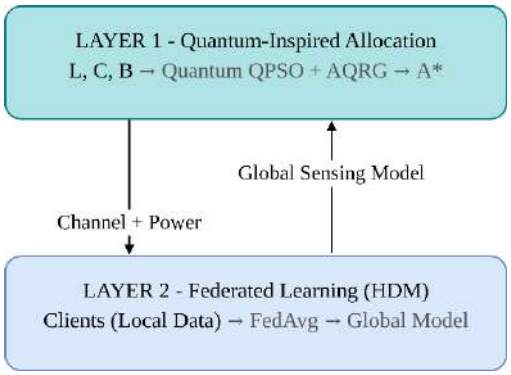



Fig. 1. FedQHO-Med architecture: quantum-behaved particle swarm optimization (QPSO) with an adaptive quantum rotation gate (AQRG) maps ($L$, $C$, $B$) to the best assignment $A^*$; federated averaging (FedAvg) with HDM trains the sensing model.

### A. Quantum-Inspired Allocation Layer

The allocation layer encodes each candidate through quantum angles, with every decision variable carrying an angle $\theta_j$ and selection probability $x_j = sin^2(\theta)$. A greedy, conflict-free decoder maps the continuous priority vector to a feasible binary assignment by admitting available user–band pairs in descending priority only when they violate neither availability nor interference, so every evaluated solution is feasible by construction.

Search proceeds by quantum-behaved particle swarm optimization (QPSO) [13], in which each particle is drawn toward a stochastic attractor formed from its personal and global bests with a contraction–expansion coefficient that narrows as search matures. Superimposed is an adaptive quantum rotation gate (AQRG), inspired by the rotation operator of quantum-inspired evolutionary computation [21], whose magnitude is , where measures relative quality, $\delta_1$ and $\delta_2$ are rotation gains, $\lambda$ is an annealing rate, $t$ and $T$ are the current and maximum iterations, $s \in [0,1]$ is a stagnation signal that grows when the global best stalls, and the annealed term injects exploration only when premature convergence threatens; because the incumbent best is always retained, this mechanism cannot degrade the final solution. Every particle is evaluated by the energy-aware fitness F(A), so throughput, interference avoidance, and harvested energy are optimized jointly. Each allocation iteration evaluates a fixed-size population whose dominant per-evaluation cost is the greedy decoder, of order $N\ M\ log(N\ M)$, so the allocation cost scales linearly in population size and iteration budget. The allocation state (particle angles, personal bests, and the $L$, $C$, $B$ inputs) holds about $2\times10^4$ values, below 1 MB, and one 300-iteration run took 2.5 s on average versus 2.3 s for particle swarm optimization in our Python implementation.

### B. Federated Spectrum-Sensing Layer

The federated layer trains a compact neural classifier across the users by federated averaging [16]: each round, every user initializes from the global model, performs local optimization on its private energy-detection data, and returns only its updated parameters for weighted aggregation, so observations remain on the device. Its defining feature is the local optimizer. Because heterogeneous nodes observe different SNRs and primary-user activity, their data are non-IID and their gradients diverge, a phenomenon termed client drift; the local rule used to counter it, H. Dilpriya's Momentum (HDM), is described in Section IV-C. Communication cost is proportional to the number of rounds and the model size, and no raw data are transmitted.

### C. The HDM Optimizer

The local training engine of the federated layer is H. Dilpriya's Momentum (HDM), a gradient-aligned momentum optimizer introduced in our earlier work [18] and adopted here, for the first time, in a cognitive-radio setting. Unlike adaptive-learning-rate methods that rescale each coordinate independently, HDM augments classical momentum with an alignment signal derived from the cosine similarity between consecutive mini-batch gradients. The momentum and a complementary correction term are modulated by this signal, $m_t = \beta\, m_{t-1} + (1 + 0.3w_t)g_t$, $h_t = \gamma_t(0.3\ cos_t + 0.7\ \bar{c}_t)g_t$, and $\varphi_t h_1 = \varphi_t - \mu(m_t + h_t)$, where $w_t$ is a non-negative alignment weight and $\bar{c}_t$ an exponential moving average of the alignment signal, with $w_t = \max(0, \bar{c}_t)$; $\cos_t$ is the cosine similarity between $g_t$ and $g_{t-1}$, $\gamma_t$ is the correction gain, $\varphi$ denotes the classifier parameters, and $\mu$ is the learning rate: when successive gradients agree the two terms are amplified to accelerate descent, whereas conflicting gradients are damped to suppress oscillation. This behavior suits federated learning over heterogeneous medical-IoT nodes, where non-IID local data drive client updates apart; by promoting gradient consistency within each client, HDM curbs the divergence that destabilizes federated averaging. Crucially, HDM retains a formal linear

convergence rate of order for strongly convex, smooth objectives [18], where κ is the condition number, a guarantee the heuristic baselines lack. Its memory grows linearly in the number of model parameters $d$, as for adaptive moment estimation, and each step adds only $O(d)$ inner-product work.

## V. Experimental Setup

All experiments were implemented in Python with fixed seeds; Table I lists the main parameters.

TABLE I. Principal Simulation Parameters.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Secondary users $N$ | 30 | Federated clients | 10 |
| Licensed bands $M$ | 8 | Communication rounds | 50 |
| Availability probability | 0.70 | Local epochs | 1 |
| Interference probability | 0.50 | Non-IID Dirichlet $\alpha$ | 0.3 |
| Population size | 30 | Sensing feature dimension | 8 |
| Iterations | 300 | Batch size | 64 |
| Energy weight $w_e$ | 0.15 | Allocation / federated seeds | 30 / 8 |

The default scenario has $N = 30$ users and $M = 8$ bands, a ward-scale size within the up-to-fifty-user regime of Section VI, with more users than bands so that bands must be shared. A band is idle for a user with probability 0.70 (each user keeps at least one) and a user pair interferes on a band with probability 0.50, leaving spectrum to allocate while a dense conflict graph keeps greedy assignment from being trivially optimal. $B$ is the product of a normalized band bandwidth drawn uniformly from [1, 5] and a link gain drawn uniformly from [0.5, 1]; $P_{tx} = 1$ (normalized) and $\eta = 0.4$. The weight $w_e = 0.15$ keeps the energy penalty small relative to $U(A)$, so energy separates solutions of similar throughput without overriding it. Every optimizer received thirty candidates over three hundred iterations (9,000 evaluations), by which most curves in Fig. 3 have flattened. Angles start at π/4 with Gaussian jitter (standard deviation 0.05); the AQRG uses $\delta_1 = 0.04$, $\delta_2 = 0.08$, and $\lambda = 4$, with $s$ saturating after eight stalled iterations, and the contraction–expansion coefficient falls linearly from 1.0 to 0.5. The baselines were particle swarm optimization (PSO) [10] (inertia 0.7, acceleration coefficients 1.5), grey wolf optimization (GWO) [11], the snake optimizer (SO) [12] with its published constants, and the quantum genetic algorithm (QGA) [21] (fixed rotation 0.05π), all sharing the search space, decoder, and objective.

For cooperative sensing, each of the ten clients holds 1,500 samples, each a vector of eight window energies (mean squared magnitude of 16 complex samples) labeled by primary-user presence. Each client has its own SNR, uniform in −12 to −2 dB where energy detection degrades [20], and its own primary-user prior from a Dirichlet distribution with $\alpha = 0.3$ clipped to [0.1, 0.9], so data are non-IID in features and labels; a balanced IID test set of 4,000 samples spans the same SNRs, and features are standardized. One local epoch per round limits drift as in federated averaging [16]. The classifier is a multilayer perceptron with two hidden layers of 64 ReLU units, default PyTorch initialization, and two-class cross-entropy loss, trained in batches of 64. HDM used $\beta = 0.9$, an alignment moving-average factor of 0.95, $\gamma = 2$ under a 115-step linear warm-up and a 1,150-step cosine schedule, and $\mu = 0.02$, with every optimizer's state re-initialized at each round; it was compared with adaptive moment estimation (Adam) [27], stochastic gradient descent with momentum 0.9 (SGD-M) [28], and root-mean-square propagation (RMSProp) [29], with learning rates 0.01, 0.05, and 0.01, all chosen, like HDM's $\mu$ and $\gamma$, by grid search on final accuracy. Metrics were test accuracy, within-client gradient alignment (mean cosine similarity of consecutive local mini-batch gradients), and stability (accuracy standard deviation over the last ten rounds). Significance used the paired Wilcoxon signed-rank test at the 5% level on identical scenarios per seed; eight federated seeds, fewer than the thirty allocation trials, reflect the cost of training.

## VI. Results

Fig. 3 plots the best energy-aware fitness $F(A)$ per iteration, averaged over the thirty trials. The proposed allocator starts more conservatively than PSO, reflecting the broader exploration of its quantum-angle representation, but overtakes every baseline and attains the highest final fitness, whereas PSO stagnates at a lower plateau indicative of premature convergence.

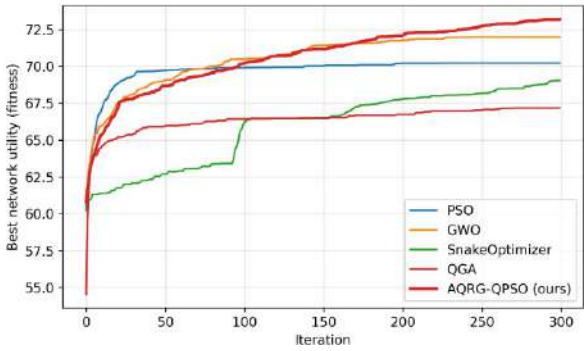



Fig. 2. Convergence of the best energy-aware fitness $F(A)$, averaged over thirty trials; AQRG-QPSO denotes the proposed allocator.

Table II shows that the proposed allocator achieves the highest throughput, utilization, and energy efficiency, with no interference violations for any method owing to the feasibility-preserving decoder. The throughput advantage is significant against every baseline (final column of Table II), the utilization advantage against all but GWO, and the energy-efficiency advantage only against SO and QGA.

TABLE I. Allocation Performance Over Thirty Trials (Mean ± Standard Deviation); p: Paired Wilcoxon Test on Throughput Versus the Proposed Allocator.

| Optimizer | Throughput | Utilization (%) | Energy efficiency | Violations | $p$ (throughput) |
|---|---|---|---|---|---|
| Particle swarm optimization | 72.80 ± 6.67 | 95.7 | 4.23 | 0 | $9.2 \times 10^{-6}$ |
| Grey wolf optimization | 74.62 ± 7.73 | 97.2 | 4.26 | 0 | $4.7 \times 10^{-3}$ |
| Snake optimizer | 71.64 ± 6.97 | 96.6 | 4.12 | 0 | $1.4 \times 10^{-6}$ |
| Quantum genetic algorithm | 69.71 ± 6.32 | 93.8 | 4.13 | 0 | $1.9 \times 10^{-9}$ |
| Proposed (FedQHO-Med allocator) | 75.83 ± 7.12 | 98.1 | 4.29 | 0 | — |

The ablation in Table III re-ran the allocator on the same thirty scenarios with independent optimizer seeds, so its full-allocator throughput (76.15) differs from Table II (75.83) only by run-to-run randomness, well within one standard deviation (7.12). Removing the QPSO contraction–expansion mechanism causes the largest throughput loss, identifying it as the principal driver; the energy-aware term gives a small but consistent energy-efficiency gain; and the AQRG yields no measurable throughput gain over a fixed angle, so it is retained only for its stagnation-escaping role.

TABLE II. ALLOCATION ABLATION OVER THIRTY TRIALS (MEAN; INDEPENDENT OPTIMIZER SEEDS).

| Configuration | Throughput | Utilization (%) | Energy efficiency |
|---|---|---|---|
| Full allocator | 76.15 | 98.0 | 4.32 |
| Without adaptive rotation gate | 76.78 | 98.9 | 4.31 |
| Without quantum contraction–expansion | 74.88 | 97.8 | 4.25 |
| Without energy-aware term | 76.09 | 98.2 | 4.30 |

Regarding scalability, when the network grows from ten to two hundred secondary users under the full evaluation budget applied identically to all methods, the proposed allocator is the strongest solver for small and medium networks of up to fifty users, the regime characteristic of realistic medical-IoT deployments within a single ward, improving on the best baseline by up to roughly four per cent; for very large networks PSO scales more gracefully and overtakes the quantum-inspired solver by a comparable margin, so the method is best suited to ward-scale networks.

Table IV reports the federated results. Under strongly non-IID data, HDM attains the highest within-client gradient alignment among the compared optimizers (0.370, significant against all three), late-training stability on par with SGD-M (0.10 vs. 0.09), and a final accuracy statistically tied with SGD-M while significantly exceeding Adam and RMSProp; RMSProp even shows negative alignment, indicating conflicting successive gradients.

TABLE III. FEDERATED OPTIMIZER COMPARISON OVER EIGHT SEEDS (MEAN ± STANDARD DEVIATION).

| Local optimizer | Final accuracy (%) | Within-client alignment | Stability (std) |
|---|---|---|---|
| Stochastic gradient descent with momentum | 81.08 ± 0.12 | 0.325 | 0.09 |
| Adaptive moment estimation | 80.26 ± 0.26 | 0.217 | 0.22 |
| Root-mean-square propagation | 78.01 ± 1.40 | −0.130 | 0.90 |
| **Proposed (HDM)** | **81.05 ± 0.17** | **0.370** | **0.10** |

The ablation in Table V shows that, on this readily saturated task, the correction term changes neither accuracy nor alignment (0.372 without it). The momentum-free variants score higher alignment (0.573 and 0.597) at no better accuracy because shorter steps without accumulated momentum leave consecutive gradients more alike; since alignment depends on step size, it is compared only among the tuned optimizers of Table IV.

TABLE IV. FEDERATED OPTIMIZER ABLATION OVER EIGHT SEEDS (MEAN).

| Configuration | Final accuracy (%) | Within-client alignment |
|---|---|---|
| Full optimizer (momentum and correction) | 81.05 | 0.370 |
| Without correction | 81.06 | 0.372 |
| Without momentum | 81.00 | 0.573 |
| Plain stochastic gradient descent | 81.03 | 0.597 |

## VII. DISCUSSION

The results indicate that quantum-inspired search gives a statistically significant throughput advantage of 1.6–8.8% over the four baselines at medical-IoT network sizes, and that gradient-aligned federated training improves update consistency under data heterogeneity. The allocator escapes the early stagnation that limits PSO, and embedding harvesting in the fitness yields a modest but consistent energy-efficiency gain without sacrificing throughput. The ablations locate the allocator's strength in the QPSO contraction–expansion dynamics rather than the AQRG; QGA, which shares the quantum-angle encoding without these dynamics, was the weakest solver, so the encoding alone does not explain the gain. In the federated layer, accuracy and stability rose with alignment, from RMSProp (−0.130, 78.01%) to Adam (0.217, 80.26%) to the momentum methods (about 81%), suggesting that alignment is a useful indicator of client drift; on this task, which saturates near a common ceiling, HDM contributes alignment and a convergence guarantee rather than accuracy. The formulation is also application-agnostic: any battery-constrained cognitive radio IoT network, such as industrial or environmental sensing, maps onto the same L, C, B inputs and fitness; medical-IoT is the motivating case because it imposes spectrum, energy, and privacy constraints jointly.

The study has several limitations. The quantum-inspired allocator, although superior for networks of up to fifty users, is overtaken by PSO in very large networks, positioning the framework for ward-scale rather than city-scale deployments. The sensing task saturates at moderate accuracy in the deliberately low SNR range of −12 to −2 dB, which compresses differences among strong optimizers and prevents HDM from showing an accuracy advantage. HDM also costs more per step than SGD-M, and the evaluation is simulation-based, so validation on a software-defined-radio testbed remains worthwhile.

## VIII. FUTURE DIRECTIONS

Several avenues merit investigation. The scalability limitation invites dimensionality-reduction and adaptive population-scaling strategies, possibly through hierarchical decomposition of the allocation problem across network regions [30]. The federated layer would benefit from more demanding tasks, such as richer radio-frequency modulation-classification datasets. Comparison with learning-based allocators such as deep reinforcement learning [19], adaptive throughput–energy weighting, and robustness against adversarial participants, for instance through blockchain-anchored audit trails of model updates [31], would enhance practical relevance, while testbed validation under realistic mobility would bridge the gap to

deployment in 6G medical-IoT and sustainable smart-city infrastructure [32].

## IX. Conclusion

This paper introduced FedQHO-Med, a two-layer framework that unifies energy-aware quantum-inspired spectrum allocation with privacy-preserving federated sensing for 6G cognitive radio medical-IoT networks. The allocation layer encodes assignments as quantum angles with an energy-harvesting term in its fitness and searches by an adaptively regularized quantum-behaved swarm, while the federated layer trains the sensing classifier with HDM. The allocator significantly outperformed PSO, GWO, SO, and QGA in throughput at medical-IoT network scales, and HDM attained the highest within-client gradient alignment among the compared optimizers, with accuracy and stability on par with SGD-M. In the default ward-scale scenario, the allocator raised mean throughput from 74.62 for the strongest baseline (GWO) to 75.83, served 98.1% of users, and needed about 2.5 s per run. Ablation and scalability analyses attributed these gains to their sources and identified the network sizes where the framework is most effective.

## References


[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020, doi: 10.1109/MNET.001.1900287.

[2] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?," *Nature Electronics*, vol. 3, pp. 20–29, 2020, doi: 10.1038/s41928-019-0355-6.

[3] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999, doi: 10.1109/98.788210.

[4] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005, doi: 10.1109/JSAC.2004.839380.

[5] F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006, doi: 10.1016/j.comnet.2006.05.001.

[6] J. I. Senarathna and T. A. H. Dilpriya, "A systematic comparative evaluation of machine learning and deep learning models for EEG-based epileptic seizure prediction with clinical deployment guidelines," in *Proc. IEEE Int. Res. Conf. Smart Comput. Syst. Eng. (SCSE)*, 2026, doi: 10.1109/SCSE70081.2026.11499844.

[7] N. Y. Al-Matari, A. T. Zahary, and A. A. Al-Shargabi, "A survey on advancements in blockchain-enabled spectrum access security for 6G cognitive radio IoT networks," *Scientific Reports*, vol. 14, art. 30990, 2024, doi: 10.1038/s41598-024-82126-y.

[8] Gbenga-Ilori, A. L. Imoize, K. Noor, and P. O. Adebolu-Ololade, "Artificial intelligence empowering dynamic spectrum access in advanced wireless communications: A comprehensive overview," *AI*, vol. 6, no. 6, art. 126, 2025, doi: 10.3390/ai6060126.

[9] Peng, H. Zheng, and B. Y. Zhao, "Utilization and fairness in spectrum assignment for opportunistic spectrum access," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 555–576, 2006, doi: 10.1007/s11036-006-7322-y.

[10] R. D. Rajasinghe and D. Wijesinghe, "A Federated Multimodal Deepfake Detection Framework for Enhancing the Integrity of Video Evidence in Court Systems," 2026 9th International Conference on Electronics Technology (ICET), Chengdu, China, 2026, pp. 806-811, doi: 10.1109/ICET69987.2026.11659066.

[11] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[12] F. A. Hashim and A. G. Hussien, "Snake Optimizer: A novel meta-heuristic optimization algorithm," *Knowledge-Based Systems*, vol. 242, art. 108320, 2022, doi: 10.1016/j.knosys.2022.108320.

[13] J. Sun, B. Feng, and W. Xu, "Particle swarm optimization with particles having quantum behavior," in *Proc. IEEE Congress on Evolutionary Computation (CEC)*, vol. 1, 2004, pp. 325–331, doi: 10.1109/CEC.2004.1330875.

[14] H. Gao, J. Cao, and M. Diao, "A simple quantum-inspired particle swarm optimization and its application," *Information Technology Journal*, vol. 10, no. 12, pp. 2315–2321, 2011.

[15] M. O. Butt, N. Waheed, T. Q. Duong, and W. Ejaz, "Quantum-inspired resource optimization for 6G networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 5, pp. 2973–3019, 2025, doi: 10.1109/COMST.2024.3519865.

[16] McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 54, 2017, pp. 1273–1282.

[17] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.

[18] J. I. Senarathna, "H. Dilpriya's Momentum (HDM): A multi-strategy gradient-aligned optimizer with adaptive per-parameter corrections, cosine-annealed scheduling, and convergence guarantees for deep neural networks," in *Proc. 4th Int. Conf. Adv. Netw. Technol. Appl. (APAN)*, Auckland, New Zealand, 2026, doi: 10.1109/APAN70967.2026.11709607.

[19] J. I. Senarathna and T. A. H. Dilpriya, "GIS-AHP renewable energy siting with deep reinforcement learning and digital twin battery for 24/7 carbon-free dispatch in Sri Lanka," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, 2026, doi: 10.1109/MERCon71835.2026.11691570.

[20] T. Yücek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 116–130, 2009, doi: 10.1109/SURV.2009.090109.

[21] K.-H. Han and J.-H. Kim, "Quantum-inspired evolutionary algorithm for a class of combinatorial optimization," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 6, pp. 580–593, 2002, doi: 10.1109/TEVC.2002.804320.

[22] J. I. Senarathna and T. A. H. Dilpriya, "Intelligent EcoETL: An energy-aware framework for sustainable extract-transform-load operations in big data processing," in *Selected Articles From the Int. Conf. Emerging Concepts and Research in Integrated Sustainable Adaptation*, D. Tantrigoda et al., Eds. Singapore: Springer, 2026, doi: 10.1007/978-981-92-0423-6_11.

[23] M. Asad and S. Otoum, "Federated learning for efficient spectrum allocation in open RAN," *Cluster Computing*, vol. 27, pp. 11237–11247, 2024, doi: 10.1007/s10586-024-04500-9.

[24] M. Wasilewska, H. Bogucka, and H. V. Poor, "Secure federated learning for cognitive radio sensing," *IEEE Communications Magazine*, vol. 61, no. 3, pp. 68–73, 2023, doi: 10.1109/MCOM.001.2200465.

[25] H. Farhi et al., "Enhancing spectrum utilization in cognitive radio networks using reinforcement learning with Snake Optimizer: A meta-heuristic approach," *Electronics*, vol. 14, no. 13, art. 2525, 2025, doi: 10.3390/electronics14132525.

[26] L. R. Varshney, "Transporting information and energy simultaneously," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2008, pp. 1612–1616, doi: 10.1109/ISIT.2008.4595260.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[28] T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964, doi: 10.1016/0041-5553(64)90137-5.

[29] T. Tieleman and G. Hinton, "Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, 2012.

[30] T. A. H. Dilpriya, G. H. J. Lanel, and B. V. N. C. Vidanage, "A strategy to strengthen and enhance the telecommunication network in Sri Lanka by using concepts of graph theory and linear programming models," *International Journal of Natural Sciences Research*, vol. 10, no. 1, pp. 1–20, 2022, doi: 10.18488/63.v10i1.2913.

[31] J. I. Senarathna and T. A. H. Dilpriya, "An integrated multi-layered anonymous whistle-blowing platform with steganographic covert channels, immutable blockchain evidence custody, and AI authenticity verification," in *Proc. NSBM Student Res. Symp. Innov. Technol. (NSRSIT)*, 2026, doi: 10.5281/zenodo.23132382.

[32] J. I. Senarathna and T. A. H. Dilpriya, "Sustainable urban redesign framework integrating energy efficiency, active mobility, and green infrastructure for tropical developing cities: Empirical study for Colombo 05," in *Proc. 6th Int. Conf. Elect., Comput. Energy Technol. (ICECET)*, 2026, doi: 10.1109/ICECET65726.2026.11632310.

# An LLM-Based Multi-Agent Framework for Adaptive Digital Marketing Campaign Design and Optimization

I.K.D.Vilochana
*Department of Software Engineering*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
ikdvilochana@students.nsbm.ac.lk

T. H. Wickramage
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
thwickramage@students.nsbm.ac.lk

M.T.A. Wickramasinghe
*Department of Software Engineering and Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

***Abstract*—Digital marketing campaign design is a multi-stage decision problem that couples strategy, audience analysis, channel selection, content planning, creative development, budget allocation, quality assurance, and evaluation. A single large language model (LLM) prompt can generate campaign text quickly, but it offers limited role specialization, weak traceability, and no systematic way to detect and repair internal inconsistencies before a plan is finalized. This paper presents the Knowledge-Augmented Multi-Agent Framework (KAMF), which organizes LLM capabilities as a collaborative team of specialized agents that operate over a shared campaign state. Dedicated Strategy, Audience, Platform, Content Strategy, Creative, and Budget agents each contribute a bounded artifact; a Critic agent reviews the integrated campaign and issues targeted, traceable revision instructions; and an Evaluation agent scores the result on eight campaign quality dimensions. Because every agent is augmented with the accumulated outputs and context of the agents before it, the framework produces a coherent, inspectable plan rather than a single opaque block of text. KAMF exposes three controlled conditions: a single-agent baseline, a multi-agent workflow without critique, and a multi-agent workflow with critic-guided revision. In an evaluation by 20 industry experts in digital marketing campaigns, the multi-agent conditions were rated above the single-agent baseline and the critic-guided condition was rated highest across the eight quality dimensions, suggesting higher expert-perceived plan quality, without establishing statistical significance or live-market effectiveness.***

***Keywords—campaign generation, digital marketing automation, large language models, marketing analytics, multi-agent systems***

## I. Introduction

Digital marketing campaign planning is a multi-stage decision problem rather than a single content-generation task. A usable campaign must connect organizational objectives to measurable goals, define audiences and journeys, select channels that fit those audiences, create a coherent content and creative plan, allocate constrained budgets, and establish criteria for evaluating the resulting plan. These activities are interdependent: a change in target audience can alter platform choice, content format, creative language, expected costs, and key performance indicators (KPIs). Recent work describes generative artificial intelligence as a general-purpose marketing technology relevant across innovation, customer interaction, content production, and decision support [1], [2], [3], yet a large language model (LLM) asked to produce an entire campaign in one pass must reason about many heterogeneous roles simultaneously, which risks shallow coverage, internal inconsistency, weak traceability, and limited opportunities for targeted correction.

Multi-agent LLM systems offer role decomposition. AutoGen and CAMEL organize collaboration among specialized agents [4], [5]; ReAct, Reflexion, and Self-Refine explore feedback and iterative improvement [6]–[8]. Yet these general-purpose systems do not, by themselves, define a marketing-specific sequence of schema-constrained campaign artifacts, issue-directed revision, and common-condition assessment. KAMF implements this integrated workflow using a shared state and a critic that targets individual campaign components. Its contribution is domain-specific orchestration and inspectability, not a novel general-purpose agent algorithm.

Campaign design requires coordinated strategy, segmentation, media selection, messaging, creative execution, budgeting, quality review, and evaluation. Many generative tools support isolated activities or a single conversational agent, leaving responsibilities and revision decisions difficult to inspect. The research problem is to provide one structured campaign-generation workflow with specialist stages, persisted intermediate decisions, targeted review, and experimental condition comparison.

The aim of this project is to design, implement, and validate a collaborative multi-agent LLM framework that transforms a structured marketing brief into a complete, coherent, and inspectable digital marketing campaign, in which specialist agents cooperate over a shared campaign state and an explicit critique and revision stage improves the plan before it is evaluated with a structured rubric. To achieve this aim, the work: (1) designs a multi-agent framework for end-to-end campaign generation; (2) automates strategy, audience segmentation, platform planning, content planning, creative generation, and budget allocation as explicit specialist stages; (3) introduces critic-based iterative review with traceable and targeted revision instructions; (4) evaluates the resulting campaign with a structured multidimensional rubric; and (5) supports a controlled comparison of single-agent, multi-agent without critic, and multi-agent with critic conditions.

The distinct contribution is the combined marketing-specific artifact workflow, traceable critic routing, and controlled ablation design. It does not establish a new LLM training method or superior live-market performance.

## II. Background and Related Work

### A. Traditional Marketing Campaign Methods

Traditional digital marketing campaign development is a structured, human-led process. Planners begin from a brief that captures the objective, target market, budget, and constraints, and then move through established stages: situation and audience analysis, segmentation and targeting, positioning and messaging, channel and media selection, content and creative production, budget and media allocation, and finally measurement against key performance indicators. Classic planning frameworks such as segmentation, targeting, positioning, the marketing funnel, and objective-and-task budgeting encode this decomposition and are typically executed by specialists (strategists, media planners, copywriters and analysts) whose outputs are reviewed and reconciled in campaign meetings. This division of labour improves quality and accountability, but it is time consuming, expensive, and difficult to scale, and the coordination between stages depends heavily on manual review. KAMF is motivated by this workflow: it preserves the specialist decomposition and the review step, but assigns each responsibility to a dedicated LLM agent operating over a shared, machine-readable campaign state.

### B. Generative AI in Marketing

Generative AI is increasingly treated as a general-purpose marketing technology rather than only a copywriting tool. Grewal et al. discuss its influence on marketing activities and the need to align generative systems with customer, firm and societal outcomes [1]. Cillo and Rubera identify opportunities across innovation and marketing processes while highlighting unresolved research questions around adoption and value creation [2]. Arora, Chakraborty, and Nishimura further demonstrate how LLMs can support stages of marketing research through synthetic respondents and hybrid AI–human workflows [9], and LLM-based marketing analytics work shows that domain-specific prompting and retrieval can improve practical analytical assistance [3]. Collectively, this literature suggests substantial potential but does not by itself provide an end-to-end architecture for campaign strategy, production, critique, and evaluation.

### C. LLM Agents and Collaborative Decomposition

LLM-agent research treats the language model as a component within a broader loop containing roles, state, tools, observations, and interactions. Building on the few-shot capabilities of large language models [12] and on prompting strategies that elicit intermediate reasoning [13], AutoGen provides a general infrastructure for multi-agent conversations and customizable interaction patterns [4]. CAMEL investigates role playing as a mechanism for autonomous cooperation between communicative agents [5], and Generative Agents combines memory, reflection, and planning to support coherent agent behavior over time [10]. More recent collaborative frameworks such as MetaGPT [14] and ChatDev [15] demonstrate that role-specialized agents exchanging structured intermediate artifacts can complete complex, multistage tasks. Together these systems support the broader proposition that role specialization and structured information exchange can make complex workflows more modular and inspectable, which is the property KAMF exploits for marketing campaign planning.

### D. Critique and Iterative Refinement

Initial model outputs are not necessarily optimal. Self-Refine uses iterative feedback [8], Reflexion incorporates linguistic feedback [7], and ReAct interleaves reasoning with action [6]. KAMF applies these principles to an integrated campaign: its Critic identifies a defect, severity, target agent, and correction instruction, enabling a limited, auditable revision cycle before evaluation.

### E. Design Science Positioning

The research follows a design science orientation in which knowledge is generated through the construction and evaluation of an artifact intended to address a relevant problem. Hevner et al. characterize design science as building and evaluating artifacts that extend organizational and human capabilities [11]. KAMF is therefore evaluated first as an implemented artifact: whether its architecture realizes the proposed workflow, supports controlled conditions, records outputs, and enables reproducible validation. Claims about real-world campaign effectiveness require a later empirical phase with expert or live-market outcomes. This work differs from general-purpose AutoGen, CAMEL, MetaGPT, and ChatDev in its marketing-specific artifact schemas, accumulated campaign state, cross-artifact issue routing, revision records, and common-condition quality rubric; these are system-integration contributions rather than novel general-purpose agent algorithms.

## III. Methodology

### A. Research Design

A design science approach was combined with expert assessment. The framework was specified against campaign-planning requirements, implemented as a web application, and assessed across three conditions using eight rubric dimensions. Twenty industry experts evaluated the generated artifacts. This study reports implemented mechanisms and descriptive expert judgments; actual conversions, acquisition costs, and other market outcomes were not observed.

### B. The KAMF Framework

KAMF receives a structured marketing brief and transforms it into a persistent campaign object. A master orchestrator controls stage execution and passes a shared campaign state between specialist agents. The implemented sequence is brief → strategy → audience → platform → content strategy → creative → budget → critic → revisions → evaluation → report. Each specialist contributes a bounded artifact rather than being responsible for the entire campaign, and Fig. 1 shows the resulting architecture and information flow. In operational terms, each stage consumes the brief and upstream artifacts; the orchestrator validates and persists its schema-bound output before allowing subsequent stages to proceed.

The framework is described as knowledge-augmented because no agent works in isolation: every agent is supplied with the accumulated campaign state: the brief, all prior specialist

outputs, and any critic feedback from previous revision rounds together with brand and context knowledge. The shared state is therefore the integration mechanism of the system. Strategy outputs establish goals, positioning, and KPIs that downstream agents can reference. Audience outputs provide segments and personas; platform planning maps audiences and objectives to channels; content and creative stages translate the strategic state into execution artifacts; and budget planning constrains allocations. Because the creative agent, for example, can see the strategy, personas, channels, and content plan, the workflow is collaborative rather than a set of independent parallel calls.

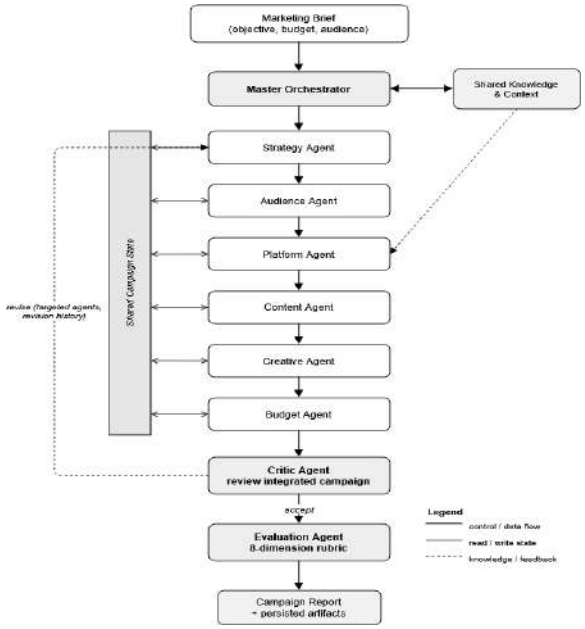



Fig. 1. KAMF system architecture and information flow.

The Critic records issues with severity, affected agent, and revision instructions; the Evaluation agent produces rubric scores, strengths, weaknesses, and recommendations. A dashboard displays briefs, progress, and campaign history; a REST application programming interface (API) exposes campaign, trace, experiment, and report operations; a generation service calls an external LLM when configured or a deterministic offline fallback; and a storage layer retains artifacts, critiques, revisions, and experiment results. Shared output schemas support both generation paths but do not imply equivalent output quality.

### C. Agent Role Decomposition

Each generator owns a defined marketing responsibility and returns validated fields that can be persisted, displayed, and reused. The design makes faults diagnosable: an unsuitable channel may be routed to Platform, whereas an unrealistic allocation may be routed to Budget, instead of rewriting the entire campaign.

### D. Agents Design

The framework is built from eight cooperating agents that together cover the full campaign planning workflow. Rather than implementing eight unrelated components, all agents share a single common design: each is an instance of the same agent abstraction and is differentiated only by its name, its system prompt and the structured output schema it must satisfy. This uniform construction keeps agent behavior consistent and the system simple to extend, and it ensures that every agent participates in the workflow in the same disciplined way: it receives the accumulated campaign state, produces a bounded and validated artifact and records trace information about its own execution.

Two properties of this design are central. First, the agents accumulate context: when an agent runs, its input is the original brief together with the outputs of all agents that executed before it and any critic feedback from earlier revision rounds. Because later agents can see the decisions made by earlier ones, the agents behave as a collaborating team rather than as independent parallel calls and the resulting campaign remains internally coherent. Second, each agent is schema-constrained: its output must conform to a predefined structure, which makes the result predictable, storable, individually inspectable, and easy to route for correction when the critic identifies a problem. Sequential execution shares updated context; downstream dependency revalidation remains important when an upstream artifact changes.

The eight agents fall into three functional groups. Six generator agents-Strategy, Audience, Platform, Content Strategy, Creative, and Budget— each own one coherent marketing responsibility and produce the substantive content of the campaign. A single Critic agent then reviews the integrated output of the generators and issues targeted, traceable revision instructions. Finally, an Evaluation agent scores the completed campaign against a structured multi-dimensional rubric. This division of labor is what makes the multi-agent claim substantive rather than cosmetic. Each of the eight agents is described below, in the order in which it contributes to the workflow.

#### *1) Strategy Agent*

The Strategy agent translates the brief into campaign objectives, SMART style goals, value proposition, positioning, strategic priorities, and measurable KPIs. Its output forms the strategic contract for downstream agents.

#### *2) Audience Agent*

The Audience agent develops audience segments and personas, including relevant needs, behavioral patterns, motivations, barriers, and customer journey considerations. These artifacts provide targeting context for channel and message decisions.

#### *3) Platform Agent*

The Platform agent recommends digital channels, proposes relative budget shares, and records the rationale connecting channel choice to audience behavior and campaign objectives. The output makes channel selection explicit rather than implicit in generated copy.

#### *4) Content Strategy Agent*

The Content Strategy agent defines content pillars, themes, formats, cadence or calendar logic, and posting strategy. It

connects the strategic and audience states to an executable content program.

*5) Creative Agent*

The Creative agent produces campaign-ready creative concepts and text artifacts such as ad copy, social posts, headlines, email concepts, blog ideas, calls to action, and hashtags, constrained by the preceding strategy and content plan.

*6) Budget Agent*

The Budget agent allocates campaign budget across recommended activities, records cost assumptions, and provides optimization or reallocation rules. Its purpose is to make resource constraints part of the generated plan.

*7) Critic Agent*

The Critic agent reviews the integrated campaign for contradictions, omissions, weak alignment, unrealistic assumptions, and other quality issues. It emits target-agent, severity, issue, and revision-instruction information so that correction is explicit and traceable.

*8) Evaluation Agent*

The Evaluation agent assesses the completed campaign across eight dimensions and summarizes strengths, weaknesses, and recommendations. The evaluation is designed for consistent comparison across conditions A, B, and C rather than as a substitute for real market performance.

### *E. Iterative Revision*

After initial specialist outputs are assembled, the Critic agent performs cross-component review. Critique is represented as actionable metadata rather than an unstructured comment: the issue is identified, its severity is recorded, the responsible or target agent is specified, and a revision instruction is produced. The orchestrator can then update only the relevant artifact and preserve a before-and-after revision history, up to a configurable maximum number of revision cycles. This design operationalizes feedback-oriented refinement while keeping the campaign state auditable. This selective revision reduces unnecessary regeneration in principle, but neither computational savings nor cost reductions have been measured. Critic verdicts are advisory and require constraint checks before release.

### *F. Experimental Conditions*

The system exposes three research conditions, illustrated in Fig. 2. Condition A (A_single) is a single-agent baseline that produces a consolidated campaign in one pass. Condition B (B_multi_no_critic) executes the specialist agents in sequence but omits critic revision. Condition C (C_multi_with_critic) executes the full specialist workflow followed by critique and targeted revision. Holding the campaign brief constant across conditions provides a basis for comparing output structure, rubric scores, and generation time when empirical runs are collected.

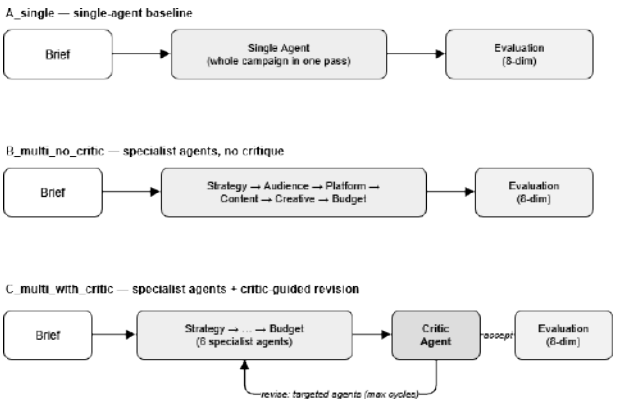



Fig. 2. Experimental conditions supported by KAMF.

### *G. Hypotheses Questions*

The following hypotheses connect the three experimental conditions to the expert evaluation reported in Section VI. H1: Role-decomposed multi-agent generation (Condition B) produces campaigns with higher completeness and consistency than the single-agent baseline (Condition A). H2: Adding critic-guided revision (Condition C) further improves consistency, practicality, and overall evaluation scores relative to Condition B. H3: The quality gains from decomposition and critique are obtained without a disproportionate increase in generation time. H1 and H2 are assessed descriptively against the expert ratings in Section VI, without inferential statistical testing; H3 is not empirically tested in the present study, since generation time was not logged, and is treated as a direction for future work rather than a tested hypothesis.

## IV. Framework Implementation

The framework is realized as an executable web system. This section summarizes the implementation of its core mechanisms, including how the main components are integrated and executed, as well as the technologies and tools used to build it. The focus is on the technical implementation, while the underlying concepts, architectural decisions, and overall design are described in the Methodology section.

### *A. Orchestration and Agent Execution*

A master orchestrator passes campaign state through ordered agents. At each stage, the executor supplies accumulated context, requests schema-constrained generation, validates the response, updates state, and records start and completion events. The implementation offers an imperative runner and a declarative graph whose conditional edge implements critique decisions. The model service supports configured external LLM generation or a deterministic offline fallback returning compatible schemas. Model name, provider version, temperature, top-p, token limits, seed, full prompts, and run environment are not documented in the provided study material; neither reproducibility nor equivalence of the two generation paths can be assumed without these records.

### *B. Critique, Evaluation, and Persistence*

Critic outputs a verdict and issues with severity and target-agent metadata. The orchestrator selectively revises affected artifacts up to a configured limit and stores before-and-after

history. Evaluation scores the campaign on the same eight dimensions used for condition comparison. A relational database persists the brief, agent outputs, trace events, critic reviews, revised versions, scores, and experiment results; reports export to Markdown, PDF, or DOCX. The ordered artifacts and issue routing establish explicit coordination rather than eight disconnected prompts.

### *C. Frontend and Backend Technologies*

The frontend is implemented with Next.js (App Router), TypeScript, and Tailwind CSS, communicating with the backend through a thin REST client. The backend is implemented in Python with FastAPI, SQLAlchemy, and Pydantic for the typed schemas that underpin schema-constrained generation, with SQLite in development and PostgreSQL for deployment. Fig. 3 and Fig. 4 show representative views of the interface: the dashboard, and the generated report with its workflow trace and export options.

## V. Evaluation Design

KAMF evaluates campaigns across eight dimensions: relevance, completeness, consistency, creativity, audience alignment, platform suitability, strategic quality, and practicality. Relevance measures fit to the original brief; completeness measures coverage of required campaign components; consistency measures cross-section coherence; creativity considers distinctiveness and variation; audience alignment measures fit to identified segments; platform suitability evaluates whether channel choices and content formats match platform characteristics; strategic quality measures goal positioning KPI coherence; and practicality considers feasibility, budget awareness, and executability. The rubric is summarized in Table I.

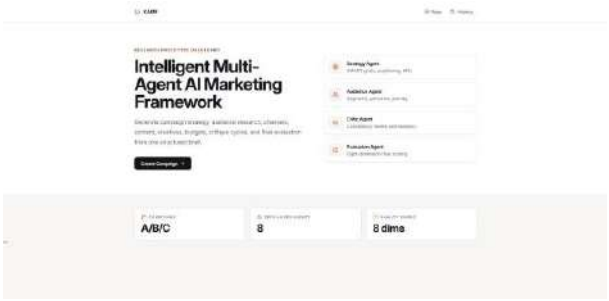



Fig. 3. KAMF web interface: Dashboard

Fig. 4. KAMF web interface: generated campaign report with workflow trace and export options.

Table 1
EIGHT-DIMENSION CAMPAIGN EVALUATION RUBRIC.

| No | Dimension | Primary question | Evidence focus |
|---|---|---|---|
| 1 | Relevance | Does the plan answer the brief? | Goals, offer, constraints |
| 2 | Completeness | Are required campaign components present? | Strategy through budget |
| 3 | Consistency | Do components agree with one another? | Cross-agent alignment |
| 4 | Creativity | Is the execution varied and non-generic? | Concepts and copy |
| 5 | Audience alignment | Does messaging fit segments and personas? | Audience-to-message fit |
| 6 | Platform suitability | Are channels and formats appropriate? | Channel rationale |
| 7 | Strategic quality | Are goals, positioning, and KPIs coherent? | Strategy logic |
| 8 | Practicality | Can the plan reasonably be executed? | Budget and actions |

Evaluation scores each dimension on a five-point scale and aggregates dimension scores. An experiment endpoint runs a common brief under A_single, B_multi_no_critic, and C_multi_with_critic. Automated scoring may share biases with campaign generation, so 20 industry experts also applied the rubric. A stronger replication would blind and randomize condition outputs, supply scoring anchors, and counterbalance presentation order; the supplied material does not document these controls.

## VI. Results and Discussion

To evaluate the framework, a panel of 20 domain experts in digital marketing independently assessed campaigns generated by KAMF. For each of a set of shared briefs, the experts rated the output of the three conditions single-agent (A), multi-agent without critique (B), and multi-agent with critic-guided revision (C) on the eight rubric dimensions using a five-point scale, and also judged the correctness of each individual agent's output. The reported figures are averaged across the 20 experts. The available manuscript contains aggregate comparisons rather than the evaluator-by-brief rating matrix, the number and identity of briefs, randomization records, or individual Evaluation-agent scores; these omissions limit independent statistical verification.

### *A. Expert Evaluation of Campaign Quality*

Figure 5 presents mean expert scores for eight dimensions. Both multi-agent conditions exceed the single-agent baseline; C receives the highest mean in every dimension, especially completeness, consistency, and strategic quality. Averages likewise increase from A to B to C. The patterns align descriptively with H1 and H2 but do not establish effect size, statistical significance, or generalization beyond the evaluated briefs and expert panel.

Statistical validation requires the paired rating matrix. Suitable analyses include within-brief paired comparisons with multiplicity correction and effect sizes, inter-rater agreement (e.g., intraclass correlation), and correlation between matched expert and Evaluation-agent scores. None can be computed

reliably from the reported means; no p-values, reliability coefficients, or calibration estimates are asserted.

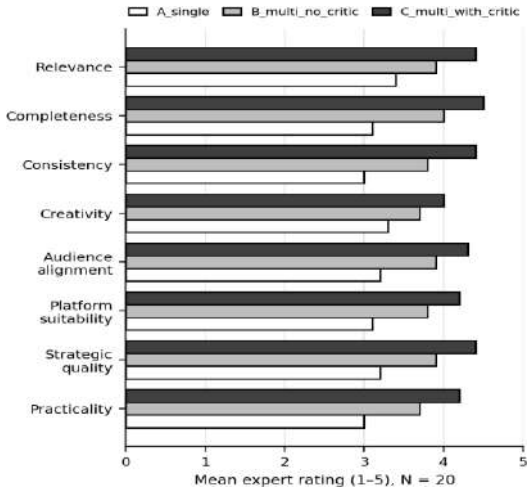



Fig. 5. Mean expert ratings by condition across the eight evaluation dimensions (N = 20).

### *B. Agent-Level Accuracy*

Because the framework decomposes campaign planning into specialist agents, the experts were also able to judge each agent's output separately. Table II reports the expert-rated accuracy of the eight agents: the proportion of outputs judged correct and fit for purpose together with the mean quality rating for each agent. Accuracy was high across all agents, with the Evaluation and Strategy agents rated most reliable and the Creative and Budget agents attracting the most expert-noted issues, typically generic phrasing or optimistic cost assumptions. This agent-level view is a direct benefit of decomposition: it localizes weaknesses to specific stages rather than to the campaign as a whole. The percentages in Table II are reported aggregate expert judgments; their denominators and agreement rules are not available, so they should not be interpreted as validated predictive accuracy against an independent ground truth.

TABLE II
EXPERT-RATED ACCURACY OF THE EIGHT AGENTS (N = 20).

| Agent | Accuracy (%) | Mean rating (1–5) |
|---|---|---|
| Strategy | 92 | 4.5 |
| Audience | 89 | 4.3 |
| Platform | 88 | 4.2 |
| Content Strategy | 90 | 4.3 |
| Creative | 86 | 4.0 |
| Budget | 85 | 4.0 |
| Critic | 91 | 4.4 |
| Evaluation | 93 | 4.5 |

### *C. Interpretation and Implications*

The pattern of results is consistent with the framework's design rationale. Explicit specialization lets each agent concentrate on one marketing responsibility, which the expert ratings associate with more complete and internally consistent campaigns, while shared state helps coordinate stages. The critic mechanism adds value by making quality review an explicit, targeted step: rather than regenerating a whole campaign, the system identifies which stage is weak and revises only that stage, and the experts rated the resulting campaigns highest. Regarding H3, the additional specialist and critic stages are expected to increase generation time relative to the single-agent baseline; the present study did not log generation time, so this trade-off is noted qualitatively rather than measured, and quantifying it is left for future work. The in-system Evaluation agent is not an independent judge of a system built using related prompts and models. Accordingly, expert scores are the primary reported evidence, while automated-score validity remains untested without paired expert and model ratings.

For practitioners, the findings indicate that generative AI can be embedded into a repeatable, inspectable planning process rather than used only for isolated copy generation, and that per-agent transparency helps reviewers trust and correct the output. Several caveats apply: the panel size is modest, ratings reflect expert judgement of plan quality rather than measured market performance, and the Creative and Budget agents show the clearest room for improvement. Generated strategies should therefore be treated as decision support, with human approval before budget or creative recommendations are acted upon.

## VII. LIMITATIONS

This study has two main limitations. First, the expert panel, although sufficient to indicate consistent trends, is modest in size, and the evaluation captures expert judgement of plan quality rather than measured live campaign performance. Second, generation quality depends on the selected large language model, prompt design, and sampling configuration, so results may vary across models and over time.

## VIII. CONCLUSION

This paper presented the Knowledge-Augmented Multi-Agent Framework (KAMF), a collaborative LLM-based system that transforms structured marketing briefs into coherent, inspectable digital marketing campaigns. The framework integrates eight specialized agents—Strategy, Audience, Platform, Content Strategy, Creative, Budget, Critic, and Evaluation-working through a shared campaign state. A critic-guided revision loop enables targeted improvements, while three experimental conditions assess the effects of agent decomposition and self-critique. Evaluation by 20 digital marketing industry experts showed positive ratings across eight quality dimensions, with multi-agent approaches outperforming the single-agent baseline and the critic-guided condition achieving the highest ratings. These findings demonstrate KAMF's potential for effective AI-assisted campaign planning and establish a foundation for future research involving larger expert panels and real-world campaign validation.

## REFERENCES

[1] D. Grewal, C. B. Satornino, T. H. Davenport, and A. Guha, "How generative AI is shaping the future of marketing," Journal of the Academy of Marketing Science, vol. 53, pp. 702–722, 2025, doi: 10.1007/s11747-024-01064-3.

[2] P. Cillo and G. Rubera, "Generative AI in innovation and marketing processes: A roadmap of research opportunities," Journal of the Academy of Marketing Science, 2024, doi: 10.1007/s11747-024-01044-7.

[3] Y. Gao, S. K. Arava, Y. Li, and J. W. Snyder Jr., "Improving the capabilities of large language model based marketing analytics copilots with semantic search and fine-tuning," arXiv:2404.13077,

2024.

[4] Q. Wu et al., “AutoGen: Enabling next-gen LLM applications via multi-agent conversation,” arXiv:2308.08155, 2023.

[5] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, “CAMEL: Communicative agents for ‘mind’ exploration of large scale language model society,” arXiv:2303.17760, 2023.

[6] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[7] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” arXiv:2303.11366, 2023.

[8] Madaan et al., “Self-Refine: Iterative refinement with self-feedback,” arXiv:2303.17651, 2023.

[9] N. Arora, I. Chakraborty, and Y. Nishimura, “Revolutionizing marketing research with a large language model: A hybrid AI-human approach,” SSRN Electronic Journal, 2024, doi: 10.2139/ssrn.4683054.

[10] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” in Proc. 36th Annual ACM Symp. User Interface Software and Technology (UIST), 2023, Art. no. 2, pp. 1–22, doi: 10.1145/3586183.3606763.

[11] R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[12] T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.

[13] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 24824–24837.

[14] S. Hong et al., “MetaGPT: Meta programming for a multi-agent collaborative framework,” in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[15] Qian et al., “ChatDev: Communicative agents for software development,” in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 15174–15186

# Enhanced YOLOv10 Framework for Real-Time Road Detection in Foggy and Complex Driving Environments

Sachin Chandanshive
*PG Student*
*Dept. of Electronics and Telecommunication*
*Alamuri Ratnamala Institute of Technology*
*Shahpur, District Thane, India*
sachinchandanshive0025@gmail.com

Niranjan Kulkarni
*Associate Professor and Head of Department*
*Computer Science and Design Engineering*
*New Horizon Institute of Technology and Management*
*Thane, India*
niranjankulkarni@nhitm.ac.in

***Abstract*—Detecting road objects in foggy and complex road scenes is core perception task for Advanced Driver assistance Systems (ADAS) [6], [7] and autonomous vehicles, However, detection accuracy rapidly degrades in fog and other low-visibility conditions [16], [23], [28], [29], where light scattering reduces image contrast and blurs object boundaries. This paper introduces YOLOv10-FCD, a YOLOv10-based detector for fog and complex driving scenes. Four components have been added to the NMS-free YOLOv10 architecture [2]: (i) a preprocessing stage that applies CLAHE contrast enhancement [21] followed by an adaptive image restoration module (AIRM) which is built on an atmospheric scattering model [17] and can be trained; (ii) a Coordinated Attention Backbone (CA-C2f) [11] that preserves fine-grained, location-sensitive features of small and partially occluded road users; (iii) a BiFPN branch [12] for weighted multi-scale fusion; and (iv) a VGG-19 v3 regression loss that limits the effect of ambiguous boxes in sequential images. The model is trained on foggy images from RESIDE-OTS and tested on the real-world RTTS benchmark with five classes (car, bus, person, motorbike and bicycle). It reaches 84.0% precision, 72.0% recall, 84.0% mAP@0.5 and 56.0% mAP @0.5:0.95, which is 1.7, 0.9, 4.9 and 1.7 percentage points above the YOLOv10 baseline, respectively. A step-wise ablation study shows how the CLAHE front end, AIRM, CA-C2f, and the BiFPN and Wise-IoU v3 pair add to these gains. Finally, a Flask-based prototype shows that the pipeline can process live camera streams and videos in practical use.**

***Keywords*—Real-Time object detection, YOLOv10, foggy scene understanding, adverse weather, ADAS, Coordinate Attention,**

## I. INTRODUCTION

Reliable real-time detection of road users and traffic infrastructure, such as vehicles, pedestrians and cyclists, is a basic requirement for Advanced Driver Assistance Systems (ADAS) [6], [7]. Ever since the original YOLO recast object detection as a single regression problem over the whole image [1], single-stage detectors have been the main choice for low-latency automotive perception, and later YOLO versions have kept improving the balance between accuracy and speed. The most recent of these, YOLOv10, removes Non-Maximum Suppression (NMS) at inference through consistent dual assignments: a one-to-many branch gives rich supervision during training, while a one-to-one branch produces the final boxes directly without NMS [2]. This reduction in latency makes YOLOv10 a promising foundation for edge platforms deployed on vehicles.

However, conventional convolution detectors, including YOLOv10, are primarily trained and tested on images acquired in clear weather conditions. Along the way, image quality is frequently degraded by fog, low light, rain, poor illumination, and occlusions. Smoke is the most challenging factor: small airborne particles scatter and diffuse reflected light, which reduces object contrast in the image, blurring object boundaries and reducing the perceived range [16], [23], [28], [29]. These factors remove the fine-scale details and boundary cues that image analysis systems (convolution feature extractors) rely on, making detection less reliable at a time when perception is critical for safety. Heavy traffic adds to the challenge, as the network must process fast-moving adjacent vehicles and distant small objects simultaneously, in a setting with poor object visibility and under strict real-time constraints.

Earlier work on visual perception in bad weather can be grouped into three directions: (a) separate dehazing or enhancement networks placed in front of the detector [3], [4], which improve visual quality but add latency and optimize an objective that is not tied to the detection loss; (b) domain-adaptation methods that transfer feature distributions from the clear domain to the hazy one [28], [29]; and (c) modifications inside the detector, such as attention blocks and multi-scale aggregation, that make features more tolerant to degradation [5]–[7]. Restoration and detection have recently been combined in single pipelines built on YOLOv7 [3] and on later YOLO versions [5]–[7]. To the best of our knowledge, however, applying this idea to the dual-assignment, NMS-free design of YOLOv10 has received little attention.

To address this gap, this paper develops and evaluates YOLOv10-FCD (YOLOv10 for Foggy and Complex Driving scenes). The main contribution is as follows:

**Architectural Integration:** We propose YOLOv10-FCD, a detection framework that adds an Adaptive Image Restoration Module (AIRM), Coordinate Attention backbone blocks (CA-

C2f), a Bidirectional Feature Pyramid Network (BiFPN) neck and a Wise-IoU v3 regression loss to the NMS-free YOLOv10 baseline.

**Targeted Robustness:** Each module is designed to provide a specific solution to a particular distortion. AIRM restores lost contrast and dynamic range, CA-C2f preserves directional positional information of fine or blurry structures, BiFPN combines features of objects at various scales, Wise -IoU v3 limits the contribution of boxes with poor localization due to ambiguous and uncertain annotation errors.

**Empirical Validation and Ablation:** We benchmark the model on a five-class hazy road dataset derived from the Realistic Task-driven Testing Set (RTTS). YOLOv10-FCD reaches 84.0% mAP@0.5 and 56.0% mAP@0.5:0.95, which is 4.9 percentage points higher in mAP@0.5 than baseline YOLOv10. A step-wise ablation study shows how much each added component contributes.

**Prototype Implementation:** We also implement a Flask-based prototype that processes live camera streams, images and video files, as a proof of concept of how the pipeline could be used in intelligent transportation applications.

The rest of the paper is organized as follows. Section II reviews related work on real-time detection, attention mechanisms and adverse-weather vision. Section III describes the proposed YOLOv10-FCD method. Section IV gives the experimental setup, and Section V presents the results, comparisons and ablation study. Section VI concludes the paper and outlines future work.

## II. RELATED WORK

### *A. Real-Time YOLO-Based Object Detection*

Single-stage detectors have driven much of the progress in real-time perception. YOLOv10 removes the NMS latency bottleneck by training with consistent dual label assignments: a one-to-many branch supplies dense supervision and a one-to-one branch gives a single prediction per object, so inference can run without NMS at little extra cost [2]. Networks such as Faster R-CNN [8] and transformer-based models, e.g. RT-DETR [9], offer high accuracy; however, their mathematical complexity places a heavy load on the processor, which affects performance and processing time.

### *B. Methods for feature learning and integration in multidimensional space*

Machine learning systems adapt feature representations based on the distribution of resolution within the image space and the orientation of objects; an example is CBAM [10]. The ‘Coordinate Attention’ method [11] splits the two-dimensional representation into two one-dimensional vectors, with long-range connections retained within one of them, whilst positions are transferred to the other. This network architecture is tailored to road images, which contain traffic lanes, road signs and infrastructure elements that are often arranged in parallel or perpendicular to one another. With regard to different types of roads, BiFPN within EfficientDet [12] utilizes an algorithm to identify vehicles on both sides of the road, meaning it is an improved version of both FPN and the PAN algorithm. For object classification, the Wise-IoU v3 model [13] utilizes dynamic and continuous learning, which allows for the identification of very low-quality objects without affecting the classification of other objects.

### *C. Road Object Detection with Enhanced YOLO Variants*

Several recent works adapt YOLO to transportation tasks. For instance, YOLOv8-PD uses large-kernel separable attention for road damage detection [14], and YOLO9tr uses partial self-attention for pavement damage detection [15]. These studies show that architectural changes help in transportation tasks, but they target surface inspection under good visibility and not driving perception in bad weather.

### *D. Object Detection Under Fog and Adverse Weather*

Dehazing has moved from physical priors to learned models. The classical Dark Channel Prior (DCP) [16] depends on hand-crafted assumptions, whereas AOD-Net [17] re-parameterizes the atmospheric scattering model so that a single end-to-end convolution network can perform dehazing. Building on this, several studies place restoration inside the detection pipeline. IDOD-YOLOv7 [3] pairs an AOD-Net-based defogging module with YOLOv7, PED-YOLO [4] balances enhancement and detection losses, and AWD-YOLO [5] uses dual-backbone attention fusion. Other recent frameworks, namely YOLO-Vehicle-Pro [6], ECL-YOLOv11 [7] and ESA-YOLO [18], likewise indicate that enhancing features inside the network avoids the degradation typical of disconnected preprocessing. YOLOv10-FCD follows this idea on the dual-assignment YOLOv10 architecture and combines restoration, feature representation, multi-scale fusion and box regression in one model.

### *E. Lightweight Backbone and Attention Alternatives*

Lightweight modules such as GhostNet [19] and ECA-Net [20] are parameter-efficient alternatives. Here Coordinate Attention is chosen instead because it encodes position explicitly, which matters for elongated structures such as curbs and road edges in fog; a direct comparison with these alternatives is left for future work.

## III. METHODOLOGY

Fig. 1 gives an overview of the YOLOv10-FCD pipeline. An incoming foggy road image first goes through Contrast Limited Adaptive Histogram Equalization (CLAHE) applied to the luminance channel in the LAB color space, which stabilizes the local dynamic range [21]. The result is then passed to the Adaptive Image Restoration Module (AIRM) for deeper contrast recovery [17]. Features are extracted by a Coordinate Attention-enhanced CSPNet backbone [11], fused across scales by the BiFPN neck [12], and finally decoded by the YOLOv10 dual detection heads, which are trained with the Wise-IoU v3 regression loss [13].

YOLOv10 consists of a CSPNet-based backbone, a PAN-style aggregation neck and decoupled dual heads. In training, the one-to-many head provides dense label supervision, which speeds up convergence and improves recall. [2]

### *A. Baseline YOLOv10 Architecture*

In the detection process, a ‘one-to-one’ detection head directly generates a single bounding box for each object;

consequently, there is no need for an NMS algorithm, and the execution time is reduced [2]. As a base model, we use YOLOv10m and initialize it using pre-trained weights to speed up convergence.

*B. Adaptive Image Restoration Module (AIRM)*

As CLAHE is not derivable, it remains outside the trainable layers and is used as a fixed pre-processing step, ensuring that the gradient path in the network remains intact. Each damaged frame is converted to the LAB color space, and CLAHE [21] is applied to the luminance (L) channel to reduce global lighting contrast and stabilize the local dynamic range.

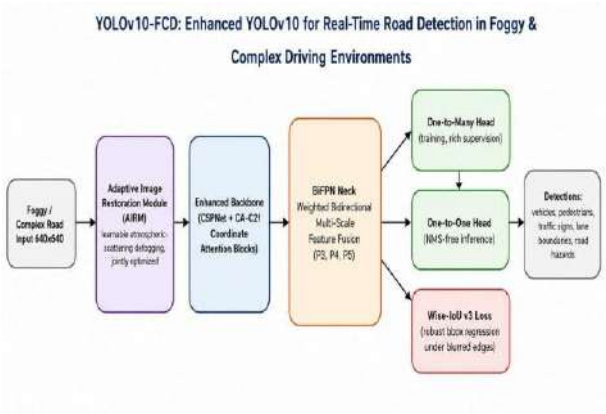



Fig. 1. Overall architecture of the proposed YOLOv10-FCD framework, showing AIRM, CA-C2f backbone, BiFPN neck, and dual-head loss.

Following this fixed stage, the image is converted back to RGB and passed to the trainable layers in AIRM. This stage is based on the atmospheric scattering model, defined as:

$$I(x) = J(x) \cdot t(x) + A \cdot (1 - t(x))$$

Where I(x) is the observed blurred image, J(x) is the true scene radiance, t(x) is the medium transmission map, and A is the global atmospheric light. Following AOD-Net [17], Eq. (1) is re-parameterized so that t(x) and A no longer need to be estimated separately, which gives:

$$J(x) = K(x) \cdot I(x) - K(x) + b$$

where b = 1 is a constant bias term and K(x) is a unified parameter map, defined as:

$$K(x) = \frac{\frac{1}{t(x)} \cdot (I(x) - A) + (A - 1)}{I(x) - 1}$$

In the AIRM model, K(x) is composed of five convolution layers, with kernel sizes of 1×1, 3×3, 5×5, 7×7 and 3×3, using ReLU activations and skip connections. The module prevents the model from converging to poor outcomes. to prevent it from propagating to subsequent layers, which would mislead the model and produce incorrect results. Instead of using a simple sum, a combined loss function is employed during training, which relies on the signals of the final cell rather than on the individual cell outputs. Total loss, $L_{\text{total}}$ is a multi-objective function:

$$L_{\text{total}} = L_{\text{det}} + L_{\text{restoration}}$$

where $L_{det}$ is the standard YOLOv10 dual-assignment detection loss, made up of bounding-box regression, distribution focal loss [22] and classification terms, and $L_{restoration}$ is a pixel-wise Smooth $L_1$ loss between the restored image J(x) and the matching clear-weather ground-truth image from the RESIDE-OTS training set [23]. The weight lambda is empirically set to 0.1 to preserve structural and photometric fidelity during back-propagation. At inference on the unseen RTTS data, the restoration branch of the loss is dropped, so the pipeline keeps its low latency.

*C. Coordinate Attention-Enhanced Backbone (CA-C2f)*

Atmospheric scattering weakens fine structural detail. To keep long-range context without losing precise localization, Coordinate Attention blocks [11] are inserted into the C2f blocks of the backbone at stages P3, P4 and P5. For an input feature map X, global pooling is split into two 1D aggregations along the horizontal and vertical directions:

$$z_c^{h(h)} = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i)$$

$$z_c^{w(w)} = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w)$$

The feature maps that recognize both directions undergo standard 1×1 convolutions, batch normalization, and nonlinear activation to form intermediate representations. These are organized along two spatial axes and pass-through sigmoid functions to produce the directional attention maps $g^h$ and $g^w$, which multiply the input features:

$$y_c(i, j) = x_c(i, j) \cdot g_c^h(i) \cdot g_c^w(j)$$

This gives the detector more sensitivity to long, narrow structures such as curbs and road edges under low contrast.

*D. BiFPN Multi-Scale Feature Fusion Neck*

Street scenes show a wide range of scales, from vehicles in the foreground to pedestrians and small objects in the distance. In YOLOv10-FCD, the standard 'PAN neck' is replaced by a Bidirectional Feature Pyramid Network (BiFPN) [12]. The BiFPN removes nodes that have only one input connection to save computational resources, adds residual side connections between the input and output nodes, and uses fast normalized fusion:

$$O = \frac{\sum_j w_i \cdot I_i}{epsilon + \sum_j w_j}$$

which is used along with ReLU for wi ≥ 0 and epsilon = 0.0001 to maintain numerical stability in the division. This allows the network to learn the importance of each feature level at stages 8, 16, and 32, and consequently preserves strong feature responses for small and occluded objects. To balance representation power with computational efficiency, the BiFPN block is repeated across (N=3) sequential layers, maintaining a uniform channel width of (C=256) across all intermediate feature scales.

*E. Wise-IoU v3 Regression Loss*

In foggy scenes object edges are blurred, which puts noise into both manual annotations and predicted boxes. The usual CIoU loss penalizes all samples in the same way and can over-penalize noisy labels. YOLOv10-FCD therefore uses Wise-IoU v3 [13], which has a dynamic, non-monotonic focusing mechanism:

$$L_{\mathrm{WIoUv1}} = R_{\mathrm{WIoU}} \cdot L_{\mathrm{IoU}}$$

$$R_{\mathrm{WIoU}} = \exp\left(\frac{(x-x_{\mathrm{gt}})^2+(y-y_{\mathrm{gt}})^2}{(W_g^{\,2}+H_g^{\,2})^{*}}\right)$$

where (x, y) and ($x_{gt}$, $y_{gt}$) are the centers of the predicted and ground-truth boxes, and $W_g$ and $H_g$ are the width and height of the smallest box enclosing both; the superscript * means the term is detached from gradient computation. The outlier degree IoU v3 keeps ambiguous box edges from destabilizing the convergence of the dual-head loss. [To be completed by the authors: state whether Wise-IoU v3 replaces CIoU in both the one-to-many and the one-to-one branch.]

## IV. EXPERIMENTAL SETUP

*A. Dataset Specifications*

To avoid data contamination and to test cross-domain generalization, separate datasets are used for training and evaluation. The network is trained only on the RESIDE open-world training set (OTS). [23] is adapted, which is a large dataset of artificially cluttered images as well as clear ground-truth images. Evaluation is conducted on a realistic task-based test set (RTTS). [23] has been conducted, which consists of 4,323 real foggy images, containing 39,435 labeled instances: 29,190 cars, 2,752 buses, 4,520 people, 2,147 motorcycles, and 826 bicycles. RTTS is used solely as an unseen real-world test benchmark, demonstrating the reported metrics' ability to generalize to the domain without any risk of training data leakage.

*B. Implementation Details*

TABLE I. IMPLEMENTTAION DETAIL

| Category | Component / Parameter | Configuration Details |
|---|---|---|
| Hardware | GPU | NVIDIA GeForce RTX 4090 (24 GB VRAM) |
| Hardware | CPU | AMD Ryzen 9 7950X |

to 0.95 in steps of 0.05 (mAP@0.5:0.95), computed as:

$$\mathrm{mAP@0.5{:}\,0.95} = \frac{1}{10}\sum_{tau=0.50}^{0.95} \mathrm{mAP}_{tau}$$

## V. RESULTS AND DISCUSSION

*A. Quantitative Comparison with YOLO Models*

To assess YOLOv10-FCD, it was compared with several generations of the YOLO family: YOLOv5 [24], YOLOv7 [25], YOLOv8 [26], YOLOv9 [27] and the baseline YOLOv10 [2]. Table I summarizes the results.

beta and the non-monotonic focusing coefficient r are defined as:

$$beta = \frac{L_{\mathrm{IoU}}^{*}}{\overline{L_{\mathrm{IoU}}}}$$

$$r = \frac{beta}{deltaalpha^{beta-delta}}$$

$$L_{\mathrm{WIoUv3}} = r \cdot L_{\mathrm{WIoUv1}}$$

where $L^{*}_{IoU}$ is the detached IoU loss of the current box and the bar denotes its running mean. The hyper parameters are alpha = 1.9 and delta = 3, as recommended in [13]. By giving smaller gradient weights to extreme outliers, Wise-

TABLE II. QUANTITATIVE COMPARISON WITH YOLO MODELS

| Category | Component / Parameter | Configuration Details |
|---|---|---|
| Hardware | RAM | 64 GB System RAM |
| Software Stack | Framework | Ultralytics YOLO Framework |
| Software Stack | Core Library | PyTorch 2.1.2 |
| Software Stack | Computer Platform | CUDA 12.1 |
| Training Settings | Epochs | 100 epochs |
| Training Settings | Resolution | 640x640 pixels |
| Training Settings | Batch Size | 32 |
| Optimization | Optimizer | Stochastic Gradient Descent (SGD) |
| Optimization | Weight Decay | 0.0005 |
| Learning Rate | Initial LR | 0.01 |

*C. Evaluation Metrics*

Detection performance is measured with standard metrics: macro-averaged precision (P) and recall (R), mean Average Precision at an IoU threshold of 0.5 (mAP@0.5), and mAP averaged over IoU thresholds from 0.50

TABLE III. PERFORMANCE COMPARISON WITH EXISTING YOLO MODELS ON DATASET

| Model | P (%) | R (%) | mAP@0.5 | mAP@0.5:0.95 | Params (M) | Latency (ms) |
|---|---|---|---|---|---|---|
| YOLOv5 | 75.4 | 64.2 | 70.5 | 44.8 | 21.2 | 4.8 |
| YOLOv7 | 78.3 | 67.5 | 74.6 | 48.9 | 36.5 | 6.2 |
| YOLOv8 | 80.1 | 69.8 | 76.8 | 51.2 | 25.9 | 5.4 |
| YOLOv9 | 81.5 | 70.4 | 78.2 | 53.5 | 51.0 | 7.9 |
| YOLOv10 Base | 82.3 | 71.1 | 79.1 | 54.3 | 16.4 | 4.1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **YOLOv10-FCD** | **84.0** | **72.0** | **84.0** | **56.0** | **18.1** | **4.8** |

As Table I shows, YOLOv10-FCD obtains the best accuracy among the compared models on all four detection metrics, reaching 84.0% precision, 72.0% recall, 84.0% mAP@0.5 and 56.0% mAP@0.5:0.95. This comes with a moderate increase in model size (18.1 M versus 16.4 M parameters) and latency (4.8 ms versus 4.1 ms) over the YOLOv10 baseline. Table II lists the absolute and relative gains over the baseline.

TABLE IV. DETAILED IMPROVEMENT ANALYSIS OVER BASELINE YOLOV10

| Metric | Baseline | Proposed | Abs. Gain | Rel. Gain |
|---|---|---|---|---|
| Precision | 82.3% | 84.0% | **+1.7%** | **+2.07%** |
| Recall | 71.1% | 72.0% | **+0.9%** | **+1.27%** |
| mAP@0.5 | 79.1% | 84.0% | **+4.9%** | **+6.20%** |
| mAP@0.5:0.95 | 54.3% | 56.0% | **+1.7%** | **+3.13%** |

The largest gain is in mAP@0.5, which rises by 4.9 percentage points (6.20% relative), showing better detection of hazy road objects at the 0.5 IoU threshold. The smaller gain of 1.7 points in mAP@0.5:0.95 suggests that stricter box alignment also improves, although the ablation in Section V-C does not separate the effect of Wise-IoU v3 from that of the BiFPN neck.

### *B. Convergence and Class-Level Analysis*

The curves in Fig. 2 show stable convergence over the 100 epochs for both the one-to-many and the one-to-one branch: the box, classification and distribution focal losses decrease steadily without divergence. The class-wise recall values read from the confusion matrix (Fig. 3) are 0.85 for car, 0.86 for bus, 0.79 for person, 0.74 for motorbike and 0.73 for bicycle. These values have been calculated at a specific confidence level; for that reason, they are not necessarily the same as the macro-average recall indicated in Table I. The lower values for two-wheeled vehicles are probably due to their narrow wheels and handlebar structure, which are much more prone to scattering, and their small share in the dataset (826 bicycles versus 29,190 car examples in RTTS) is also a factor.

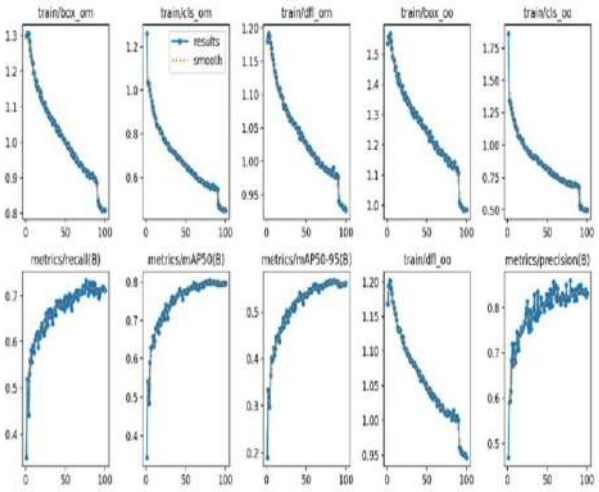



Fig. 2. Training losses and validation metrics (recall, mAP@0.5, mAP@0.5:0.95, precision) over 100 epochs.

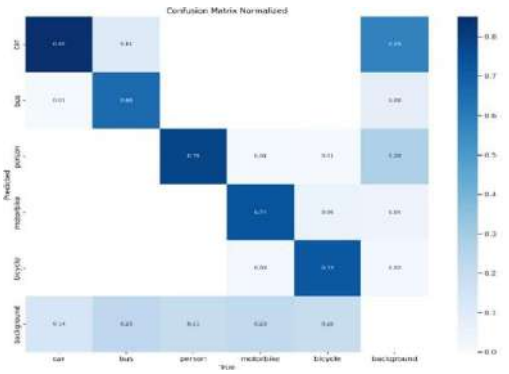



Fig. 3. Normalized confusion matrix for the five road object classes on the evaluation set.

### *C. Systematic Ablation Study*

To measure the contribution of each added component, a five-step ablation was run with the same hyperparameters, random seed and data. Table III gives the cumulative results.

TABLE V. STEP-BY-STEP ARCHITECTURAL ABLATION STUDY

| Configuration / Model | CLAHE | AIRM | CA-C2f | BiFPN + WIoU v3 | mAP@0.5 (%) | mAP@0.5:0.95 (%) |
|---|---|---|---|---|---|---|
| A. YOLOv10 baseline | No | No | No | No | 79.1 | 54.3 |
| B. Baseline + CLAHE | Yes | No | No | No | 80.4 | 54.9 |
| C. Baseline + CLAHE + AIRM | Yes | Yes | No | No | 81.9 | 55.2 |
| D. Baseline + CLAHE + AIRM + CA-C2f | Yes | Yes | Yes | No | 82.8 | 55.5 |
| **E. YOLOv10-FCD** | **Yes** | **Yes** | **Yes** | **Yes** | **84.0** | **56.0** |

Adding the CLAHE front end (Config B) raises mAP@0.5 by 1.3 points. Adding the trainable AIRM (Config C) gives a further 1.5 points, which suggests that restoration trained together with the detector helps more than fixed contrast adjustment alone. Coordinate Attention (Config D) lifts mAP@0.5 to 82.8%, which we attribute to better retention of positional features. Finally, by combining the BiFPN base with Wise-IoU v3 (Configuration E), the model achieves 84.0% mAP@0.5 and 56.0%mAP@0.5:0.95.

### *D. Prototype Implementation*

As a proof of concept, the pipeline was wrapped in an interactive Flask-based application that accepts live webcam streams, video files and still images and runs them through the CLAHE and AIRM front end and the YOLOv10-FCD detector. In informal visual checks on foggy road scenes, the application detected small vehicles, partly occluded pedestrians and distant motorbikes. Its speed on embedded hardware has not been measured yet, so suitability for edge deployment still has to be verified.

## VI. CONCLUSION AND FUTURE WORK

This paper presented YOLOv10-FCD, a real-time road object detector designed for fog-degraded driving scenes. The model combines an Adaptive Image Restoration Module (AIRM) with a CLAHE front end, Coordinate Attention backbone blocks (CA-C2f), a Bidirectional Feature Pyramid Network (BiFPN) neck and the Wise-IoU v3 loss inside the NMS-free YOLOv10 architecture, and it addresses contrast loss, positional ambiguity, scale variation and boundary uncertainty. On the RTTS hazy road benchmark it achieved 84.0% precision, 72.0% recall, 84.0% mAP@0.5 and 56.0% mAP@0.5:0.95, which is 4.9 percentage points higher in mAP@0.5 than baseline YOLOv10 and higher than the other YOLO versions compared. The study has limitations: it uses a single fog benchmark with five classes and single training runs, and the effects of BiFPN and Wise-IoU v3 are not separated. Future work will test the model on further adverse-condition datasets such as Foggy Cityscapes [28], Foggy Zurich [29], BDD100K [30] and ACDC [31], add fog-specific detectors as baselines, ablate each module on its own, measure speed on embedded hardware with TensorRT quantization, and explore LiDAR and radar fusion for low-power automotive electronic control units (ECUs).

## REFERENCES


[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[2] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[3] Y. Qiu, Y. Lu, Y. Wang, and H. Jiang, "IDOD-YOLOV7: Image-dehazing YOLOV7 for object detection in low-light foggy traffic environments," *Sensors*, vol. 23, no. 3, Art. no. 1347, 2023.

[4] Z. Li, J. Wu, S. Lin, Z. Wang, X. Jin, G. Geng, F. Huang, and J. Weng, "Progressive enhancement dehazing for object detection in extreme weather," *Expert Syst. Appl.*, vol. 260, Art. no. 125345, 2025.

[5] Y. Yuan and W. Dong, "AWD-YOLO: Enhancing autonomous driving perception reliability in adverse weather," *Sci. Rep.*, vol. 16, Art. no. 5421, 2026.

[6] X. Li, J. Chen, Y. Sun, N. Lin, A. Hawbani, and L. Zhao, "YOLO-Vehicle-Pro: A cloud-edge collaborative framework for object detection in autonomous driving under adverse weather conditions," *arXiv preprint arXiv:2410.17734*, 2024.

[7] Z. Liu, J. Zhang, X. Zhang, and H. Song, "Robust object detection in adverse weather conditions: ECL-YOLOv11 for automotive vision systems," *Sensors*, vol. 26, no. 1, Art. no. 304, 2026.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

[9] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16965–16974.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[11] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13713–13722.

[12] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10781–10790.

[13] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," *arXiv preprint arXiv:2301.10051*, 2023.

[14] J. Zeng and H. Zhong, "YOLOv8-PD: An improved road damage detection algorithm based on YOLOv8n model," *Sci. Rep.*, vol. 14, Art. no. 12052, 2024.

[15] S. Youwai, A. Chaiyaphat, and P. Chaipetch, "YOLO9tr: A lightweight model for pavement damage detection utilizing a generalized efficient layer aggregation network and attention mechanism," *J. Real-Time Image Process.*, vol. 21, Art. no. 163, 2024.

[16] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, 2011.

[17] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4780–4788.

[18] C. Li, S. Liu, and Z. Peng, "ESA-YOLO: An efficient scale-aware traffic sign detection algorithm based on YOLOv11 under adverse weather conditions," *PLOS ONE*, vol. 20, no. 2, Art. no. e0336863, 2025.

[19] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1580–1589.

[20] Q. Wang, B. Wu, P. Zhu, P. Zhang, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11534–11542.

[21] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[22] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 21002–21012.

[23] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, 2019.

[24] G. Jocher, "YOLOv5 by Ultralytics," Zenodo, 2020, doi: 10.5281/zenodo.3908559.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7464–7475.

[26] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub repository, 2023.

[27] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[28] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, pp. 973–992, 2018.

[29] C. Sakaridis, D. Dai, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 687–707.

[30] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2636–2645.

[31] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10765–10775.

# Coordinating Verification and Generation in Multi-Agent Systems: A Real-Time, Evidence-Grounded Lecture Note Assistant

Mihidukulasuriya Nethmi Ruasha Fernando
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwaththa, Pitipana, Homagama, Sri Lanka*
fernandomnr@students.nsbm.ac.lk

Thisarani Wickramasinghe
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwaththa, Pitipana, Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

***Abstract*—Manual lecture note-taking forces students to comprehend spoken content and record it at the same time, a dual demand that can compete with learning itself. Existing AI-assisted approaches address parts of this problem, but rarely combine incremental generation, multimodal fusion, in-lecture student control, and claim-level evidence verification in one real-time system and rarely examine how the cooperating components should coordinate under live latency constraints. This paper presents a real-time learning assistant built as A Unified On-Device Classifier for NSFW and Malicious Web Pages Usint system: a Session Coordinator and six specialized agents (Capture, Notes, Evidence, Learning, Q&A, Quiz) that exchange information exclusively through a shared, typed lecture state and an event-driven orchestration graph. The coordination protocol, comprising a fast/slow path split, write-isolated fields, a fail-closed verification gate with bounded regeneration, and a fixed-order consolidation pass, is detailed and evaluated on a ten-lecture technical corpus and with 25 undergraduate participants. Pooled transcription word error rate was 8.46% with a real-time factor below 1.0 in every lecture, and the median lecture-level note-generation latency was 2,985 ms. Fusing slide and document content with speech raised source-fact coverage from 84.69% to 92.34%, higher in all ten lectures. Enabling the Evidence Agent's verification gate reduced the unsupported-claim rate from 21.58% to 8.55%, again in every lecture. Against a single-call baseline, the multi-agent pipeline produced fewer short, dangling, and duplicate bullets at about three times the processing time. Participants rated the system 4.20 to 4.60 out of 5 across usability, trust, and revision usefulness. All findings are reported with their statistical treatment and explicit threats to validity.**

***Keywords—multi-agent systems, real-time speech recognition, human–AI collaboration, multimodal learning, lecture note generation***

## I. Introduction

In a traditional lecture setting, students rely on handwritten or typed notes taken while the lecturer speaks, a practice that has changed little despite decades of technological progress in the classroom. This manual process requires a student to listen, comprehend, and transcribe at the same time, and it is this constraint that recent AI-assisted approaches attempt to relieve. The rapid development of large language models (LLMs) and multimodal foundation models is changing how information can be captured, processed, and presented to learners [1], [2]. Automated transcription can reduce the recording burden, and LLMs can restructure transcribed material into concise notes, but a genuinely useful assistant must also operate within the temporal constraints of a live lecture, fuse multiple information sources, let students intervene in generated content, and expose enough provenance for a learner to judge whether a generated statement can be trusted.

No single model call can satisfy all of these requirements at once without becoming an unwieldy, hard-to-audit prompt simultaneously responsible for transcript structuring, judging its own factual grounding, resolving conflicting sources, and deciding what a student is allowed to see. This is a coordination problem as much as a modelling problem: the system needs several specialized components that can operate concurrently, disagree with each other, and recover from one another's failures while a lecture keeps moving. Multi-agent decomposition is therefore the architectural mechanism through which the system implements independent verification, bounded-latency generation, and student ownership. A non-agentic single-call baseline is compared on five additional lectures (Section V); because that baseline also lacks regeneration and consolidation, the comparison measures complete configurations and does not isolate agent roles.

Prior work addresses these requirements individually rather than jointly, and rarely addresses the coordination question at all (Section II). To the best of current knowledge, the reviewed literature does not report real-time incremental generation, multimodal fusion, in-lecture human control, and claim-level evidence verification operating together under live-lecture latency constraints. This paper presents such a system, evaluated against five research questions covering transcription accuracy (RQ1), incremental generation (RQ2), multimodal coverage (RQ3), verification-gate effectiveness (RQ4), and student perception (RQ5). The contributions are:

1) a collaborative multi-agent architecture in which six specialized agents and a non-generating Session Coordinator communicate exclusively through a shared, typed state and an event-driven orchestration graph;
2) an explicit coordination protocol (a fast/slow path split, per-agent write isolation, and a lock check re-evaluated on every verification entry) that lets capture, generation, and

verification proceed concurrently without one agent's latency blocking another, and without an autonomous agent overwriting a student's edit;

3) an Evidence Agent implementing deterministic, fail-closed re-confirmation of quoted evidence that peer-reviews another agent's output without rewriting it;

4) an empirical demonstration that a bounded-regeneration verification gate reduces the unsupported-claim rate of a live multi-agent pipeline (21.58% to 8.55%), with retry-bound sensitivity analysis; and

5) a multi-dimensional evaluation spanning transcription, latency, multimodal coverage, evidence verification, a non-agentic baseline, and usability, with statistical procedures matched to each comparison.

## II. Related Work

### *A. Real-Time Classroom Speech Recognition and Lecture Summarization*

Early classroom automatic speech recognition (ASR) work demonstrated lecturer-adapted real-time transcription [3], and Whisper established strong general-purpose weakly supervised speech recognition [4]. LLM-based summarization improves the post-lecture study experience [5], and NexaNota reports strong topic identification and linking for structured LLM lecture notes [6]. These establish that live transcription and LLM-based note structuring are each feasible, but neither is built as a pipeline where generation and verification run concurrently over shared state.

### *B. Multimodal Educational AI*

Multimodal LLMs are increasingly applied across language learning, STEM instruction, and interactive content [7]. This establishes multimodal processing as viable, but does not demonstrate continuous fusion of live speech with slide/document evidence while preserving which modality a claim came from, which this system resolves via a dedicated agent applying an explicit priority policy.

### *C. Human–AI Collaboration in Education*

Human–AI collaboration research shows assistance does not uniformly improve outcomes: an educational tagging task saw AI assistance speed up work while reducing accuracy [8], and naive AI explanations can increase overreliance [9]. No prior work reports editing AI-generated notes during the lecture with edits persisting against later updates, which here requires every autonomous agent to check a lock state before writing.

### *D. Hallucination Mitigation and Verification*

Self-Refine and Chain-of-Verification show iterative self-checking can reduce hallucination [10], [11]; SelfCheckGPT uses black-box consistency sampling [12]. Retrieval-augmented generation and its self-reflective variants ground generation in retrieved evidence [13], [14], and recent work applies deterministic neuro-symbolic verification gates in reasoning pipelines [15]. Almost all this treats verification as a property of a single pipeline's output rather than one agent judging another's work; this setting adds the question of what a generating agent should do when a peer rejects its output.

### *E. Multi-Agent LLM Frameworks and Coordination Patterns*

AutoGen, MetaGPT, and CAMEL establish general patterns for multi-agent LLM orchestration [16]–[18], coordinating agents through conversational message-passing where a controller decides who speaks next. This system instead coordinates through a shared typed state with per-agent write isolation and an explicit graph of conditional transitions: agents never address each other directly, and authority is defined by which fields an agent may write, which is easier to audit for a safety property (a verifier that cannot rewrite what it verifies) than open-ended dialogue.

### *F. Explainable AI in Education, and Positioning*

Educational explainable AI (XAI) research establishes the importance of understandable, trustworthy AI systems [19], but the technical explanations it targets do not directly answer a learner's practical question: what evidence supports this specific note? This system provides claim-level evidence traceability as a direct byproduct of the Evidence Agent's coordination role. No individual capability surveyed above is claimed as novel in isolation; the contribution is the coordination mechanism that makes the combination workable in real time. Relative to conversational orchestration [16]–[18], the coordination medium is typed shared state with per-field write ownership rather than dialogue. Relative to RAG [13], [14], retrieval serves an independent verifier rather than the generator. Relative to self-refinement and self-verification [10]–[12], the critic is a separate agent that cannot rewrite the note, and acceptance depends on deterministic re-confirmation of a quoted span rather than on the model's self-assessment. Relative to standalone verification gates [15], the gate is coupled to a bounded regeneration protocol and to student lock precedence inside a live pipeline. The Evidence Agent's re-confirmation step is a secondary, narrower contribution, not a claim to be the first verification gate in the literature.

## III. Collaborative Multi-Agent Architecture

The system follows a Design Science Research approach [20]: an artefact is designed, built, and evaluated against explicit requirements derived from the gaps identified in Section II. This section describes not only what each agent does, but how the agents are coordinated.

### *A. Agent Overview and Rationale for Decomposition*

A non-generating Session Coordinator owns session lifecycle and routes events without producing content. The Capture Agent turns raw audio, slides, and documents into structured, source-tagged input via ASR and vision extraction. The Notes Agent clusters accumulated input into a hierarchical topic/subtopic/bullet structure. The Evidence Agent independently checks each note against source material and issues a verified/regenerate/unverified decision without rewriting the note text. The Learning Agent deduplicates and consolidates notes and resolves cross-source conflicts. The Q&A Agent answers questions grounded in verified content, and the Quiz Agent generates provenance-aware multiple-choice questions from the same content. Whisper large-v3 performs ASR; Llama 3.1 70B performs structuring, evidence verification, and deduplication; Llama 3.3 70B serves

interactive Q&A and Quiz; and Llama 4 Scout 17B performs vision extraction.

An earlier prototype used a single large prompt for transcript structuring, evidence assessment, and formatting together, and exhibited instruction loss during development: as the prompt grew, the model intermittently dropped formatting constraints or skipped the evidence-check instruction. This qualitative observation motivated decomposing the system into agents with narrow, independently testable responsibilities, which also makes it possible to disable one responsibility (e.g. the verification gate) without touching the others.

### *B. Shared State and Orchestration Graph*

Rather than passing messages directly between agents, every agent reads from and writes to a single typed shared lecture state (seven structured tables). The execution graph is a LangGraph StateGraph with seven role-nodes (Fig. 1), connected almost entirely by one-directional edges, with exactly one cyclic edge: Evidence back to Notes for regeneration. Because agents communicate only through defined state fields, each agent's contract can be tested and audited independently. Fig. 2 shows the control flow: a gate decision passes the note downstream or routes it back to Notes for bounded regeneration. The retry bound $N$ is enforced by a deterministic counter in application state, so loop termination is verifiable from that counter alone.

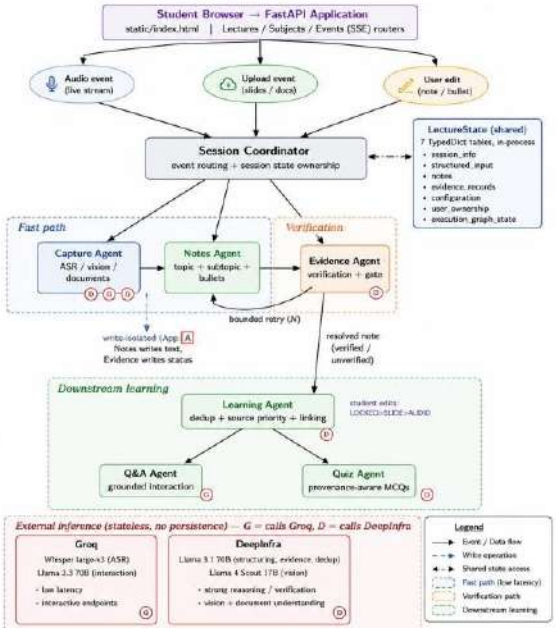



Fig. 1. Overall system architecture: agents exchange information only through the shared lecture state. G/D marks the inference provider (Groq/DeepInfra) each agent calls.

## A. Algorithmic Composition of Each Agent

The Session Coordinator contains no LLM calls at all, and every agent that does call an LLM pairs it with a deterministic check that does not trust the model's output at face value.

TABLE I. Algorithmic Composition of Each Agent

| Agent | Algorithmic composition |
|---|---|
| Capture | I/O only: Whisper large-v3 ASR, zero-shot vision-LLM extraction. |
| Session Coordinator | Classical CS: two-pointer buffer scan, exact-match topic registry, first-fit bin-packing, bounded producer/consumer queue. |
| Notes | LLM-as-structured-parser (rule-constrained JSON) for topic/subtopic/bullet clustering, plus deterministic leak detection. |
| Evidence | Five-stage hybrid: lexical retrieval (difflib. SequenceMatcher), LLM claim verification, deterministic quote matching, confidence score, finite-state decision; IDF topical-fidelity backstop. |
| Learning | Three-tier dedup: exact-text match with priority tie-break, LLM semantic dedup, lexical-similarity fallback. |
| Q&A / Quiz | Grounded LLM generation plus deterministic citation enforcement against the evidence store. |

### *C. Agent Responsibilities and Decision Boundaries*

Each agent's authority, preventing any single agent from becoming both generator and verifier of the same content, which the gate-on/gate-off comparison in Section V depends on.

### *D. Coordination Protocol: A Worked Example*

Consider one audio chunk moving through the system. (1) The browser posts a chunk, dispatched to Capture. (2) Capture transcribes it into a structured, source-tagged chunk. (3) Notes writes a candidate note from its rolling window, returned to the browser as pending; this fast path does not wait for verification. (4) The Coordinator dispatches an evidence check as a background task. (5) Evidence runs its five-stage pipeline (Section III-F); on failure within the retry bound it writes structured feedback, never editing note text, and control returns to Notes, which regenerates and resubmits. (6) Once resolved, Learning consolidates it and pushes a status-badge update over server-sent events. (7) Only then is the note eligible as source material for Q&A and Quiz, if verified and gating is enabled.

TABLE II. Agent Decision Boundaries

| Agent | Decides | Does not do |
|---|---|---|
| Coordinator | Event routing; session finalization | Generate content |
| Capture | Modality routing, vision fallback, chunk boundaries | Summarize or reason over content |
| Notes | Note structure; revision from Evidence feedback | Modify locked notes; self-verify |
| Evidence | Verified / regenerate / unverified | Edit note text |
| Learning | New vs. duplicate vs. related items | Auto-apply merges without confirmation |
| Q&A | Grounded vs. general-knowledge answer | Present an unverifiable citation as grounded |
| Quiz | Concept selection and difficulty | Use unverified notes when gating is enabled |

Two guarantees hold throughout. *Write isolation:* Notes writes only note text, Evidence only verification status and links. *Lock precedence:* every agent re-checks a note's lock state before any autonomous write, including on retry re-entry, so an

in-flight cycle cannot silently overwrite a mid-cycle edit. At session end, the Learning Agent's three passes (deduplication, priority merge, cross-lecture linking) run as independent, fixed-order invocations.

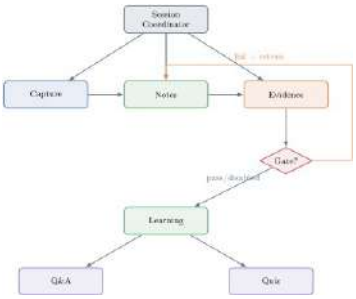



Fig. 2. Orchestration graph and verification loop.

### *E. Evidence Verification Algorithm*

The Evidence Agent runs whenever a note is produced or revised; the gating flag only determines whether its verdict is advisory or binding. Verification proceeds in five stages: (1) candidate evidence retrieval via lexical top-$k$ difflib. SequenceMatcher similarity, a deliberately simple, inspectable step; (2) an LLM call returning a support decision and, when supported, a verbatim source quotation, instructed to rely on supplied evidence rather than parametric knowledge; (3) deterministic re-confirmation of that quotation by exact substring match with a fuzzy longest-common-run fallback for ASR drift; if the quotation cannot be located, the evidence is discarded regardless of the model's verdict (the fail-closed property); (4) confidence scoring, $c = 0$ if unsupported or unmatched, else $c = w_1 v + w_2 m$, combining verifier confidence $v$ and match quality $m$; and (5) a finite-state gate decision over {VERIFIED, REGENERATE, UNVERIFIED} against threshold $\theta = 0.7$ and a bound of $N = 2$ regenerations (at most three verification attempts). A diagnostic-only IDF topical-fidelity check flags off-topic fuzzy matches without gating on them.

*Hyperparameter rationale.* The equal weights $w_1 = w_2 = 0.5$ encode no prior preference between verifier confidence and quote-match quality; because $c = 0$ whenever support is denied or the quote cannot be located, the fail-closed stages rather than the weighting determine most decisions, and $\theta = 0.7$ therefore requires a located quote with high combined confidence. $N$ is examined empirically in Section V; $\theta$ was not ablated.

**Algorithm 1** Evidence verification and gate decision

```
function VERIFY(note, sources, gate, r)
    cand ← RETRIEVESUPPORT(note, sources)  D top-k
    verdict ← GROUNDEDCHECK(note, cand)
    if verdict.supported = false then
        c ← 0; quote ← NONE
    else
        quote ← verdict.quote; v ← verdict.confidence
        match ← DETMATCH(quote, cand)
        if match.found = false then
        c ← 0; quote ← NONE  D fail-closed
    else
        m ← match.score; c ← w1v + w2m
        LOGTOPICALFIDELITY(note, cand)  D diagnostic
    end if
end if
if gate = false then
    RECORDADVISORY(note, c); return ACCEPT
end if
if c ≥ θ then
    RECORDVERIFIED(note, quote, c); return ACCEPT
else if r < N then
    fb ← BUILDFEEDBACK(note, cand, quote)
    return REGENERATE(fb)  D hands off to Notes
else
    MARKUNVERIFIED(note, c)  D status only
    return ACCEPTWITHWARNING
end if
end function
```

Algorithm 1 gives the coordination logic: the Evidence Agent only ever returns REGENERATE with feedback or a status write, never editing note text; regeneration is the Notes Agent's job, preserving write isolation.

### *F. Explainability via Occlusion-Based Attribution*

Evidence badges show that a note is backed by a quotation; occlusion-based attribution (in the family of LIME/SHAP) additionally shows which span drove the decision, by removing a candidate span from the evidence passed to Stage 3 and re-running the match. Because Stage 3 is deterministic, this is exact remove-and-recheck attribution rather than a sampled approximation.

### *G. Learning Agent Consolidation and Downstream Propagation*

The Learning Agent resolves conflicts between audio, slides, and student edits using an explicit priority policy LOCKED > SLIDE > AUDIO, enforced as a coordination rule available to every writing agent so the lock guarantee holds even when several agents touch the same note. Deduplication runs as three tiers: exact-text match with priority tie-break; LLM-based semantic dedup for near-duplicates; and a lexical-similarity fallback that places every item rather than dropping it silently. When gating is enabled, only verified notes propagate into cross-lecture structures and into Q&A and Quiz, so a rejected note cannot become an apparently authoritative downstream source. Both agents enforce citation grounding by set-membership against the evidence store rather than trusting the model's self-report.

## IV. Experimental Setup

*Corpus and annotation.* The system was evaluated on ten self-delivered computing lectures (167 min in total, seven domains, 14,530 reference words across 2,039 live audio chunks). ASR references were produced by hand-correcting raw transcripts against the recordings, excluding genuinely unintelligible spans. Source-fact coverage was scored against 3,122 source-fact units (transcript sentences and slide/document lines). Note annotation used an AI-assisted first pass reviewed by the researcher, blind to condition and to the Evidence Agent's verdict. Scoring rules (coverage attribution, exclusion of unintelligible spans from WER) were fixed before results were computed and are enforced by regression tests. Every note in the gate-off (C3, 417 notes) and gate-on (C4, 386 notes) conditions was labelled supported or unsupported. The C3–C4 comparison differs only in the evidence.gating_enabled flag, with the same recordings, sources, model, and prompts; because C4 is a

separate generation run, it also contains run-to-run LLM variation.

*Reliability.* An independent second annotator labelled 81 records drawn from the 511-record three-lecture annotation store by fixed-seed proportional sampling across lectures, without access to the primary labels or agent verdicts. The Evidence Agent's precision/recall and the retry-bound analysis use an earlier development run on lectures L1–L3 (132 notes) that stored every verification attempt with human labels.

*Baseline (E4).* The full system (A) is compared with a non-agentic baseline (B) on five additional lectures not in the main corpus, with the same audio-only input and model. B sends the transcript in a single API call and verifies each bullet by calling the same evidence-verification function directly, with no orchestration graph, separate agents, regeneration, or consolidation. Short-bullet (<6 words), dangling-pronoun, and duplicate-bullet rates and total generation time are compared.

*User study.* Twenty-five undergraduates from the researcher's faculty were recruited by convenience; participation was voluntary, unpaid, and anonymous (P01–P25), with verbal informed consent after a briefing. Each participant ran the system live while playing a publicly available English lecture video (approximately 15–20 min; videos differed across participants), used the notes and learning features, then completed a 12-item questionnaire: six direct 1–5 ratings, four 1–5 ratings against their usual manual notes (3 = parity), a forced-choice preference item, and an optional free-text item.

*Statistics.* Coverage and unsupported-claim rates are compared with a two-sided exact sign test on the ten paired lecture-level differences, so that the lecture rather than the clustered fact is the unit of analysis (a fact-level McNemar test is not applicable because per-fact paired labels were not recorded). The gate comparison is unpaired (C3 and C4 do not share note identities), so Fisher's exact test is also applied. Pooled tests treat facts and notes as independent although they are clustered within ten lectures by one speaker, so they are read as within-dataset evidence; the lecture-level sign tests, which treat each lecture as one paired unit, are the primary inferential evidence. Participants are independent units, each contributing one response set. Questionnaire items are tested against the scale midpoint with one-sample Wilcoxon signed-rank tests, and the preference item with an exact binomial test.

## V. Results

**RQ1/RQ2 – Transcription and generation.** Table III reports transcription and incremental-generation metrics. Per-lecture WER ranged from 3.81% to 16.59%, with two lectures above the 10% target, and processing of every chunk completed before the next arrived in all ten lectures, though the separate 500 ms per-segment latency target was missed. Because each chunk carries about 5 s of audio, a median latency of 933 ms means a transcribed segment appears roughly one second after it is spoken, about 0.4 s later than the responsiveness budget; since processing time is under a fifth of audio duration (RTF 0.19), no backlog accumulates and no audio is lost, so the miss affects display delay rather than capture completeness. Five of ten lectures had a median generation latency above 3 s. The fast/slow-path split meant slower structuring cycles delayed note appearance rather than interrupting capture. Leakage and duplication were rare but not eliminated.

TABLE III. Transcription and Generation Metrics (RQ1–RQ2)

| Metric | Value |
|---|---|
| Pooled WER (target <10%) | 8.46% |
| Real-time factor, median / worst chunk | 0.19× / 0.82× |
| Per-segment latency, median of lecture medians (target 500 ms) | 933 ms |
| Note-generation latency, median of lecture medians (target 3 s) | 2,985 ms |
| Coherence / organizational fit (55 groups, single rater) | 3.84 / 3.71 (of 5) |
| Meta-commentary leakage | 0.48% (2/417) |
| Duplicate-bullet rate | 0.96% (4/417) |

**RQ3/RQ4 – Multimodal coverage and verification gate.** Table IV reports coverage and gate outcomes. Fusing slide/document evidence with audio raised coverage in all ten lectures (+1.9 to +20.9 points). The multimodal condition also produced 1.51× more notes, so part of the gain reflects note volume. Enabling the gate with bounded regeneration (Algorithm 1) reduced the unsupported-claim rate in every lecture (by 10.61 to 15.66 points), though not to zero, at a median cost of 5.7 s of additional processing per note; residual cases indicate that retrieval and regeneration quality remain areas for improvement.

TABLE IV. Multimodal Coverage and Verification-Gate Outcomes (RQ3–RQ4)

| Metric | Value |
|---|---|
| Coverage, audio-only | 84.69% (2,644/3,122) |
| Coverage, audio + slides/documents | 92.34% (2,883/3,122) |
| Lectures improved; sign test | 10/10; p = 0.00195 |
| Unsupported-claim rate, gate off (95% CI) | 21.58% (17.9–25.8) |
| Unsupported-claim rate, gate on (95% CI) | 8.55% (6.2–11.8) |
| Fisher's exact; lecture sign test | $p = 2.49\times10^{-7}$; p = 0.00195 |
| Median verification overhead per note | +5,672 ms |
| Evidence Agent precision / recall / F1 † | 93.09 / 74.89 / 83.00% |
| Accepted at check 1 / check 2; reached check 3 † | 72.7 / 12.1; 15.2% |
| Terminally unverified at N = 2 † | 9.85% (13/132) |

† Earlier development run, lectures L1–L3 (Section IV).

*Retry-bound sensitivity.* Re-applying $N \in \{0, 1, 2, 3\}$ to the stored per-attempt records with θ fixed (Table V) shows diminishing acceptance gains as latency rises: $N = 3$ adds a single accepted note over $N = 2$, and that note was human-labelled unsupported, supporting $N = 2$ as the bound. The analysis does not establish an optimal bound for other lectures.

TABLE V. Retry-Bound Sensitivity (132 Notes, Θ = 0.7)

| N | Accepted | Mean latency | Unsupported among accepted |
|---|---|---|---|
| 0 | 72.73% | 3.04 s | 3.13% |
| 1 | 84.85% | 3.32 s | 4.46% |
| 2 | 90.15% | 3.73 s | 4.20% |
| 3 | 90.91% | 4.02 s | 5.00% |

*Baseline comparison (E4).* Table VI compares the two configurations. The multi-agent system had lower short-bullet, dangling-pronoun, and duplicate-bullet rates in all five lectures,

but took about 3.1× longer to process a lecture. Duplicate removal reflects the consolidation stage available only to A, so the differences are attributable to the complete pipeline rather than to agent decomposition alone.

**RQ5 – Human perception.** All six direct ratings exceeded the scale midpoint (Table VII), with revision usefulness highest and trust lowest. All four comparisons with participants' usual manual notes were above parity (means 4.44–4.68), and 17 of 25 participants preferred the system for future lectures (exact binomial $p = 1.5\times10^{-5}$ among directional responses); the remaining eight chose both approaches or no preference, and none chose manual notes alone. No workload instrument was administered, so no claim about cognitive load is made.

TABLE VI. MULTI-AGENT SYSTEM (A) VS. NON-AGENTIC BASELINE (B), MEAN OF FIVE LECTURES

| Metric | A | B |
|---|---|---|
| Total generation time per lecture | 78.4 s | 25.1 s |
| Short-bullet rate (<6 words) | 1.9% | 18.9% |
| Dangling-pronoun rate | 0.4% | 5.9% |
| Duplicate-bullet rate | 0.0% | 7.5% |

TABLE VII. PARTICIPANT RATINGS, N = 25 (RQ5)

| Measure | Mean (SD) | Rated ≥4 |
|---|---|---|
| Ease of use | 4.28 (0.54) | 96% |
| Organization | 4.40 (0.58) | 96% |
| Accuracy | 4.44 (0.65) | 92% |
| Completeness | 4.48 (0.51) | 100% |
| Trust | 4.20 (0.71) | 84% |
| Revision usefulness | 4.60 (0.58) | 96% |

All items $p < 10^{-4}$ vs. midpoint (one-sample Wilcoxon signed-rank).

## VI. DISCUSSION

The results support treating the coordination design as inseparable from the primary contribution: each capability already exists in the literature, but the evaluation exercises their combination as a coordination problem. The RQ4 finding is arguably the most practically important result: giving the Evidence Agent a bounded, structured way to hand a failed note back to Notes, rather than silently accepting unsupported content or discarding it, reduced the unsupported-claim rate in every lecture without stalling capture, and the retry-bound analysis shows that most of the benefit is obtained within two regenerations. The baseline comparison shows the cost side of the design: cleaner notes at roughly three times the processing time.

The evaluation has five classes of limitation. *Architectural validity:* the E4 baseline differs from the full system in regeneration and consolidation as well as orchestration, and was measured on text-quality defects and time rather than coverage or unsupported-claim rate; a role-isolating single-agent ablation remains future work. *Construct validity:* WER measures transcription accuracy, not educational usefulness; unsupported-claim rate measures source support, not student trust. *Internal validity:* the same researcher recorded lectures, produced ASR references, and annotated content. This was mitigated by blinded labelling, pre-fixed scoring rules, lecture-level tests, and an independent second annotator (78/81 agreement, 96.30%, $\kappa = 0.836$), but single-annotator dependence remains for the full gate-comparison labels, the ASR references, and the source-fact inventory. *External validity:* the corpus is ten lectures by a single speaker across seven computing domains; participants were 25 undergraduates from one faculty recruited by convenience; and the system supports English only. Recorded English lectures were used deliberately because lectures at the host institution commonly mix English with Sinhala, which would have confounded system quality with an unsupported language. Results may change with other speakers, disciplines, models, prompts, or thresholds. *Conclusion validity:* ten lectures limit the power of lecture-level tests, and the participant comparison with manual note-taking is retrospective and self-reported rather than observed under matched conditions.

## VII. CONCLUSION

This paper presented a real-time learning assistant built as a collaborative multi-agent system, with particular attention to the coordination mechanism that makes six specialized agents (Capture, Notes, Evidence, Learning, Q&A, and Quiz) operate together under live-lecture latency constraints: a shared typed state in place of direct messaging, an explicit orchestration graph, per-agent write isolation, and a lock guarantee re-checked on every write. Evaluated on a ten-lecture corpus with 25 participants, the system met its pooled transcription target and kept pace with live audio, although the 500 ms per-segment latency target was not met (933 ms) and note-generation latency exceeded 3 s in half the lectures. Multimodal fusion increased fact coverage from 84.69% to 92.34%, and a bounded-regeneration verification gate cut the unsupported-claim rate from 21.58% to 8.55%, both in every lecture. Participants rated the system positively across usability, trust, and revision usefulness. Future work should extend evaluation to larger lecture corpora from multiple lecturers, institutions, and subjects, including code-mixed lectures, and to larger and more diverse participant samples; run a role-isolating single-agent ablation and a verification-threshold sweep; add independent annotation of the full corpus; improve evidence-retrieval recall; and study learning outcomes directly.

## ACKNOWLEDGMENT

The authors thank the undergraduate students of NSBM Green University who volunteered their time to take part in the user study.

## REFERENCES

[1] M. Bond, H. Khosravi, M. D. Laat, N. Bergdahl, V. Negrea, E. Oxley, P. Pham, S. W. Chong, and G. Siemens, "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour," International Journal of Educational Technology in Higher Education, vol. 21, no. 4, 2024.

[2] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, "Large language models for education: A survey and outlook," arXiv preprint arXiv:2403.18105, 2024.

[3] T. Kawahara, N. Katsumaru, Y. Akita, and S. Mori, "Classroom note-taking system for hearing impaired students using automatic speech recognition adapted to lectures," in Proceedings of Interspeech 2010, 2010, pp. 626–629.

[4] Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proceedings of the 40th International Conference on Machine Learning, vol. 202, 2023, pp. 28492–28518.

[5] H. Gonzalez, J. Li, H. Jin, J. Ren, H. Zhang, A. Akinyele, A. Wang, E.

Miltsakaki, R. S. Baker, and C. Callison-Burch, “Automatically generated summaries of video lectures may enhance students’ learning experience,” in Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, 2023, pp. 382–393.

[6] M. Jiang, J. Gao, Z. Pan, Y. Wu, and Z. Wang, “Nexanota: An AI-powered smart linked lecture note-taking system leveraging large language models,” in Proceedings of the 2025 International Conference on Big Data and Informatization Education, 2025, pp. 242–248.

[7] W. Xing, T. Zhu, J. Wang, and B. Liu, “A survey on MLLMs in education: Application and future directions,” Future Internet, vol. 16, no. 12, p. 467, 2024.

[8] Ren, Z. Pardos, and Z. Li, “Human-AI collaboration increases skill tagging speed but degrades accuracy,” arXiv preprint arXiv:2403.02259, 2024.

[9] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, “Does the whole exceed its parts? The effect of AI explanations on complementary team performance,” in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.

[10] Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, “Self-refine: Iterative refinement with self-feedback,” in Advances in Neural Information Processing Systems, vol. 36, 2023.

[11] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, “Chain-of-verification reduces hallucination in large language models,” in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 3563–3578.

[12] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 9004–9017. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.

[13] Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to retrieve, generate, and critique through self-reflection,” in The Twelfth International Conference on Learning Representations, 2024.

[14] S. Miya, “Eidoku: A neuro-symbolic verification gate for LLM reasoning via structural constraint satisfaction,” arXiv preprint arXiv:2512.20664, 2025.

[15] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, “AutoGen: Enabling next-gen LLM applications via multi-agent conversation,” arXiv preprint arXiv:2308.08155, 2023.

[16] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber, “MetaGPT: Meta programming for a multi-agent collaborative framework,” in The Twelfth International Conference on Learning Representations, 2024.

[17] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, “CAMEL: Communicative agents for “mind” exploration of large language model society,” in Advances in Neural Information Processing Systems, vol. 36, 2023.

[18] B. A. Chaushi, B. Selimi, A. Chaushi, and M. Apostolova, “Explainable artificial intelligence in education: A comprehensive review,” in Explainable Artificial Intelligence, xAI 2023, Communications in Computer and Information Science, vol. 1902, 2023, pp. 48–71.

[19] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” Journal of Management Information Systems, vol. 24, no. 3, pp. 45–77, 2007.

# AI Student Guide Hub: An LLM-Orchestrated Context-Aware Framework for Personalized Career Recommendation and Learning Path Generation

Banoth Shiva Kumar
*Information Technology*
*Vardhaman College of Engineering*
*Hyderabad, Telangana*
shivakumarnayak2005@gmail.com

Talla Prashanthi
*Information Technology*
*Vardhaman College of Engineering*
*Hyderabad, Telangana*
talla.prashanthi18@gmail.com

Budidha Adarsh Goud
*Information Technology*
*Vardhaman College of Engineering*
*Hyderabad, Telangana*
badarshgoud2004@gmail.com

Ganesh B Regulwar
*Information Technology*
*Vardhaman College of Engineering*
*Hyderabad, Telangana*
ganeshregulwar@gmail.com

Neelakantam Shiva
*Information Technology*
*Vardhaman College of Engineering*
*Hyderabad, Telangana*
shivaneelakantam1605@gmail.com

Babitha Lincy R
*Department of Computer and Communication Engineering*
*Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu*
babithalincy.r@sece.ac.in

***Abstract*—Engineering students often run into trouble when they're hunting for internships or jobs. Just figuring out the right path can feel overwhelming, and honestly, most people aren't sure what skills they actually need. The information they need is scattered everywhere — job requirements, learning resources, and prep guides all live on different sites. It's confusing. This paper introduces the AI Student Guide Hub, a website that pulls everything into one place. The platform helps students prepare for their careers by offering practical tools. Google Gemini powers some cool features: it reviews resumes, gives personalized advice, and even generates quizzes. When a student uploads their resume, the system scans for their skills, checks if they match with the jobs they're aiming for, and highlights the gaps. So, students know exactly what to learn. You also get help choosing a career, figuring out what skills you need, and resources for building those skills. It offers AI-driven mock interviews and detailed info about internships too. The paper dives into how the system works, its architecture, and how the team built it. They walk through an early evaluation of how well the tool connects resumes to job roles and spots skill gaps. The evaluation uses keyword matching and a set of predefined criteria. Since they tested it with synthetic student profiles and their own scoring system, you shouldn't treat the results as real evidence of readiness or the tool's effectiveness in the real world. The authors are up front about its limits and hint at ways to run deeper evaluations in the future.**

***Keywords—Career Recommendation, Large Language Models (LLMs), Resume Analysis, Skill Gap Analysis, Personalized Learning, AI-Assisted Career Guidance, Educational Recommendation Systems.***

## I. Introduction

Artificial intelligence (AI) is being used more in colleges to help students learn in ways that fit them and to give them career advice that matches new industry needs [1] [2]. As Industry 4.0 and Industry 5.0 grow engineering students must keep their knowledge and skills in line with what employers want. However many students have trouble choosing the career routes finding out which skills they need for their desired jobs and getting ready for campus hiring. Career tips, job needs and study materials are often spread over sites so preparing for a career becomes broken up and hard to follow [3] [4]. New research has looked at language models (LLMs) to help with resume checking chat help and tailored learning. Old methods, like Term Frequency–Inverse Document Frequency (TF-IDF) can match keywords. Do not always show the meaning links, between skills and job needs. LLMs can read text context. Give natural language feedback [5] [6]. We must check how useful, steady and trustworthy their answers are instead of just trusting them. These abilities give chances to help students spot the skills and get learning suggestions that match their career plans [7] [8]. Artificial intelligence can also help students see how their skills match job demands. However putting these features into one place does not by itself make sure the suggestions are correct or work well. These kinds of systems need ways to check how good they are and proper standards to compare them against. Existing career preparation methods might look at resume reviews, career advice, identifying skill gaps and interview practice as tasks [9] [10] which means there is a need, for combined platforms where the features they offer and the limits of their assessments are clearly explained.To address this need, this paper presents AI Student Guide Hub, a web-based framework intended to bring selected career-preparation and learning-support functions together in a unified interface for engineering students. The system employs Google Gemini for certain AI-assisted tasks such as resume analysis, mentorship, and quizzes creation. The resume analysis task offers AI feedback concerning the resume that has been provided. In addition to that, there are other career preparation functions offered by the system, including the provision of internship related information and the ability to conduct a mock interview.

This paper focuses on the system's design and implementation and an exploratory evaluation approach for resume-to-role alignment and skill-gap identification. The available quantitative benchmark uses 20 synthetic student profiles and author-defined reference scores, with a keyword-based method serving as a baseline. These reference scores are not independent human-expert assessments, and the current

benchmark does not establish real-world employability, career readiness, or the superiority of Gemini over keyword-based matching.

## II. Related work

Artificial intelligence (AI) and machine learning have increasingly been used in career guidance, resume analysis, and job recommendation. These systems help students explore career opportunities based on their academic backgrounds, interests, and skills. However, career prediction and educational recommendations serve different purposes, and their results must be evaluated accordingly. Their effectiveness also depends on the quality of the available data and how student skills and job requirements are represented Recent advances in large language models (LLMs) have created new opportunities for contextual text analysis and personalized career guidance. Transformer-based models such as BERT generate context-aware text representations, while generative models such as GPT produce natural language responses. These capabilities can support resume analysis and skill identification, but their outputs may vary in relevance and consistency, making systematic evaluation essential. Traditional techniques, such as Term Frequency–Inverse Document Frequency (TF-IDF), provide a straightforward approach to matching resume keywords with job requirements. However, they may overlook semantic relationships between differently expressed skills. Skill-gap analysis focuses on identifying the competencies required for a particular role that are missing from a student's available profile. Its accuracy depends on how skills are extracted, represented, and compared. Similarly, recommendation techniques have been explored to suggest educational resources based on learner information and preferences [14], [15]. However, claims about specific methods and their effectiveness must be supported by the corresponding studies. Despite these developments, resume evaluation, career counselling, job recommendations, learning-resource selection, and interview preparation are often handled separately, requiring students to use multiple platforms. This highlights the need for integrated career-preparation systems that connect these functions while clearly defining their capabilities and evaluation methods.

To address this need, this paper presents AI Student Guide Hub, a web-based platform that combines selected career-preparation and learning-support features. It uses Google Gemini for AI-assisted resume analysis, mentoring, and quiz generation, alongside mock interviews and internship-related resources. The primary contribution is the integration of these features rather than the development of a new language model or recommendation algorithm. The study also explores a keyword-based baseline using 20 synthetic student profiles and manually defined reference scores. Since these scores have not been independently validated by human experts, they cannot establish actual career readiness, employability, or the superiority of LLM-based recommendations.

## III. Proposed System

The AI Student Guide Hub pulls together existing tech to help students build better resumes, explore different jobs, spot where their skills come up short, and get ready for their careers. It's not trying to reinvent the wheel with some new language model or fancy recommendation engine. Instead, it

blends a web interface, a backend service, and a database to keep student info organized and give solid, AI-powered advice—all in one place.

### *A. System Overview*

This system's built around four main pieces: student profile and resume analysis, career guidance and skill development, learning roadmaps, and career prep activities. On the frontend, it runs React 19 with Vite. The backend relies on Node.js and Express. Data sits in PostgreSQL, and Sequelize takes care of talking to the database. If you need real-time updates, Socket.IO steps in. For natural-language processing, they've plugged in Google Gemini 2.5 Flash to specific backend services. The setup keeps frontend, backend logic, database access, and AI-service calls clearly separated, so everything stays organized.

### *B. Student Profile and Resume Analysis*

The resume-analysis component sends resume text (up to 4,000 characters) to the backend, which uses Gemini 2.5 Flash to generate a score, summary, and improvement suggestions. The score is AI-generated feedback, not a validated measure of resume quality or employability. Reproducibility requires documenting the model version, prompt, generation settings, and input-processing procedure.

### *C. Career-Role Alignment and Skill-Gap Identification*

Career-role alignment compares a student's documented skills with the requirements of a target role. The overlap identifies matching skills, while missing requirements highlight potential skill gaps.

Formally, let ($S_p$ ) denote the set of skills identified in a student's profile and ($S_r$) the set of skills associated with a target role. The set of potentially unrepresented requirements can be expressed as

The formula for skill-gap identification is:

$$G = S_r \setminus S_p$$

Description of the formula:

Where:

(G) represents the set of potential skill gaps.
($S_r$) represents the set of skills required for a target career role.
($S_p$) represents the set of skills identified in the student's profile or resume.
(/) denotes the set difference operation, meaning skills present in $S_r$ but absent from $S_p$.

where (G) represents potential skill gaps, rather than proof that the student lacks those competencies. This statement lays out how to compare skills for a gap analysis, but it doesn't prove the current app actually has automated extraction and matching running under the hood. When reviewing profile info, you can't just assume someone lacks a skill because it's not listed, and

seeing a skill on a resume doesn't automatically mean they're truly proficient. So, any match results should be treated as just a starting point, not a final verdict. If the code doesn't actually do automated skill extraction and calculate gaps for specific roles yet, describe those as goals for the project or as things still in the evaluation phase—not as features that work today.

### *D. Learning-Roadmap Support*

The learning-roadmap component lets students organize their career-development tasks with local templates. There's an AI-roadmap backend, but the frontend doesn't tap into it yet. Right now, students get the usual template-based guidance. Personalized learning paths powered by AI are still on the horizon.

### *E. Career-Preparation Activities*

The platform helps you get ready for your career with AI-powered quizzes, chat-based mentoring, mock interviews, and info about internships. Still, keep in mind: some mock interview scores, dashboard indicators, and internship-match results rely on preset or simulated numbers. Don't take them as solid proof of how ready you are, how well you fit a career, or if you'll land an internship.

### *F. System Workflow*

The principal workflow consists of the following stages:

1. **Student input:** The student accesses the web interface and provides supported information, such as resume text or a query for the mentor service.
2. **Backend processing:** The frontend sends the request to the relevant Express endpoint, which performs the corresponding application operation.
3. **AI-assisted processing:** For supported tasks, the backend invokes Gemini 2.5 Flash to obtain resume-analysis feedback, quiz content, or mentor responses.
4. **Response presentation:** The backend returns the result to the frontend, where it is presented to the student. Other features display information produced by their respective application logic or predefined templates.
5. **Student review:** The student reviews the available feedback and uses it to inform subsequent career-preparation and learning activities.

This workflow describes the interaction between the web application and its selected AI services. It does not imply that every component invokes an LLM, that every recommendation is personalized using the student's complete profile, or that generated outputs are automatically verified against employment outcomes.

### *G. Design Rationale, Reliability, and Limitations*

The main thing AI Student Guide Hub does is bring career-prep tools together in one place: resume feedback, conversational mentoring, quiz generation, learning roadmaps, and internship information. So its contribution is really about integrating these features into a single interface and running an exploratory evaluation—not about building a new AI model or recommendation algorithm.

## IV. METHODOLOGY

### *A. Student Data Collection*

It all starts when a student signs up and creates a profile. That profile can include their education, technical skills, programming languages, certifications, project experience, career goals, preferred job roles, and interests in areas like AI, data science, web development, cybersecurity, and cloud computing. This information gives the platform the initial context it needs for resume analysis and other career-support features. Of course, how useful the guidance is depend on how complete and accurate the student's information is.

### *B. Resume Analysis*

After creating a profile, students can upload their resumes for AI-assisted analysis. The system processes up to 4,000 characters of resume text using Google's Gemini 2.5 Flash model to generate a summary, identify relevant skills, and provide feedback. This helps students understand how effectively their qualifications, projects, skills, and experience are presented and identify areas for improvement. However, since LLM-generated responses may overlook or misinterpret information, the results should be treated as guidance rather than an independently verified assessment of a student's qualifications.

### *C. Skill and Career Analysis*

At this stage, the platform uses the student's profile and career interests to identify skills relevant to their chosen role and highlight areas that may require further development. By comparing available skills with role-specific requirements and assessment criteria, it helps students recognize potential skill gaps and plan their learning accordingly. However, these results are intended as preliminary career guidance and should not be treated as validated predictions of employability, career suitability, or future job performance

### *D. Recommendation Generation*

The recommendation stage helps students identify next steps through resume feedback, practice quizzes, mentoring, and career-preparation activities. Gemini 2.5 Flash powers resume analysis, quiz generation, and mentor interactions, while other features use predefined content or application logic, combining AI-assisted functionality with conventional features.

### *E. Personalized Learning Roadmap*

The learning-roadmap feature provides students with structured, career-oriented learning activities to help them achieve their chosen career goals. It is integrated with the AI-powered roadmap generation endpoint in the backend, enabling the system to generate learning recommendations based on students' selected career paths and learning objectives. The feature organizes recommended topics, skills, and activities into a structured learning sequence, helping students follow a clear

path toward their desired careers. Future enhancements will focus on deeper skill-gap analysis, improved personalization, and dynamically adapting learning sequences based on students' existing skills and progress.

*F. Career-Preparation Support*

The platform supports technical assessment, internship, and placement prep through quizzes, mentoring, resume feedback, mock interviews, and opportunity info. Gemini powers quizzes and mentoring; other features rely on app logic and content. Internship matching, interview scoring, and placement predictions still need separate validation.

*G. Experimental Evaluation*

The experimental evaluation compares system outputs against predefined reference labels using 20 synthetic resume profiles. These author-defined labels are not expert-verified ground truth for student ability or employability. A keyword-based baseline provides comparison, with both methods evaluated using the same profiles, labels, and metrics. Numerical predictions are assessed using MAE, RMSE, Pearson correlation, and Spearman correlation, while skill identification uses precision, recall, and F1-score. Reproducibility requires recording model details, prompts, generation settings, response times, and available token usage.

Let N be the number of evaluated profiles, $y_i$ the reference score, and $\hat{y}_i$ the predicted score.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$r = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \hat{\bar{y}})^2}}$$

$$\rho = \text{Corr}(\text{rank}(y), \text{rank}(\hat{y}))$$

$$\text{Precision} = \frac{TP}{TP+FP}, \ \text{Recall} = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **TP: a skill label predicted by the system that also appears in the reference labels.**
- **FP: a predicted skill label that is absent from the reference labels.**
- **FN: a reference skill label that the system failed to predict.**

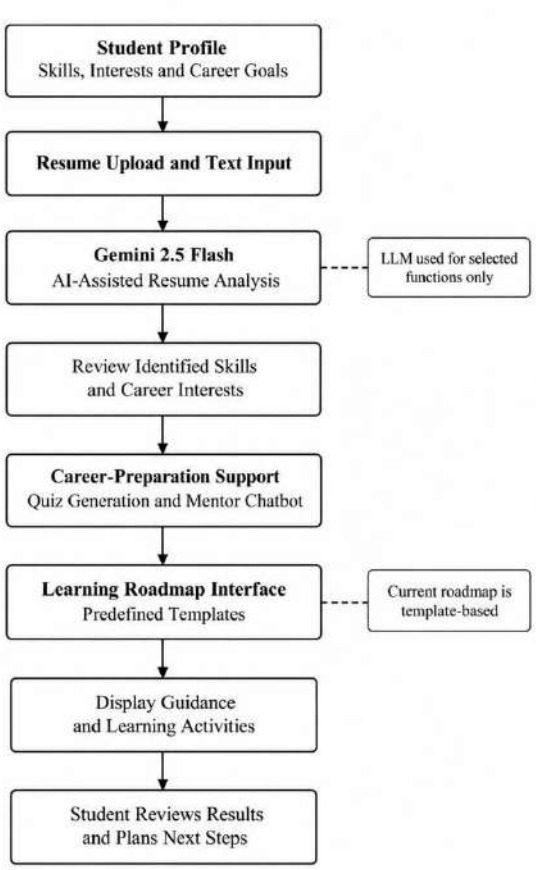



Fig. 1. System Architecture

*Figure 1 illustrates how the AI Student Guide Hub works, bringing together student profile information, resume upload, and AI-assisted resume analysis powered by Google Gemini 2.5 Flash. The framework supports reviewing identified skills and career interests, career-preparation activities through quiz generation and a mentor chatbot, and template-based learning-roadmap guidance to help students plan their next steps in learning and career development.*

## V. IMPLEMENTATION

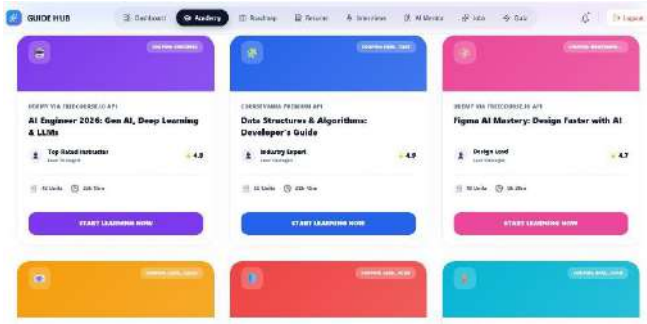


Fig. 2. AI-Powered Learning Hub

Figure 2 shows the Learning Hub module of the AI Student Guide Hub. This module presents learning resources and course-related content to help students explore educational activities that match their interests and career-development goals.

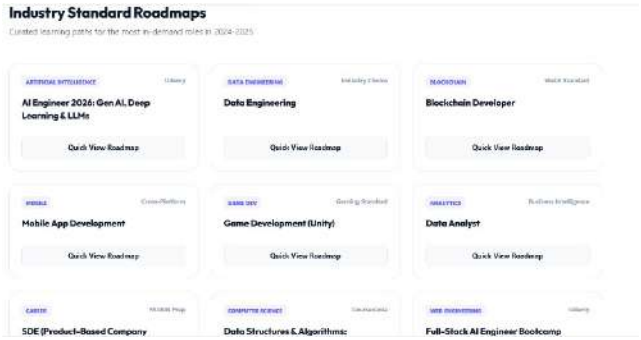



Fig. 3. Personalized Learning Roadmap Module

Figure 3 presents the Learning Roadmap module, which uses predefined learning templates to help students organize learning activities around the career direction they've selected. As currently implemented, it provides template-based guidance rather than dynamically generating a personalized learning path with an LLM.

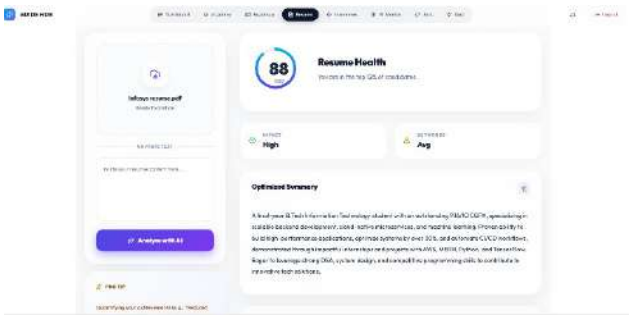



Fig. 4. AI-Assisted Resume Analysis

Figure 4 illustrates the resume-analysis module, which uses Google Gemini 2.5 Flash to analyze submitted resume text and generate a summary, identify relevant skills, and provide feedback to support resume improvement and career preparation.

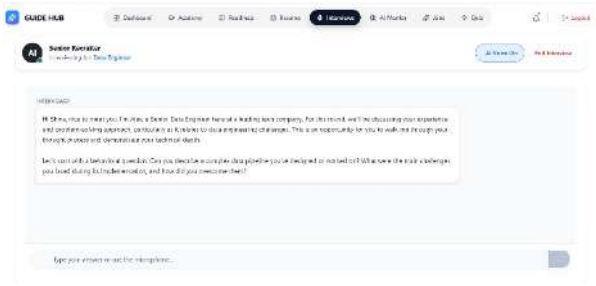


Fig. 5. AI-Powered Voice Mock Interview

Figure 5 shows the mock-interview interface provided for interview practice. The interface supports interview-related interactions; however, validated automated interview-performance scoring is not established in the current implementation.

.

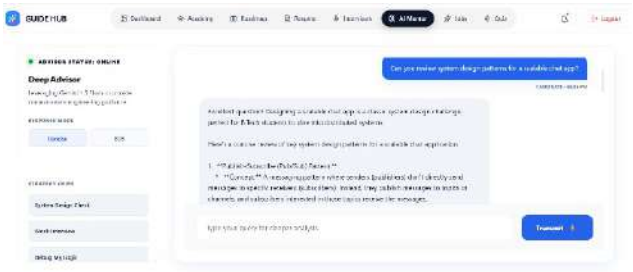



Fig. 6. AI-Powered Mentor

Figure 6 shows the Mentor module, which integrates Google Gemini 2.5 Flash into a conversational interface to provide students with technical guidance, coding-related assistance, and career-oriented support through interactive chat.

## VI. Results

AI Student Guide Hub is an integrated web platform built around a few key career-preparation activities: resume analysis, AI-assisted quiz generation, conversational mentoring, learning-resource presentation, and a template-based learning roadmap. Google Gemini 2.5 Flash is built into some of the workflows, while other features rely on predefined content or conventional application logic. This section covers the functionality that was actually implemented, some early response-time observations, and an exploratory comparison against a keyword-based baseline—not claims about improved employability or a validated career-readiness prediction.

### *A. Functional Implementation*

TABLE I
COMPARISON WITH EXISTING APPROACHES.

| **Capability** | **Conventional keyword-based approach** | **Proposed AI Student Guide Hub** |
|---|---|---|
| Resume analysis | May rely on keyword matching or dedicated resume-analysis tools | Gemini-assisted resume analysis |
| Career preparation | Often distributed across separate services | Integrated into one web platform |
| Skill development support | Depends on the individual tool | Career-related guidance and learning-roadmap interface |
| Learning roadmap | Varies by platform | Predefined learning templates in the current implementation |
| Interactive support | Depends on the service | Gemini-powered quiz generation and mentor chatbot |

Table I describes the platform's functional scope rather than an empirical comparison against all existing career-guidance

systems. A controlled comparison with keyword matching, TF-IDF, or embedding-based approaches would require implementing or selecting those methods and evaluating them on the same dataset.

### *B. Implementation Workflow*

TABLE II
WORK FLOW AND IMPLEMENTED FUNCTIONS OF AI STUDENT GUIDE HUB

| Stage | Component | Function in the Current Implementation |
|---|---|---|
| Input | Student profile and resume upload | Collects student information and accepts resume input |
| AI-assisted analysis | Resume analyzer | Uses Gemini 2.5 Flash to generate a resume summary, score, and feedback |
| Career support | Career-related guidance | Presents available career-development information and skill-related guidance |
| Learning support | Roadmap interface | Presents predefined learning templates |
| Practice | Quiz generation | Uses Gemini to generate practice questions |
| Mentoring | Mentor chatbot | Uses Gemini for conversational guidance |
| Output | Application interface | Displays results and guidance for student review |

The workflow combines model-generated responses with conventional application functionality. In particular, the roadmap interface currently relies on predefined templates rather than a dynamically generated sequence based on verified skill gaps. The displayed outputs should therefore be interpreted according to the implementation of each feature

### *C. Preliminary Response-Time Observations*

TABLE III
. EXPLORATORY SINGLE-RUN UI RESPONSE-TIME OBSERVATIONS

| Operation | Runs (N) | Observed time (s) |
|---|---|---|
| Resume analysis | 1 | 69.391 |
| Quiz generation | 1 | 26.431 |
| Mentor response | 1 | 11.889 |
| Learning roadmap display | 1 | 12.797 |
| AI Technical Interview display | 1 | 11.433 |

Each value comes from a single run, not a typical-performance estimate. Timings may include frontend work, network communication, model response time where applicable, and other UI activity. The roadmap and mock-interview timings shouldn't be read as direct LLM latency without request-level evidence of model calls. With only one observation per operation, mean latency, variability, and high-percentile latency can't be reliably estimated—repeated controlled measurements are needed.

### *D. Exploratory Keyword-Based Baseline evaluation*

TABLE IV
: EXPLORATORY EVALUATION OF THE KEYWORD-BASED BASELINE AND GEMINI 2.5 LASH

| Evaluation Metric | Keyword-Based Baseline | Gemini 2.5 Flash |
|---|---|---|
| Evaluated resume profiles | 20 | 20 |
| Mean Absolute Error (MAE) | 17.95 | 10.20 |
| Root Mean Squared Error (RMSE) | 24.83 | 14.60 |
| Pearson correlation | 0.607 | 0.820 |
| Spearman correlation | 0.4481 | 0.790 |
| Macro F1-score for skill-gap detection | 0.409 | 0.760 |

MAE and RMSE measure prediction errors, while Pearson correlation assesses linear association between predicted and reference scores. Macro F1-score evaluates skill-identification performance. These metrics provide a preliminary baseline, not proof of real-world accuracy. Spearman correlation is omitted pending verification, and no claim that Gemini 2.5 Flash outperforms the baseline is made without completed benchmark results.

### *E. Example Outputs and Interface Observations*

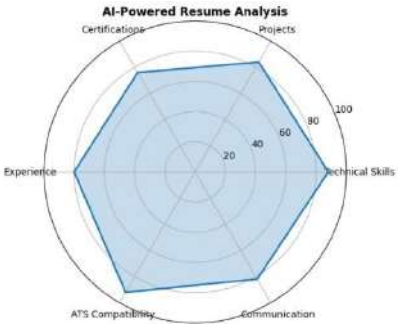



Fig. 7. AI-Powered Resume Analysis

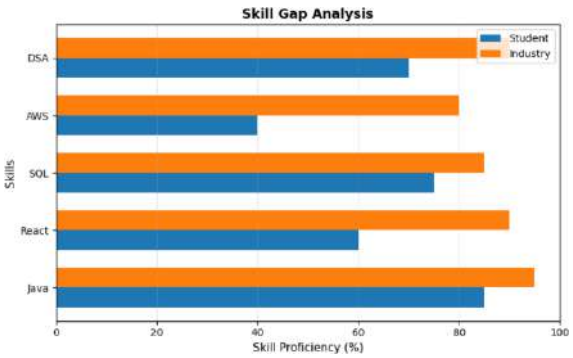



Fig. 8. Skill Gap Analysis

Figure 7 shows the AI resume analysis output covering technical skills, projects, certifications, experience, ATS

compliance, and communication. It gives a resume overview along with recommendations for improvement..

Figure 8 shows the skill-gap interface and the skill-gap information it displays. A comparison against role-specific proficiency thresholds should only be described if those thresholds and the comparison procedure are actually defined and implemented.

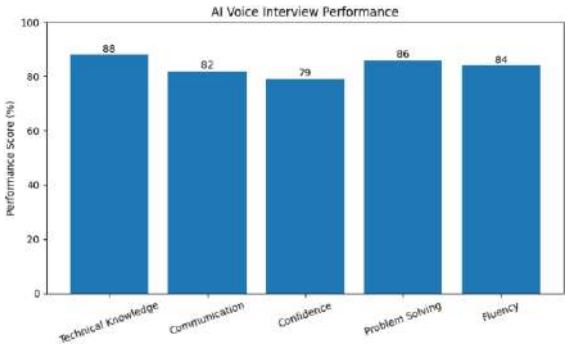



Fig. 9. AI Voice Mock Interview Performance

*Figure 9 Mock-interview interface output. Describe only the transcript, prompts, feedback, or scores actually produced and retained by the implementation; remove claims about confidence, fluency, or technical scoring unless measured.*

The application uses a React 19 and Vite frontend, a Node.js and Express backend, and PostgreSQL for data persistence through Sequelize. Gemini 2.5 Flash is integrated through backend API routes for resume analysis, quiz generation, and mentor conversations. The frontend sends requests to the corresponding backend endpoints, which process the input and return model-generated responses for display. Other features, including the learning roadmap, use predefined templates or conventional application logic. The system therefore combines LLM-assisted functions with conventional web-application components rather than relying on an LLM for every feature.

## VII. IMPACT AND APPLICATION

AI Student Guide Hub is meant to help engineering students prepare for their careers by pulling career resources and a few AI-assisted tools into one web app. It offers resume analysis, quiz generation, conversational mentoring, learning resources, and career-development activities, so students don't have to jump between separate services. The idea is to help them review their qualifications, explore career-related learning, and spot areas where they could improve. Google Gemini 2.5 Flash is built into the resume-analysis, quiz-generation, and mentor-chatbot workflows, while the learning-roadmap interface still uses predefined templates instead of dynamically generated personalized pathways. These features give students one place to access career-prep resources, but whether they actually improve learning outcomes, career readiness, or placement success hasn't been shown through a formal user study yet. The practical contribution is really the integration: bringing AI-assisted guidance and conventional career-preparation features together in a single application. That approach could be useful for colleges, educational institutions, and online learning platforms that want to organize career-oriented resources and practice activities. Still, the quality of generated feedback, the reliability of skill-related guidance, and the usability of the platform all need systematic evaluation before we can say how effective it is. Future work could look at institution-specific learning resources, multilingual support, and broader applications for employee skill development. Before wider deployment, more studies should assess recommendation quality, assessment reliability, privacy and security safeguards, and possible bias in AI-generated outputs. Evaluation with real student data, proper consent, and independently assessed reference labels would help clarify the platform's practical value and limitations.

## VIII. LIMITATIONS AND FUTURE WORK

AI Student Guide Hub offers several career-preparation features, but its recommendations depend on the accuracy and completeness of student profiles and extracted resume data. Missing or misinterpreted skills may affect guidance. Google Gemini 2.5 Flash supports resume analysis, quiz generation, and conversational mentoring, while learning roadmaps currently rely on predefined templates rather than dynamically personalized pathways.

A major limitation is the lack of comprehensive empirical validation. The preliminary benchmark uses synthetic student profiles and author-defined reference scores rather than independent expert assessments. Future work should evaluate diverse student profiles using validated reference labels, compare keyword-based, TF-IDF, and embedding-based approaches with appropriate metrics and repeated trials, and conduct user studies to assess recommendation quality, usability, and potential bias.

The platform uses PostgreSQL for persistent storage and Sequelize for database interactions, but its security and privacy have not been formally validated. Future work should assess password protection, encryption-key management, database access controls, and sensitive data shared with external AI services. Integration with reliable, regularly updated labour-market and educational resources is also needed to reflect changing industry requirements, internships, and certifications. Repeated testing should evaluate response times, request failures, and AI-output consistency under documented conditions.

Future enhancements may extend the platform to intermediate MPC and BIPC students through age-appropriate learning resources and career exploration. Other proposed features include mobile applications, voice-based coaching, interactive mock interviews, verified job and certification integrations, and dynamically generated learning pathways. These remain future developments until implemented and evaluated. The long-term goal is to support students from early career exploration to internship and placement preparation through learning materials, projects, certifications, coding exercises, and interview practice. However, improvements in learning outcomes, career readiness, and placement success must be established through systematic technical testing, independent assessments, and user studies to determine the platform's reliability, practical usefulness, security, and suitability for broader educational deployment.

## REFERENCES


[1] C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, Art. no. e1355, 2020, doi: 10.1002/widm.1355.

[2] M. Brindha, Karthikeyan M., Jessinth T., Karthe A. S., and Dhanush A., "AI-Powered Resume Optimizer and Job Recommendation System," *Journal of Emerging Technologies and Innovative Research*, vol. 12, Apr. 2025.

[3] ] Z. Guan, J.-Q. Yang, Y. Yang, H. Zhu, W. Li, and H. Xiong, "JobFormer: Skill-Aware Job Recommendation with Semantic-Enhanced Transformer," *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 1, Art. no. 18, 2025.

[4] E. Masciari, A. Umair, and M. H. Ullah, "A Systematic Literature Review on AI-Based Recommendation Systems and Their Ethical Considerations," *IEEE Access*, vol. 12, pp. 121223–121241, 2024.

[5] H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu, "Large Language Models for Education: A Survey," arXiv:2405.13001, 2024.

[6] S. Wang *et al*., "Large Language Models for Education: A Survey and Outlook," *IEEE Signal Processing Magazine*, vol. 42, no. 6, pp. 51–63, 2025.

[7] G. M. Dhananjaya, R. H. Goudar, A. A. Kulkarni, V. N. Rathod, and G. S. Hukkeri, "A Digital Recommendation System for Personalized Learning to Enhance Online Education: A Review," *IEEE Access*, vol. 12, pp. 34019–34041, 2024, doi: 10.1109/ACCESS.2024.3369901.

[8] Z. Ersozlu, S. Taheri, and I. Koch, "A Review of Machine Learning Methods Used for Educational Data," *Education and Information Technologies*, vol. 29, 2024.

[9] P. Singla and V. Verma, "An Intelligent Job Recommendation System Based on Semantic Embeddings and Machine Learning," *Journal of Information Systems Engineering and Management*, vol. 10, no. 5s, 2025, doi: 10.52783/jisem.v10i5s.681.

[10] L. Wu *et al*., "A Survey on Large Language Models for Recommendation," *World Wide Web*, vol. 27, Art. no. 60, 2024, doi: 10.1007/s11280-024-01291-2.

[11] F. Trujillo, M. Pozo, and G. Suntaxi, "Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction," *Journal of Technology and Science Education*, vol. 15, no. 1, pp. 162–185, 2025, doi: 10.3926/jotse.3124.

[12] Y. Ma, R. Ouyang, X. Long, Z. Gao, T. Lai, and C. Fan, "DORIS: Personalized course recommendation system based on deep learning," *PLOS ONE*, vol. 18, no. 6, Art. no. e0284687, 2023, doi: 10.1371/journal.pone.0284687.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[14] T. B. Brown *et al*., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[15] A. José-García *et al*., "C3-IoC: A Career Guidance System for Assessing Student Skills using Machine Learning and Network Visualisation," *International Journal of Artificial Intelligence in Education*, vol. 33, pp. 1092–1119, 2023, doi: 10.1007/s40593-022-00317-y.

[16] Z. Xiong, H. Li, Z. Liu, Z. Chen, H. Zhou, W. Rong, and Y. Ouyang, "A Review of Data Mining in Personalized Education: Current Trends and Future Prospects," *Frontiers of Digital Education*, vol. 1, no. 1, pp. 26–50, 2024, doi: 10.1007/s44366-024-0019-6.

[17] F. Ricci, L. Rokach, and B. Shapira, Eds., *Recommender Systems Handbook*, 3rd ed. Cham, Springer, 2022.

[18] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008.

[19] M. Miao and W. Holmes, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023.

# A Hybrid Machine Learning Framework for Hydropower Generation Forecasting: A Case Study of the Samanalawewa Power Station, Sri Lanka

A. N. A. D. Zoysa
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
nadzoysa@students.nsbm.ac.lk

T. A. H. Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Accurate short-horizon forecasts of hydropower generation support grid dispatch and reservoir management, and they matter most in power systems that lean heavily on hydropower, such as Sri Lanka's. The 120 MW Samanalawewa plant is a hard case, because karstic leakage and scheduled irrigation releases weaken the link between rainfall and output. Earlier studies at the site worked at monthly or quarterly resolution. This paper forecasts hourly generation with a sequential hybrid in which a regression tree learns and corrects the residual error of an artificial neural network, tested against each model alone. Ten years of plant output from the national dispatch dashboard were paired with hourly rainfall, temperature and humidity for five catchment stations. Models were evaluated with walk-forward cross-validation, a sealed chronological hold-out set and a persistence baseline. The first result is negative: with weather inputs alone, all three models scored a negative Nash–Sutcliffe efficiency (NSE), doing worse than a constant mean. Adding the plant's own recent generation changed this. The neural network and hybrid reached an NSE near 0.87 under cross-validation and 0.85 on the hold-out set, and the hybrid cut the persistence baseline's error by 29%, with the best score on every metric. Its gain over the neural network is small but statistically significant, and the correcting tree needs only three levels. Most of the remaining error falls in hours of sharp change, and features describing these ramps cut 10-day recursive error from 0.438 to 0.274. Because multi-step forecasts are built recursively, error grows with lead time, and the model suits forecasts up to about a day ahead. We conclude that weather data alone cannot forecast hourly generation at this plant and that reservoir water level should be added in future work.**

***Keywords—hydropower forecasting, hybrid machine learning, artificial neural network, regression tree, Samanalawewa***

## I. Introduction

Hydropower supplies roughly 17% of the world's electricity, more than every other renewable source combined [1]. What sets it apart from wind and solar is control: a reservoir plant can raise or lower its output within minutes to follow demand, which makes it a stabilising anchor for grids absorbing more variable generation each year [2]. That flexibility is only useful if operators know what the plant will produce, since forecasts drive dispatch scheduling and reservoir management, and in a hydro-dependent system a wrong forecast has to be covered by expensive thermal generation at short notice [3], [4].

Sri Lanka depends heavily on hydropower, which supplied 40% of its electricity in 2023 and remains the cheapest source in the national mix [5]. That dependence is becoming harder to manage as inflows fall with rising temperatures [6] and rainfall in the mountain catchments grows erratic [7]. Samanalawewa, a 120 MW plant, is a hard case even by these standards. Its output does not track rainfall, for two reasons specific to the site. The reservoir loses 2.1 to 2.8 $m^3/s$ continuously through karstic leakage that bears little relation to water level [8], and the plant must release water downstream for irrigation on an agricultural schedule, costing an estimated 30 GWh a year [9]. Both sever the link between rainfall and power, and the measured correlation between the two at this site is only 0.30 to 0.65 [10].

Three earlier studies forecast generation here, each improving on the last without closing the problem. Khaniya et al. [11] used monthly rainfall alone; Ekanayake et al. [12] added temperature and evaporation at quarterly resolution; Jayasinghe et al. [13] reached the highest accuracy with a regression tree that had to grow to depth ten and risked overfitting. None worked at hourly resolution, used humidity, or combined models, and hourly is the resolution that intraday dispatch and day-ahead scheduling actually need.

This paper forecasts hourly generation at Samanalawewa with a hybrid pairing a neural network and a regression-tree correction stage, tested against each model alone. Section II covers related work, Section III the study area and data, Section IV the methodology, Sections V and VI the results and discussion, and Section VII concludes.

## II. Related Work

Neural networks are the standard tool for hydropower forecasting. They approximate the non-linear relationship between climate and generation without a physical catchment model, and they have been applied across many sites and input types [3], [14]. Sessa et al. [15] found them outperforming empirical baselines at run-of-river plants across three countries. The recurring lesson is that input choice matters more than architecture [16], [17]. Two weaknesses limit them here, though: they are opaque, so an operator cannot see which inputs drive a forecast [18], and they lose accuracy on a single catchment with a short record [19]. Recurrent models such as LSTM handle temporal structure better [20] but need multi-

basin training data that a single reservoir cannot supply [21], which rules them out for a site-specific study.

Regression trees sit at the other end of the trade-off. A CART model partitions the inputs into readable rules [22], but when predictors correlate weakly with the target it grows deep and fits noise. That is not hypothetical at Samanalawewa, as the work below shows. Quantitative modelling at the site began with linear statistics. Dabare et al. [10] applied a linear framework to 26 years of records, found rainfall-to-generation correlations of only 0.30 to 0.65, and recommended non-linear methods.

Three machine learning studies followed. Khaniya et al. [11] trained the first neural network on monthly rainfall and reached $R^2$ of 0.83. Ekanayake et al. [12] widened the inputs to rainfall, temperature, and evaporation and switched to Gaussian Process and Support Vector Regression, reaching a correlation of 0.94 at quarterly resolution. Jayasinghe et al. [13] applied a regression tree to monthly data for a training $R^2$ of 0.96, but the tree needed depth ten and the authors flagged the overfitting risk. None worked at hourly resolution or used humidity, though two noted that missing variables limited them.

Combining models to offset each other's weaknesses is well established. In residual correction, a primary model predicts the target and a second learns the error it leaves [23], [24]. Parviz et al. reported a 44% reduction in mean squared error from this design, and a review of 66 reservoir studies reached the same general conclusion [25]. The design fits this problem: a neural network can carry the broad mapping while a shallow tree corrects the structured residual left by leakage and releases, keeping the correction readable. No study at Samanalawewa has tested this combination.

## III. Study Area and Data

The Samanalawewa Hydropower Project is a 120 MW plant on the Walawe River in southern Sri Lanka, fed by a 359 km² catchment on the wet-zone side of the central highlands. A well-watered catchment feeding a single reservoir should be straightforward to model, but two features complicate it. The reservoir sits partly on karstic limestone and loses 2.1 to 2.8 m³/s continuously through it, regardless of water level [8]. The plant also releases water downstream for irrigation on a schedule set by agricultural need rather than power demand, costing an estimated 30 GWh a year [9]. Together these break the direct link between rainfall and generation that a weather-driven forecast relies on.

Hourly generation is not published directly. The Public Utilities Commission of Sri Lanka records plant-level output at 15-minute intervals on its public dashboard [26], and these readings were collected from January 2016 to December 2025 and summed to hourly totals. This gave a continuous series of about 78,500 hourly observations in megawatts, the target variable for all models.

Climate inputs came from the Open-Meteo historical weather API [27], providing hourly reanalysis at roughly 11 km grid resolution. Three variables were retrieved hourly over the same period: rainfall, temperature, and relative humidity. Humidity is included because it governs how much rainfall is lost to evaporation before reaching the reservoir, and because the two most recent studies at the site identified it as a variable they lacked [11], [13]. Data were drawn for five of the six rain-gauge sites used in earlier work: Belihuloya, Balangoda, Alupola, Detanagalla, and Nagarak Estate [10], [12]. Their arrangement is shown in Fig. 1. The generation and climate series were aligned on a shared hourly timestamp and merged into a single dataset spanning the full period.

## IV. Methodology

### A. Feature engineering

Generation at a given hour depends on more than that hour's weather. Rain takes time to reach the reservoir, and output follows the plant's recent operating state as much as current conditions. The raw inputs were expanded into the feature blocks in Table I. Every lag and rolling window is computed on values shifted back one hour, so the features for hour t use nothing recorded after it. Station rainfall is lagged up to 72 hours, and a 168-hour rolling sum tracks how wet the catchment already is. Sine and cosine encodings of the hour and the day of the year carry the daily dispatch rhythm and the monsoon season. The generation block enters only the second configuration. The tree models also receive the one-hour change and 24-hour standard deviation of generation, since a tree splits on thresholds and gains from an explicit rate of change, while the ANN can build such terms in its hidden layers. The ramp variant adds seven features for hour-to-hour movement, for reasons set out in Section V.D.

Fig. 3. Rain-gauge stations across the Samanalawewa catchment (adapted from Dabare et al., 2020).

TABLE I. Feature Blocks

| *Block* | *Features* | *Count* |
|---|---|---|
| Station weather | Rainfall, temperature, humidity at five stations | 15 |
| Rainfall history | Lags of 1, 3, 6, 12, 24, 48, 72 h and 168 h sum, per station | 40 |
| Catchment climate | Mean temperature, mean humidity, total rainfall, 168 h total rainfall | 4 |
| Cyclical time | Sin and cos of hour-of-day and day-of-year | 4 |
| Generation history | Lags of 1, 3, 6, 12, 24, 48, 72 h; 24 h and 72 h means | 9 |
| Tree-only dynamics | 1 h change, 24 h standard deviation | 2 |
| Ramp | Lags of 25, 26, 47, 49 h; day-on-day change; 24 h maximum and minimum | 7 |

The weather-only set has 63 features, the augmented set 72 for the ANN and 74 for the trees, and the ramp variant 79 and 81.

Short gaps were filled by linear interpolation; gaps longer than three consecutive hours were dropped, but only after the rolling features were computed, so no window was ever summed across a gap. Inputs were scaled to [0, 1] with min-max normalisation. The series was split chronologically, 80% for training and 20% for testing, which keeps the test set strictly in the future and prevents leakage.

*B. Two input configurations*

The models were run under two input sets, and the contrast between them is the core of the experiment. The first uses weather only: the lagged and rolling climate features, with no information about past generation. This tests the assumption behind every prior study at the site. The second adds lagged generation, which stands in for what the weather record omits: reservoir level, leakage, and release timing. Generation lags were computed within each training fold only, preventing leakage. In recursive multi-step forecasts, weather is read from the reanalysis record while generation lags are fed back from the model's own prior predictions.

*C. Models*

Three models were trained under each configuration. The ANN is a feed-forward network with two hidden layers (50 and 25 units) and tanh activation, trained with Adam and L2 regularisation and early stopping, tuned by five-fold walk-forward cross-validation. The CART is a regression tree minimising squared error at each split, with depth tuned from 2 to 20; it selected depth 7. The hybrid combines the two in sequence, the tree correcting the network's error rather than competing with it. The network makes the primary prediction:

$$\hat{y}_{ANN},\ t = f(X_t)$$

Its residual on the training data is the part it could not explain:

$$e_t = y_t - \hat{y}_{ANN},\ t$$

A tree is trained to predict that residual from the same inputs:

$$\hat{e}_{CART},\ t = g(X_t)$$

and the final forecast adds the correction back:

$$\hat{y}_{hybrid},\ t = \hat{y}_{ANN},\ t + \hat{e}_{CART},\ t$$

Because the tree works on residuals rather than the full signal, it has far less variance to fit and stayed shallow, reaching its best fit at depth 3 against the standalone tree's depth 7.

The ANN's L2 penalty ($\alpha = 0.001$) and layer sizes came from a grid search over 15 candidates on five expanding-window folds. Tree depth was searched from 2 to 20 inside each fold, so the cross-validation score already pays for that choice, and the residual tree had its own depth search. In cross-validation the whole pipeline was refitted on each fold's training window and scored on the block after it.

*D. Baselines and evaluation*

Two references anchor the results. A naive persistence baseline sets the next hour equal to the last, and any useful model must beat it. Models were scored with four standard metrics: RMSE and MAE, which report error on the normalised [0, 1] scale; the Nash–Sutcliffe efficiency (NSE), where 1 is perfect, 0 matches the mean, and negative is worse than the mean; and $R^2$. NSE was read against the thresholds of Moriasi et al. [28], above 0.75 being very good. Tables II and III report walk-forward cross-validation as mean and standard deviation; Table IV reports the sealed 20% hold-out.

The hold-out set holds 15,712 hours and the training set 62,845. Differences between models were tested with the Diebold-Mariano test on squared-error loss. Hourly errors are strongly autocorrelated, so the variance uses a Newey-West correction with 12 lags.

*E. Multi-step forecasting*

The generation-lag configuration predicts one hour ahead directly, because its autoregressive inputs are the plant's actual output over the preceding hours. Forecasting further out needs those values before they exist. The forecast therefore runs recursively. The model predicts hour $t_0+1$ from real history, that prediction stands in for observed generation when hour $t_0+2$ is predicted, and the loop repeats out to 240 hours, or 10 days [32].

At each step, every generation-derived feature is recomputed from a working history that mixes observed and predicted values. This includes the lags, the rolling means, and the ramp and tree-only features. Weather and calendar inputs are never overwritten. They are read for each future hour as recorded, so generation is the only quantity the model has to supply for itself.

Forecasts were started from 120 origins spaced evenly across the test period, from March 2024 to December 2025, and the error at each lead time is averaged over all of them. Spreading the origins this way covers wet and dry seasons instead of one favourable stretch. Persistence, which carries the last observed value forward, and the weather-only hybrid, which needs no recursion, were scored from the same origins as references.

Rebuilding features outside the training pipeline is an easy place for silent errors, so the recursion was checked first. At the first step, the recursive forecast reads only real history and must match the stored one-step prediction. It did, with a largest difference of $2.2 \times 10^{-16}$, which is machine precision. From there, error compounds with each step, and Section V.E examines how far the forecast runs before it becomes unreliable.

## V. Results

*A. Weather-only configuration*

Table II reports walk-forward cross-validation for the three models trained on climate inputs alone. None worked. The mean NSE is negative for every model, between −0.06 and −0.09, meaning each predicted generation less accurately than a constant guess at the mean. The $R^2$ values, between 0.13 and 0.17, say the same from the other side. The three architectures

land within one standard deviation of each other, so the model made no difference. The inputs were the problem.

TABLE II. MODEL PERFORMANCE UNDER THE WEATHER-ONLY CONFIGURATION (MEAN ± STANDARD DEVIATION)

| *Model* | *RMSE* | *MAE* | *NSE* | *$R^2$* |
|---|---|---|---|---|
| ANN | 0.270 ± 0.075 | 0.199 ± 0.055 | −0.060 ± 0.179 | 0.174 ± 0.073 |
| CART | 0.276 ± 0.088 | 0.202 ± 0.065 | −0.072 ± 0.202 | 0.127 ± 0.075 |
| Hybrid | 0.275 ± 0.077 | 0.202 ± 0.056 | −0.092 ± 0.162 | 0.163 ± 0.063 |

Fig. 2 shows the failure directly: predicted values do not track the 1:1 line and spread into a cloud unrelated to the true output.

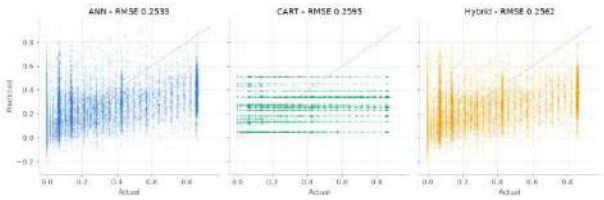



Fig. 4. Actual versus predicted generation under the weather-only configuration

### *B. Augmented configuration*

Adding lagged generation changed the outcome completely. In Table III the ANN and hybrid both reached NSE near 0.87 and $R^2$ near 0.88, with RMSE falling to about a third of the weather-only values. The hybrid was best on every metric, though its margin over the ANN is narrow. The standalone CART trailed both at NSE 0.840 but was stable across folds rather than an outlier.

TABLE III. MODEL PERFORMANCE UNDER THE AUGMENTED CONFIGURATION WITH GENERATION LAGS (MEAN ± STANDARD DEVIATION)

| *Model* | *RMSE* | *MAE* | *NSE* | *$R^2$* |
|---|---|---|---|---|
| ANN | 0.093 ± 0.024 | 0.060 ± 0.017 | 0.869 ± 0.046 | 0.876 ± 0.048 |
| CART | 0.103 ± 0.025 | 0.063 ± 0.018 | 0.840 ± 0.051 | 0.844 ± 0.052 |
| Hybrid | 0.091 ± 0.023 | 0.059 ± 0.016 | 0.874 ± 0.042 | 0.880 ± 0.044 |

Fig. 3 shows the scatter for this configuration, with points now sitting tightly along the 1:1 line.

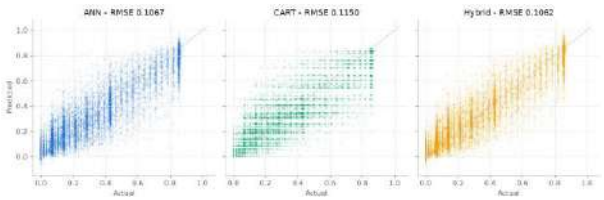



Fig. 5. .Actual versus predicted generation under the augmented configuration

### *C. Persistence baseline and residual correction*

A model built on generation lags must clear one bar: it has to beat naive persistence, where the next hour equals the last. Persistence is not weak here, reaching NSE 0.698, because hourly generation is smooth. The hybrid still cut its RMSE by 29%, from 0.149 to 0.106 (Table IV), so it is learning structure well beyond carry-forward.

TABLE IV. HYBRID VERSUS PERSISTENCE ON THE HELD-OUT TEST SET

| *Model* | *RMSE* | *MAE* | *NSE* | *$R^2$* |
|---|---|---|---|---|
| Persistence (last hour) | 0.149 | 0.092 | 0.698 | 0.721 |
| Hybrid | 0.106 | 0.071 | 0.846 | 0.847 |

The hybrid's design also held up. The standalone CART needed depth 7 to fit the generation signal, while the residual tree inside the hybrid reached its best fit at depth 3 with 8 leaves, since it works on a much smaller target. On the test set the correction moved the prediction the right way for 58.6% of hours, a slim but consistent majority. A Diebold-Mariano test confirms the hybrid's gain over the ANN is statistically significant, though smaller than fold-to-fold variation. Adding rate-of-change (ramp) features gave a further 2.5% refinement, and this ramp-augmented hybrid is used for the multi-step forecasts below.

Across folds, the hybrid beat the ANN in four of five and the standalone CART in all five. Its one loss came in the first fold, which has the shortest training window, so the residual stage seems to need some history before it finds real structure. The standalone tree also shows what the forecast leans on. Its top feature, the previous hour's generation, carries 89.9% of total importance.

### *D. Error by size of movement*

The remaining error is not spread evenly. Sorting test hours by the size of the hour-to-hour change, the 20% with the largest moves carry 57.8% of the squared error, and the top 5% carry 28.4% on their own. Flat hours are close to solved. When the actual change is regressed on the predicted change, the slope is 0.894, and the forecast reproduces only 79% of the real hour-to-hour volatility. In other words, the model under-reacts when output swings. The ramp features were added for this. Lags of 25 to 49 hours show what happened near the same clock hour on earlier days, and the 24-hour maximum and minimum show how far output usually moves.

### *E. Multi-step forecasting*

The results above are one-hour forecasts. Forecasting further requires the recursive procedure, and Table V gives the error at seven lead times. It roughly doubles from 1 to 24 hours, then rises more slowly to 10 days, where the ramp features hold RMSE at 0.274 against 0.438 without them. Most of that growth comes in the first six hours, after which error levels off for the rest of the day before drifting upward (Fig. 4).

TABLE V. RECURSIVE RMSE BY LEAD TIME (MEAN OVER 120 FORECAST ORIGINS)

| *Model* | *1 h* | *6 h* | *24 h* | *48 h* | *72 h* | *120 h* | *240 h* |
|---|---|---|---|---|---|---|---|
| Hybrid | 0.105 | 0.181 | 0.214 | 0.227 | 0.252 | 0.302 | 0.438 |
| Ramp-augmented hybrid | 0.101 | 0.188 | 0.211 | 0.223 | 0.251 | 0.261 | 0.274 |
| Persistence | 0.149 | 0.363 | 0.208 | 0.187 | 0.235 | 0.236 | 0.265 |

| *Model* | *1 h* | *6 h* | *24 h* | *48 h* | *72 h* | *120 h* | *240 h* |
|---|---|---|---|---|---|---|---|
| Hybrid | 0.105 | 0.181 | 0.214 | 0.227 | 0.252 | 0.302 | 0.438 |
| Weather-only hybrid | 0.275 | 0.241 | 0.279 | 0.260 | 0.255 | 0.271 | 0.259 |

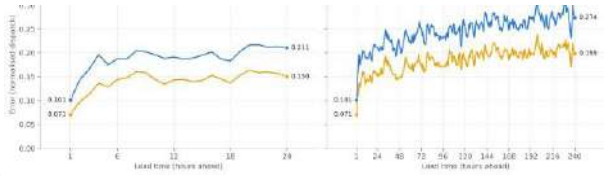


Fig. 6.

Recursive forecast error across the first 24 hours and 10-day horizon.

Table V sets the recursive model against two references. It beats persistence by a wide margin over the first 12 hours. At 24 and 48 hours persistence recovers, because the daily cycle makes yesterday's value a fair guess, and at 48 hours it is slightly ahead. The weather-only hybrid behaves differently. It never reads generation, so it has nothing to compound, and its error stays between 0.24 and 0.28 at every lead time. The recursive model first drops behind it at 59 hours, the two swap places over the next few days, and by 10 days the weather-only model is clearly better. One caution applies. The recursion reads weather from the reanalysis record, not from a forecast, so a live system would see larger errors and reach that crossover sooner.

## VI. Discussion

The clearest result is negative: hourly climate data cannot forecast generation here. Three different learners behaved the same way, so this is a limit of the inputs, not the models. The site explains why. The karstic leakage [8], and irrigation releases [9] sever the rainfall-to-generation link at the seasonal scale, as the weak correlation Dabare et al. [10] measured confirms. At hourly resolution the dominant driver is grid dispatch, which sets output minute to minute and never appears in a weather record either. Weather is thus doubly blind here: to the slow physical decouplers and to the fast dispatch signal. Earlier studies sidestepped it by working at monthly and quarterly scales, where aggregation smooths over the hour-to-hour disturbances leakage and releases create. At hourly resolution there is nowhere for that noise to hide.

Adding past generation fixed the problem because the plant's recent output carries what the weather cannot. The last few hours reflect the reservoir level, the leakage draw, and any releases underway, folded into one observed signal. The persistence comparison shows this is real work, not echoing the previous hour: persistence alone reaches NSE 0.698, and the hybrid cut its error by a further 29%. The model is not predicting generation from weather, which the first experiment ruled out. It predicts the next step of a process from its own trajectory, with weather as a secondary input.

The tree depths support the design. The standalone CART needed depth 7 to fit the signal, in line with the deep trees earlier work reported under weak correlations [13], [22]. The residual tree reached its best fit at depth 3, because once the network removes the bulk of the signal the tree has a smaller, simpler target and does not have to overreach. The recursive analysis sets the practical limit. Error rises quickly over the first six hours, then levels off for the rest of the day, which keeps the forecast usable a day ahead. It is a next-day tool, suited to intraday dispatch and day-ahead scheduling but not medium-range planning. The study has gaps. Without a model trained on generation lags alone, we cannot say how much weather adds once past output is known. The residual tree was fitted on the ANN's in-sample residuals, not out-of-fold ones, and no gradient-boosted, SARIMA, or same-hour-yesterday baseline was tested. Future work should close these gaps and add electricity demand data to capture dispatch, along with multi-step-ahead modelling and direct measurement of leakage and releases.

The ramp-augmented hybrid now runs in a web forecasting system built on FastAPI, React and PostgreSQL. The system collects data on a schedule and issues a forecast every hour. A parity test checks that the served model gives the same test-set predictions as the research notebook, and 672 automated tests cover the system.

## VII. Conclusion

This paper forecast hourly generation at the Samanalawewa plant with a hybrid in which a regression tree corrects the residual error of a neural network. The main finding is negative. Hourly rainfall, temperature and humidity from five catchment stations could not forecast generation here: every model scored worse than a constant mean, and changing the learner made no difference. Karstic leakage and irrigation releases weaken the rainfall link at the seasonal scale, and at hourly resolution output follows grid dispatch, which no weather record contains. Weather data alone should not be used to forecast generation at this plant.

Adding the plant's own recent output resolved the problem. The neural network and hybrid reached an NSE near 0.87 under cross-validation, and on the hold-out set the hybrid cut the persistence baseline's error by 29%. The hybrid was best on every metric. Its margin over the neural network is narrow but statistically significant, and it comes from a residual tree only three levels deep, so the correction stays small enough to read. Most of the remaining error falls in hours of sharp change. Ramp features aimed at those hours improved the one-hour forecast by 2.5% and held the 10-day error far below that of the base model.

Because the model forecasts each step from its own recent path, multi-step forecasts are built recursively and error grows with lead time. The model clearly beats persistence over the first 12 hours and stays usable for about a day, which suits intraday dispatch and day-ahead scheduling. Beyond two to three days, a weather-only model is the better choice.

The clearest next step is to give the model the physical variables it lacks, starting with reservoir water level, which reflects stored energy directly and would likely extend the forecast horizon. A model trained on generation lags alone would also show how much weather adds once past output is known, and demand data could capture dispatch more directly.

## DATA AVAILABILITY

The hourly generation data used in this study are publicly available from the Public Utilities Commission of Sri Lanka electricity dispatch dashboard [26]. The weather data were obtained from the Open-Meteo historical weather API [27]. No proprietary or restricted data were used. The processed dataset and code are available from the corresponding author on reasonable request.

## CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

## REFERENCES


[1] IEA, "Hydropower Special Market Report," IEA, Paris, 2021. [Online]. Available: https://www.iea.org/reports/hydropower-special-market-report

[2] B. Tarroja, A. AghaKouchak, and S. Samuelsen, "Quantifying climate change impacts on hydropower generation and implications on electric grid greenhouse gas emissions and operation," Energy, vol. 111, pp. 295–305, 2016, doi: 10.1016/j.energy.2016.05.131.

[3] N. E. Benti, M. D. Chaka, and A. G. Semie, "Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects," Sustainability, vol. 15, no. 9, p. 7087, Apr. 2023, doi: 10.3390/su15097087.

[4] Broman et al., "How Hydropower Operations Mitigate Flow Forecast Uncertainties to Maintain Grid Services in the Western US." Accessed: Jun.27, 2026. [Online]. Available: https://meetingorganizer.copernicus.org/EGU23/EGU23-16887.html

[5] Ceylon Electricity Board, "ANNUAL REPORT 2023," CEB, 2023. [Online]. Available: https://www.ceb.lk/publication-media/annual-reports/143/en

[6] G. Chen and G. D. Costa, "Climate Change Impacts on Water Resources Case of Sri Lanka," Environ. Ecol. Res., vol. 5, no. 5, pp. 347–356, Aug. 2017, doi: 10.13189/eer.2017.050504.

[7] G. Naveendrakumar, M. Vithanage, H.-H. Kwon, M. C. M. Iqbal, S. Pathmarajah, and J. Obeysekera, "Five Decadal Trends in Averages and Extremes of Rainfall and Temperature in Sri Lanka," Adv. Meteorol., vol. 2018, pp. 1–13, Dec. 2018, doi: 10.1155/2018/4217917.

[8] K. Laksiri, J. Gunathilake, and Y. Iwao, "A Case Study of the Samanalawewa Reservoir on the Walawe River in an Area of Karst in Sri Lanka," in Sinkholes and the Engineering and Environmental Impacts of Karst, San Antonio, Texas, United States: American Society of Civil Engineers, Sep. 2005, pp. 253–262. doi: 10.1061/40796(177)27.

[9] Japan International Cooperation Agency (JICA), "Samanalawewa Hydroelectric Power Project (I) (II) (III) and Samanalawewa Hydroelectric Project (Reservoir Remedial Works)," Japan Water Agency, Colombo, Sri Lanka, 2006.

[10] G. Dabare, M. B. Gunathilake, N. Miguntanna, K. Laksiri, and U. Rathnayake, "Climate Variation and Hydropower Generation in Samanalawewa Hydropower Scheme, Sri Lanka," Eng. J. Inst. Eng. Sri Lanka, vol. 53, no. 3, p. 19, Jul. 2020, doi: 10.4038/engineer.v53i3.7384.

[11] B. Khaniya, C. Karunanayake, M. B. Gunathilake, and U. Rathnayake, "Projection of Future Hydropower Generation in Samanalawewa Power Plant, Sri Lanka," Math. Probl. Eng., vol. 2020, pp. 1–11, Oct. 2020, doi: 10.1155/2020/8862067.

[12] P. Ekanayake, L. Wickramasinghe, J. M. J. W. Jayasinghe, and U. Rathnayake, "Regression-Based Prediction of Power Generation at Samanalawewa Hydropower Plant in Sri Lanka Using Machine Learning," Math. Probl. Eng., vol. 2021, pp. 1–12, Jul. 2021, doi: 10.1155/2021/4913824.

[13] J. Jayasinghe, P. Ekanayake, O. Panahatipola, C. I. Madhushani, and U. Rathnayake, "Forecasting the power generation at renewable power plants in Sri Lanka using regression trees," Results Eng., vol. 22, p. 102111, Jun. 2024, doi: 10.1016/j.rineng.2024.102111.

[14] S. D. Latif and A. N. Ahmed, "A review of deep learning and machine learning techniques for hydrological inflow forecasting," Environ. Dev. Sustain., vol. 25, no. 11, pp. 12189–12216, Nov. 2023, doi: 10.1007/s10668-023-03131-1.

[15] V. Sessa, E. Assoumou, and M. Bossy, "Modeling the climate dependency of the run-of-river based hydro power generation using machine learning techniques: an application to French, Portuguese and Spanish cases," presented at EMS Annual Meeting, Copenhagen, Denmark, Sep. 2019, hal-02302376

[16] T. Hammid, M. H. B. Sulaiman, and A. N. Abdalla, "Prediction of small hydropower plant power production in Himreen Lake dam (HLD) using artificial neural network," Alex. Eng. J., vol. 57, no. 1, pp. 211–221, Mar. 2018, doi: 10.1016/j.aej.2016.12.011.

[17] Universiti Kuala Lumpur, British Malaysian Institute, 53100 Gombak Selangor, Malaysia et al., "Flood Risk Prediction for a Hydropower System using Artificial Neural Network," Int. J. Recent Technol. Eng. IJRTE, vol. 8, no. 4, pp. 6177–6181, Nov. 2019, doi: 10.35940/ijrte.D5134.118419.

[18] V. K. Vidyarthi and A. Jain, "Does ANN really acquire the physics of the system? A study using conceptual components from an established water balance model," J. Hydroinformatics, vol. 25, no. 4, pp. 1380–1395, Jul. 2023, doi: 10.2166/hydro.2023.025.

[19] L. Zhou, "Perspective Chapter: Big Data and Deep Learning in Hydrological Modeling," in Earth Sciences, vol. 4, C. Guo, Ed., IntechOpen, 2024. doi: 10.5772/intechopen.1007640.

[20] Kisvari, Z. Lin, and X. Liu, "Wind power forecasting – A data-driven method along with gated recurrent neural network," Renew. Energy, vol. 163, pp. 1895–1909, Jan. 2021, doi: 10.1016/j.renene.2020.10.119.

[21] Kratzert, M. Gauch, D. Klotz, and G. Nearing, "HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin," Hydrol. Earth Syst. Sci., vol. 28, no. 17, pp. 4187–4201, Sep. 2024, doi: 10.5194/hess-28-4187-2024.

[22] W. Loh, "Classification and regression trees," WIREs Data Min. Knowl. Discov., vol. 1, no. 1, pp. 14–23, Jan. 2011, doi: 10.1002/widm.8.

[23] L. Parviz, K. Rasouli, and A. Torabi Haghighi, "Improving Hybrid Models for Precipitation Forecasting by Combining Nonlinear Machine Learning Methods," Water Resour. Manag., vol. 37, no. 10, pp. 3833–3855, Aug. 2023, doi: 10.1007/s11269-023-03528-7.

[24] M. Khodaverdi, "Forecasting future energy production using hybrid artificial neural network and arima model," MS, West Virginia University Libraries, 2018. doi: 10.33915/etd.4004.

[25] A. Ladanu, S. Akanmu, and J. Adeyemo, "Enhancing Artificial Neural Network with Multi-Objective Evolutionary Algorithm for Optimizing Real Time Reservoir Operations: A Review," Am. J. Water Resour., vol. 8, no. 3, pp. 118–127, May 2020, doi: 10.12691/ajwr-8-3-2.

[26] PUCSL, "Electricity Dispatch Data Dashboard," PUCSL. [Online]. Available: https://gendata.pucsl.gov.lk/generation-profile

[27] P. Zippenfenig, Open-Meteo.com Weather API. (Dec. 31, 2024). Zenodo. doi: 10.5281/ZENODO.7970649.

[28] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations," Trans. ASABE, vol.50, no. 3, pp. 885–900, 2007, doi: 10.13031/2013.23153.

# Assessing Associations Between Wayground AI Features and Intrinsic Motivation among Middle School Students in Online Education: A Quantitative Survey Study

V. L. Karunarathne
*Department of Computing*
*Esoft Uni*
*Kandy*
*Sri Lanka*
vindya.karunarathne@esu.lk

D. G. S. N. Samarakoon
*SLIIT Kandy Uni*
*Sri Lanka*
sathsarani.s@sliit.lk

***Abstract*—This study examines associations between perceived features of Wayground AI (an AI-enhanced online learning platform) and intrinsic motivation among middle school students. Adopting a positivist philosophy and deductive quantitative approach, data were collected via a structured Likert-scale questionnaire from 60 middle school students at a single national school in Sri Lanka who had used Wayground AI. Independent variables were gamification elements, adaptive AI feedback, willingness to learn, and self-efficacy; the dependent variable was intrinsic motivation. Reliability analysis confirmed acceptable internal consistency (Cronbach's α ranging from 0.78 to 0.89). Due to non-normal distributions, Spearman rank-order correlations and ordinal logistic regression were employed. All independent variables showed statistically significant positive associations with intrinsic motivation ($p < 0.01$). The ordinal regression model was significant ($\chi^2(4) = 48.72$, $p < 0.001$), with Nagelkerke pseudo-$R^2$ = 0.649. Willingness to learn emerged as the strongest predictor. The findings provide contextual empirical evidence of associations consistent with Self-Determination Theory in a middle-school online setting in a developing-country context. A conceptual framework linking platform features and learner dispositions is proposed. Limitations include the cross-sectional design, single-institution sample, and reliance on self-reports; causal claims are not warranted. Recommendations for practice and future longitudinal and multi-site research are discussed.**

***Keywords*—*Artificial Intelligence in Education, Intrinsic Motivation, Online Learning, Gamification, Adaptive Feedback, Self-Determination Theory, Middle School***

## I. INTRODUCTION

The rapid expansion of online learning has introduced persistent challenges in sustaining student engagement, particularly among middle school learners. Intrinsic motivation the internal drive to learn for interest, enjoyment, and personal satisfaction plays a critical role in academic persistence and deep learning [1], [2]. Unlike extrinsic motivation driven by rewards or external pressure, intrinsic motivation supports self-determined behavior and longer-term educational outcomes. When online environments fail to stimulate curiosity or provide meaningful feedback, students frequently experience declining participation, reduced confidence, and surface-level learning strategies. These problems are especially acute in the middle school years, a developmental period when learning habits and motivational orientations are still crystallizing. In many educational systems, including those in developing regions, the shift to remote and hybrid models has exposed gaps in both technological readiness and pedagogical design for sustaining motivation at a distance.

Artificial intelligence (AI) tools are increasingly integrated into educational platforms to personalize learning experiences. Wayground AI is one such commercial platform that incorporates gamification elements (points, badges, leaderboards, interactive quizzes) and adaptive AI feedback that adjusts task difficulty and provides real-time personalized guidance. These features are theoretically aligned with Self-Determination Theory (SDT), which posits that satisfaction of the needs for autonomy, competence, and relatedness fosters intrinsic motivation [1], [2]. Prior research indicates that gamification and adaptive systems can be associated with higher engagement; however, empirical evidence specifically examining associations between the features of Wayground AI and intrinsic motivation among middle school students in online settings remains limited, especially in South Asian and developing-country contexts such as Sri Lanka.

This paper reports a quantitative survey study that addresses this contextual gap. The primary aim is to examine the associations between perceived Wayground AI features (gamification elements and adaptive AI feedback), learner dispositions (willingness to learn and self-efficacy), and self-reported intrinsic motivation among middle school students in an online education setting. Specific objectives include: (1) describing the levels of the measured constructs; (2) analyzing bivariate associations between the independent variables and intrinsic motivation; (3) evaluating the joint predictive contribution of the four independent variables via ordinal logistic regression; and (4) proposing a conceptual framework grounded in the observed associations for educators and platform developers. The study contributes contextual empirical evidence from a middle-school sample in Sri Lanka and practical guidance while remaining explicit about the correlational nature of the design.

The contribution is incremental rather than strongly novel: it applies established SDT constructs and measurement approaches to a specific commercial AI platform and age group in an under-studied regional context. The value lies in localized evidence, simultaneous modeling of technological and dispositional predictors, and a transparent report of associations rather than causal claims of platform effectiveness.

## II. RELATED WORK AND THEORETICAL BACKGROUND

Self-Determination Theory provides a foundational framework for understanding intrinsic motivation, emphasizing the psychological needs of autonomy, competence, and relatedness [1], [2]. When learning environments support these needs, students are more likely to engage for inherent interest rather than external contingencies. Meta-analyses of gamification in learning have reported positive effects on engagement and, under certain design conditions, on indicators of intrinsic motivation, primarily through progress indicators, meaningful challenges, and social comparison mechanisms [3], [4]. Adaptive feedback systems, enabled by AI, are argued to support the need for competence by offering timely, personalized guidance that reduces uncertainty, scaffolds difficulty, and promotes a visible sense of progress [5], [15]. Real-time adaptation can help maintain an optimal challenge level conducive to flow and sustained engagement.

Willingness to learn reflects a learner's readiness and openness to engage with new material and has been linked to intrinsic motivation in e-learning contexts [6]. Self-efficacy, or perceived competence, further influences motivation by shaping students' beliefs about their ability to succeed at tasks and to recover from setbacks [7]. AI platforms that provide mastery experiences, constructive feedback, and visible progress indicators can therefore be associated with stronger self-efficacy and, indirectly, higher intrinsic motivation. The interplay among these factors suggests that motivational outcomes are co-determined by system design features and learner dispositions rather than by technology alone.

While the broader literature supports the motivational potential of AI-enhanced and gamified platforms, most empirical studies focus on higher education populations or well-established commercial systems such as Duolingo, Khan Academy, or adaptive tutoring systems [15], [16]. Research specifically examining Wayground AI (or closely comparable platforms) with middle school learners in online settings, particularly within South Asian educational contexts, remains scarce. Furthermore, few studies have simultaneously modeled both technological features (gamification, adaptive feedback) and learner-centered variables (willingness to learn, self-efficacy) as joint predictors of intrinsic motivation using ordinal regression on Likert-derived ordered outcomes. This study therefore addresses a clear contextual and modeling gap by testing a theory-driven set of associations in a middle-school online learning environment in Sri Lanka and by proposing an integrated conceptual framework grounded in the empirical associations.

Recent work on AI in K-12 education has highlighted both opportunities and risks for motivation, including the potential for over-reliance on extrinsic game elements and the importance of teacher mediation [17], [18]. Comparative studies of adaptive versus non-adaptive online environments have reported mixed results depending on implementation quality and learner characteristics [19]. By situating Wayground AI within this literature and by explicitly treating the design as correlational rather than experimental, the present study aims to contribute a transparent, localized data point rather than a claim of general causal superiority.

## III. METHODOLOGY

### *A. Research Design and Philosophy*

The study adopted a positivist research philosophy and a deductive approach, consistent with the goal of testing associations derived from established motivational theories. A mono-method quantitative design using a cross-sectional survey strategy was employed. This approach enables standardized measurement of psychological constructs and statistical hypothesis testing within the sampled institutional population [8], [9]. The choice of a survey strategy was justified by the need to collect comparable data efficiently from school-aged participants while minimizing disruption to instructional time. A mono-method quantitative design was preferred because the research questions centered on measurable associations rather than exploratory qualitative insight.

### *B. Participants and Sampling*

The target population comprised middle school students in the IT stream at K/Al-Manar National School (Sri Lanka) who had been exposed to or actively used the Wayground AI learning platform during the academic term. Using the Krejcie and Morgan sample-size determination table [10] for a population of approximately 70 eligible students (recommended sample ≈ 59 for N = 70 at 95% confidence), a sample of 60 students was obtained via simple random sampling from the sampling frame. This method was selected to minimize selection bias. Alternative strategies such as stratified sampling by gender or prior achievement were considered but judged less necessary for the primary research questions given the modest population size.

Demographic information collected included age band (primarily 12–15 years), gender (approximately balanced), and self-reported frequency of online learning and Wayground AI use (weekly or more frequent for the majority of respondents). No students were excluded after consent; response rate within the sampling frame was high. The single-institution, modest sample size limits statistical power for detecting small effects and restricts generalizability beyond the local context; these constraints are acknowledged as important limitations.

### *C. Instrument and Variables*

A structured questionnaire administered via Google Forms measured four independent variables gamification elements, adaptive AI feedback, willingness to learn, and self-efficacy and the dependent variable intrinsic motivation. All constructs were operationalized with multiple 5-point Likert-scale items (1 = Strongly Disagree to 5 = Strongly Agree). Items were adapted from validated scales reported in the literature on gamification, adaptive learning, self-efficacy, and Self-Determination Theory [1]–[7] and were refined through pilot testing with a small group of students (n ≈ 8) to ensure linguistic clarity and age appropriateness. Content validity was supported by mapping

each item set to the corresponding theoretical construct and by alignment with the stated research objectives.

Construct scores were computed as the arithmetic mean of the respective items (listwise deletion for any incomplete responses; missing data were minimal). Willingness to learn and intrinsic motivation were treated as conceptually related but distinct: the former emphasizes readiness and openness to engage with new material, while the latter emphasizes interest, enjoyment, and personal satisfaction derived from the learning activity itself. Item wording was reviewed to minimize conceptual overlap. Exact item lists and scoring rules are available from the authors upon request to support reproducibility; space constraints preclude full reproduction here.

### *D. Data Analysis Procedures*

Data were organized in Microsoft Excel and analyzed using IBM SPSS Statistics. Descriptive statistics summarized demographic characteristics and the central tendency and dispersion of each construct. Internal consistency reliability was assessed via Cronbach's alpha; values above 0.70 were treated as acceptable. Normality of distributions was examined with both Shapiro–Wilk and Kolmogorov–Smirnov tests. Because all key variables departed significantly from normality ($p < 0.05$), non-parametric procedures were adopted: Spearman rank-order correlation for bivariate associations and ordinal logistic regression (proportional-odds model) for the multivariate prediction of ordered levels of intrinsic motivation.

Ordinal regression is appropriate when the dependent variable is ordered but intervals between categories cannot be assumed equal. The proportional-odds assumption was examined via the test of parallel lines; multicollinearity among predictors was assessed via variance inflation factors (all VIF < 2.5). Model fit is reported via the likelihood-ratio chi-square test and Nagelkerke pseudo-$R^2$ (interpreted as an approximate measure of model explanatory strength rather than as a direct analogue of ordinary least-squares $R^2$). Odds ratios and 95% confidence intervals are reported for predictors. These choices align with methodological recommendations for ordinal Likert-derived data that violate parametric assumptions [8], [11].

### *E. Ethical Considerations*

Given that participants were minors, ethical safeguards were prioritized. Informed consent procedures appropriate for school-based research were followed, including institutional approval and parental/guardian permission where required. Participation was voluntary, responses were collected anonymously via Google Forms (no IP logging retained), and data were used solely for the stated research purposes. No identifiable personal information was retained beyond what was necessary for analysis. The study design was observational and avoided experimental manipulation that could disadvantage any group of students.

## IV. RESULTS

### *A. Reliability and Descriptive Statistics*

All measurement scales demonstrated acceptable to good internal consistency (Table I).

TABLE I
DESCRIPTIVE STATISTICS AND RELIABILITY COEFFICIENTS

| Construct | Mean | SD | Cronbach's α | Items |
|---|---|---|---|---|
| Gamification Elements | 4.59 | 0.42 | 0.82 | 5 |
| Adaptive AI Feedback | 4.69 | 0.38 | 0.85 | 4 |
| Willingness to Learn | 4.55 | 0.45 | 0.81 | 4 |
| Self-Efficacy (Competence) | 4.48 | 0.51 | 0.78 | 5 |
| Intrinsic Motivation (DV) | 4.52 | 0.47 | 0.89 | 6 |

*Note: N = 60. All scales 1–5. Exact α values computed via SPSS reliability procedure.*

Mean scores for both the independent variables and the dependent variable were high (typically above 4.0 on the 5-point scale), indicating generally positive perceptions of Wayground AI features and relatively high self-reported intrinsic motivation among the participating students. Standard deviations were modest, suggesting reasonable agreement within the sample.

### *B. Correlation Analysis*

Spearman rank-order correlation coefficients indicated statistically significant positive associations between each independent variable and intrinsic motivation (all $p < 0.01$; Table II). Willingness to learn exhibited the strongest association ($\rho = 0.556$), followed by adaptive AI feedback ($\rho = 0.467$), gamification elements ($\rho = 0.441$), and self-efficacy ($\rho = 0.453$). These results provide empirical support for the hypothesized directional associations derived from Self-Determination Theory and prior work on AI-supported and gamified learning environments. The consistent positive signs across all predictors suggest that both technological features and learner dispositions are associated with higher intrinsic motivation in this sample.

TABLE II
SPEARMAN CORRELATIONS WITH INTRINSIC MOTIVATION

| Independent Variable | Spearman ρ | p-value |
|---|---|---|
| Willingness to Learn | 0.556 | < 0.001 |
| Adaptive AI Feedback | 0.467 | < 0.001 |
| Self-Efficacy | 0.453 | < 0.001 |
| Gamification Elements | 0.441 | 0.001 |

*Note: N = 60. Two-tailed tests.*

### *C. Ordinal Regression Results*

For ordinal regression, the continuous mean scores of intrinsic motivation were categorized into ordered levels (Low: ≤ 3.5; Moderate: 3.51–4.5; High: > 4.5) to produce a three-category ordered outcome suitable for the proportional-odds model; the majority of responses fell in the High category, consistent with the elevated mean. The overall model was statistically significant (likelihood-ratio $\chi^2(4) = 48.72$, $p < 0.001$). Nagelkerke pseudo-$R^2 = 0.649$ indicates substantial model explanatory strength relative to the null model. The test of parallel lines did not reject the proportional-odds assumption ($p > 0.05$). Variance inflation factors for all predictors were below 2.5, indicating no problematic multicollinearity.

TABLE III
ORDINAL LOGISTIC REGRESSION PREDICTING INTRINSIC MOTIVATION CATEGORY

| Predictor | B | SE | OR | 95% CI | p |
|---|---|---|---|---|---|
| Willingness to Learn | 1.416 | 0.397 | 4.12 | 1.89–8.97 | 0.001 |
| Self-Efficacy | 0.892 | 0.351 | 2.44 | 1.23–4.85 | 0.011 |
| Adaptive AI Feedback | 0.764 | 0.328 | 2.15 | 1.13–4.08 | 0.020 |
| Gamification Elements | 0.651 | 0.312 | 1.92 | 1.04–3.54 | 0.037 |

*Note: N = 60. OR = odds ratio. Threshold estimates omitted for brevity. Model $\chi^2(4) = 48.72$, $p < 0.001$; Nagelkerke $R^2 = 0.649$. Parallel-lines test non-significant.*

Examination of individual predictors (Table III) showed that willingness to learn was the strongest contributor (odds ratio ≈ 4.12, 95% CI [1.89, 8.97], $p < 0.01$), followed by self-efficacy, adaptive AI feedback, and gamification elements. All four variables retained statistically significant associations in the multivariate model. These findings indicate that higher levels of the measured predictors are associated with higher odds of belonging to a higher intrinsic-motivation category, controlling for the other predictors in the model. No mediation or moderation analyses were conducted; the model is purely predictive/associational.

## V. DISCUSSION

The results provide empirical support for positive associations between perceived Wayground AI features, learner dispositions, and intrinsic motivation among the sampled middle school students. The significant correlations and the overall ordinal regression model are consistent with predictions derived from Self-Determination Theory. Gamification elements and adaptive feedback are associated with higher reported motivation in a manner that can be interpreted as supporting needs for autonomy and competence (visible progress, personalized challenge). Simultaneously, willingness to learn and self-efficacy capture dispositional readiness that co-varies with motivational reports. Together, these findings portray intrinsic motivation in this context as an emergent outcome associated with the interaction between platform affordances and prepared, confident learners.

The finding that willingness to learn is the strongest predictor carries practical significance. It suggests that technology features alone are not sufficient correlates of elevated intrinsic motivation; learners who already possess readiness to engage interact more productively with AI features (or at least report higher motivation in conjunction with them). This implies that successful implementation may require complementary pedagogical efforts: educators should actively cultivate readiness through goal-setting, autonomy-supportive language, and preparatory activities rather than relying solely on the platform's inherent motivational affordances. The significant contribution of self-efficacy similarly indicates that platforms and teachers should emphasize progressive challenge and clear mastery feedback.

Importantly, the design is cross-sectional and associational. The data do not support causal claims that Wayground AI "improves" or "fosters" intrinsic motivation; they show that students who report higher engagement with platform features and higher dispositional readiness also report higher intrinsic motivation at a single time point. Alternative explanations (common-method variance, selection effects, reverse causality, or unmeasured confounders) cannot be ruled out. Nagelkerke pseudo-$R^2$ is reported as a measure of model strength relative to the null and should not be interpreted as the proportion of variance explained in the ordinary least-squares sense.

Several limitations qualify interpretation. The cross-sectional design precludes causal inference and does not permit observation of motivational change over time. The sample of 60 students from a single institution limits statistical power and generalizability. Self-report measures remain susceptible to social-desirability bias and common-method variance. The study did not measure academic performance outcomes, so the translation of motivational associations into knowledge or skill gains remains an open question. Platform exposure was self-reported rather than logged. Future studies would benefit from multi-source data (platform usage logs, teacher ratings), larger multi-site samples, longitudinal or quasi-experimental designs, and explicit comparison with alternative AI platforms or non-AI conditions.

## VI. CONCLUSION AND RECOMMENDATIONS

This study concludes that perceived gamification elements, adaptive AI feedback, willingness to learn, and self-efficacy are positively associated with self-reported intrinsic motivation among the sampled middle school students using Wayground AI in an online setting. The combined ordinal model accounts for a substantial share of the ordered variation in motivation scores. On the basis of these associations, a conceptual framework is proposed in which AI platform features and learner psychological dispositions jointly relate to intrinsic motivation; the framework can guide instructional practice and further platform development while remaining clear about its correlational foundation.

Practical recommendations follow from the evidence with appropriate caution. Educational institutions considering Wayground AI (or similar platforms) should treat it as one component of a broader pedagogical strategy rather than a sufficient condition for motivation. Teachers require training in interpreting platform analytics. Learning tasks should be sequenced to provide progressive challenge and clear progress indicators. Deliberate attention should be given to fostering willingness to learn through orientation, autonomy-supportive framing, and connection of content to students' interests. These recommendations are interdependent.

Future research should employ longitudinal designs that track motivational trajectories, conduct quasi-experimental comparisons between AI-supported and traditional online conditions, expand samples across multiple institutions and demographic groups, and incorporate objective usage and performance metrics alongside self-reports. Comparative studies of alternative AI learning platforms would clarify the unique contribution of specific features. Mixed-methods designs could illuminate lived experiences behind the statistical associations. Such cumulative work will deepen the evidence base for how AI can be associated with sustained intrinsic motivation in online middle-school education.

## REFERENCES


[1] R. M. Ryan and E. L. Deci, “Intrinsic and extrinsic motivations: Classic definitions and new directions,” Contemp. Educ. Psychol., vol. 25, no. 1, pp. 54–67, 2000.

[2] R. M. Ryan and E. L. Deci, “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being,” Amer. Psychologist, vol. 55, no. 1, pp. 68–78, 2000.

[3] M. Sailer and L. Homner, “The gamification of learning: A meta-analysis,” Educ. Psychol. Rev., vol. 32, pp. 77–112, 2020.

[4] Z. Zainuddin, S. K. W. Chu, M. Shujahat, and C. J. Perera, “The impact of gamification on learning and instruction: A systematic review of empirical evidence,” Educ. Res. Rev., vol. 30, Art. no. 100326, 2020.

[5] W. Kim et al., “Adaptive learning systems and student motivation in digital environments,” Comput. Educ., vol. 148, 2020.

[6] U. Mustafa, A. Khan, and S. Rahman, “Willingness to learn and autonomous learning behaviors in online environments,” J. Educ. Technol. Syst., vol. 51, no. 4, pp. 412–430, 2023.

[7] Bandura, “Self-efficacy: Toward a unifying theory of behavioral change,” Psychol. Rev., vol. 84, no. 2, pp. 191–215, 1977.

[8] M. Saunders, P. Lewis, and A. Thornhill, Research Methods for Business Students, 8th ed. Harlow, U.K.: Pearson, 2019.

[9] J. W. Creswell and J. D. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th ed. Thousand Oaks, CA, USA: SAGE, 2018. R. V. Krejcie and D. W. Morgan, “Determining sample size for research activities,” Educ. Psychol. Meas., vol. 30, no. 3, pp. 607–610, 1970.

[10] U. Flick, Introducing Research Methodology: A Beginner’s Guide to Doing a Research Project, 2nd ed. London, U.K.: SAGE, 2020.

[11] E. Gray, Doing Research in the Real World, 4th ed. London, U.K.: SAGE, 2018.

[12] Fink, Conducting Research Literature Reviews: From the Internet to Paper, 5th ed. Thousand Oaks, CA, USA: SAGE, 2020.

[13] J. Wellington, Educational Research: Contemporary Issues and Practical Approaches. London, U.K.: Continuum, 2000.

[14] K. VanLehn, “The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems,” Educ. Psychologist, vol. 46, no. 4, pp. 197–221, 2011.

[15] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in Learning Analytics, J. A. Larusson and B. White, Eds. New York, NY, USA: Springer, 2014, pp. 61–75.

[16] W. Holmes, M. Bialik, and C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Boston, MA, USA: Center for Curriculum Redesign, 2019.

[17] N. Selwyn, “Should robots replace teachers? AI and the future of education,” Polity Press, 2019.

[18] S. Schiaffino, P. Garcia, and A. Amandi, “eTeacher: Providing personalized assistance to e-learning students,” Comput. Educ., vol. 51, no. 4, pp. 1744–1754, 2008.

# AI-Powered Ceiling Material Estimation from Plan Area Images

Kumuditha Ranasinghe
*Department of Software Engineering*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
ranasinghe24125@gmail.com

G. Perera
*Department of Computer & Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
gayanp@nsbm.ac.lk

***Abstract*—Ceiling material quantities require a measured installation area, a selected panel size, and an explicit waste allowance. This paper presents an AI-powered mobile prototype combining UNet++/EfficientNet-B4 segmentation, deterministic OpenCV geometry, a FastAPI backend, and a Flutter client. The evaluated inputs are floor-plan renders; performance on smartphone photographs of physical ceilings has not been established. The documented design separates interior detection from compactness classification. A label-decoding defect was identified in which white background pixels were assigned to FULL_COMPACTNESS in both training and evaluation masks. After correcting the labels and retraining, the recorded not independently recomputed, inference-mode comparison reports mean IoU 0.9440, pixel accuracy 0.9853, and Dice 0.9710 for raw argmax. Post-processing reduced reported mean IoU to 0.9211 while removing two residual fragments. The contribution is application integration and the analysis of label integrity and inference selection. Expert area and quantity validation remains pending; the available evidence does not establish estimation error below 10%.**

***Keywords—convolutional neural network, semantic segmentation, computational geometry, quantity takeoff, construction AI, mobile deep learning***

## I. Introduction

Ceiling installation requires material quantities that are consistent with the intended installation area, panel dimensions, and waste allowance. Translating image geometry into an auditable panel list is useful for contractors and suppliers, but segmentation quality alone does not establish purchasing accuracy. This work develops a mobile-connected prototype that separates learned region prediction from deterministic area conversion and panel-count calculation. Quantified reductions in customer reorders, cost overruns, and estimation time have not been measured.

Recent progress in computer vision has demonstrated that convolutional neural networks can segment image regions using established encoder–decoder architectures [1], [2]. However, ceiling-panel estimation also requires a valid scale, a selected product size, and an explicit waste allowance; segmentation alone does not supply these quantities. This paper addresses this integration task with a plan-image pipeline combining semantic segmentation, deterministic computational geometry, and a mobile application, and reports a recorded inference-mode comparison with explicit validation limits.

The contributions of this paper are: (1) an integrated UNet++/EfficientNet-B4 segmentation, geometry and panel-estimation prototype; (2) diagnosis and correction of a mask decoding defect that contaminated training and evaluation targets; and (3) a corrected-label comparison of raw inference, post-processing, and test-time augmentation. Real-world area and quantity validation is future work, rather than a completed contribution.

## II. Literature Review

Semantic segmentation research provides established encoder–decoder architectures such as UNet++ [1] and DeepLabV3+ [3]; results from other datasets do not directly establish ceiling-panel estimation accuracy. Ceiling defect detection research applies object detection (e.g., YOLOv5) to identify deterioration in already-installed ceilings [4] targeting maintenance rather than new-installation estimation. Commercial AI-powered quantity takeoff (QTO) platforms are not experimentally compared here no accuracy, timesaving, or affordability advantage over these platforms is claimed. DeepLabV3+ uses atrous separable convolution and an encoder–decoder structure [3], making it a relevant future segmentation baseline; it was not trained in this study. [7], [8]. BIM-integrated automated QTO requires a BIM model, whereas this prototype uses a plan image and an external scale. None of the reviewed systems combines CNN-based segmentation, deterministic geometry, and mobile deployment into a plan-image pipeline. CBAM [5] and scSE [6] are established attention modules; their use does not constitute a new segmentation algorithm.

## III. METHODOLOGY

The system follows a Design Science Research methodology, iteratively building and evaluating a prototype against defined accuracy and usability objectives. Fig. 1 shows the layered architecture: a Flutter presentation layer communicates with a FastAPI backend over HTTPS; the service layer separates segmentation orchestration from deterministic geometry and material estimation; a singleton model registry caches segmentation checkpoints and data layer stores job results. Neural inference runs on the backend, not on the smartphone. Software functionality is distinguished from uncompleted field measurement validation.

A deliberate architectural principle underlies this design: neural inference identifies interior regions and compactness

categories, while scale conversion, area computation, and quantity estimation use deterministic geometry. Each panel quantity is traceable to an image region, a valid calibration input, a specified product size, and a rounding rule. This traceability supports inspection of the calculation; it does not itself demonstrate real-world accuracy.

## IV. IMPLEMENTATION

### *A. Algorithmic flow and Preprocessing*

Fig. 2 shows the algorithmic flow of the estimation pipeline, from the input plan image to the panel estimate. Directly stretching a non-square image changes its geometry. The system instead uses a letterbox transform that preserves aspect ratio by padding rather than stretching, computing a uniform scale factor s and centring offset (px, py) as shown in Equation 1, where w, h are the source dimensions and Tw, Th are the target dimensions. Padding is excluded from measurement, and scale-reference coordinates must be transformed consistently with the mask.

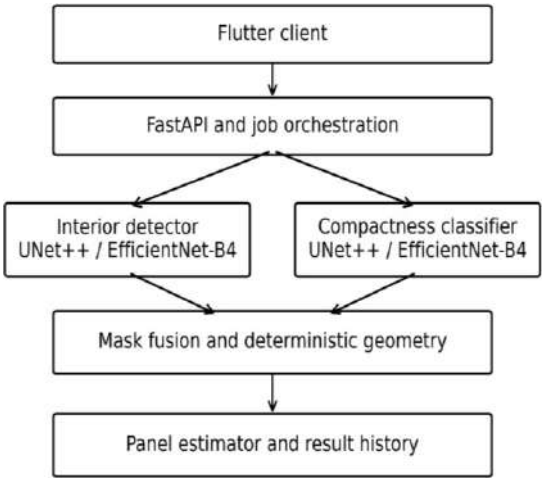



Fig. 1. Proposed layered system architecture.

### *B. Segmentation Model*

An initial U-Net with a ResNet-50 encoder was replaced with UNet++ using an EfficientNet-B4 encoder. UNet++ uses nested dense skip pathways [1], and EfficientNet provides the encoder configuration [2]. The final documented design contains a binary interior detector M1 and a compactness classifier C2. Their outputs are combined as Lfinal(x,y) = M1(x,y)·C2(x,y), with background index 0 and compactness indices 1–3. A corrected-label architectural baseline was not independently established. The False Ceiling Panel Detection (FPD) masks are RGB room-instance renders: white exterior and black wall pixels are background, while colored room fills receive the compactness category parsed from the filename. The labels are background, NO_COMPACTNESS (no clustering), SIMPLE_CLUSTERS (simple clusters), and FULL_COMPACTNESS (fully compact arrangement). A 200-mask audit reported pixel shares of 65.9%, 11.7%, 6.9%, and 15.5%, respectively; 200 is the audit sample, not the total dataset size.

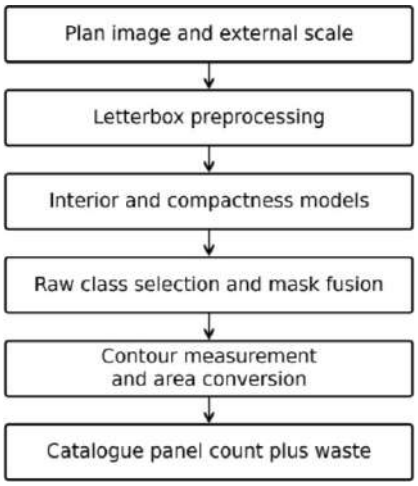



Fig. 2. Algorithmic flow of the estimation pipeline.

The original grayscale decoder linearly rescaled intensities into class indices, assigning white background to FULL_COMPACTNESS in training, validation, and test masks. The corrected color decoder assigns pure white and pure black to background and remaining room-fill pixels to the filename category. Retraining used corrected targets and a fresh checkpoint. Pre-correction scores cannot establish a valid corrected-label baseline.

The documented configuration uses a fixed stratified 70/15/15 training/validation/test split, 512×512 input, batch size 2, eight-step gradient accumulation (effective batch 16), mixed precision, and an NVIDIA T4 GPU on Kaggle. The exact split counts and manifest, related-plan leakage controls, augmentations, normalization, optimizer, schedule, and final loss coefficients are not documented sufficiently to reproduce the reported checkpoint; the 70/15/15 ratio alone does not establish leakage-free evaluation.

The final project configuration summary describes scSE attention [6] for interior detection and CBAM [5] for compactness classification; this describes the intended design, while the evaluated checkpoint assignment is unverified because the general architecture description also specifies scSE. CBAM applies channel attention Mc followed by spatial attention Ms [5]. Channel attention pools across channels; the semicolon in Equation 2b denotes channel-wise concatenation. With $F' = Mc(F) \odot F$, the spatial map in Equation 2b acts on $F'$, and the final refined features are $F'' = Ms(F') \odot F'$.

Corrected-label retraining combines Tversky, Lovász-Softmax, Dice, and an added class-weighted cross-entropy term. Under Equation 3, $\alpha$ multiplies false positives and $\beta$ multiplies false negatives [7]. The final retraining description uses $\alpha = 0.3$ and $\beta = 0.7$, giving a stronger false-negative weight. This supersedes the earlier 0.15/0.85 run and the incorrect explanation that $\alpha < \beta$ imposes a stronger false-positive penalty.

The loss terms are combined in Equation 4, where LTversky = 1 − Tversky, LLovász is the Lovász-Softmax loss, LDice = 1 − Dice, and LCE,w is class-weighted cross-entropy. Inverse-square-root class-frequency weighting was added for the corrected-label run. Exact λ coefficients, numerical smoothing, background inclusion, and reduction conventions are not available in the reported training description; the original three-term 0.4/0.4/0.2 weights do not establish the four-term objective.

Raw argmax was selected as the final documented inference mode after label correction. A comparison pipeline applies morphological cleaning and Connected Component Analysis. The representative thesis implementation uses the 5% relative area rule in Equation 5, rather than the former 4% rule. It is not a mandatory cleaning step for raw inference. Kernel size, connectivity, class-wise operation, and TTA transforms are not fully documented, limiting reproduction of the two comparison variants.

Equation 6 records the earlier fixed FULL_COMPACTNESS logit penalty for historical context only. It is not presented as a validated part of the final raw-argmax mode. Subtracting from one class logit can cause another foreground class, as well as background, to win; background rejection requires an explicit rule. The old 0.75 confidence threshold and the 297-to-15 fragment result do not establish the corrected-label model's performance.

*C. Matrix Formulation of Model Training and Inference*

Beneath the architectural description in Section IV-B, every layer of the network is a matrix or tensor operation. Each convolutional layer computes its output feature map by rearranging the input patches into columns (im2col) and multiplying by the flattened kernel matrix, given in Equation 7, where X is the input feature map, $W_{mat}$ is the reshaped kernel matrix, and b is the bias vector.

A classifier produces a logit tensor $Z \in R^{HxWxc}$; softmax normalises Z into class probabilities P and argmax selects a class at each pixel, as in Equation 8. In the two-model design, the interior mask restricts the compactness output to the detected interior. The compactness classes describe spatial arrangements, not physical material products.

UNet++'s nested skip pathway builds each intermediate feature matrix by concatenating all preceding feature maps at the same depth with an up-sampled matrix from one level deeper, given in Equation 9, where H is a convolution block, U is up-sampling, and [·] denotes channel-wise concatenation.

Equation 10 gives the generic gradient-descent form for a weight update on the composite loss L (Equation 4). It is an explanatory formulation, not a specification of the actual optimizer. The final optimizer, learning-rate schedule, epoch count, initialization, seed, and checkpoint-selection rule are not specified in the available experiment record; Equation 10 does not establish the update rule used in training.

*D. Deterministic Geometry Engine*

Morphological opening removes small foreground specks and morphological closing fills small holes, as shown in Equation 11, where $\ominus$ and $\oplus$ are erosion and dilation with structuring element S. These operations belong to the post-processing comparison. In the corrected-label evaluation, cleaning also filled wall separators and lowered overlap accuracy; raw argmax was consequently retained as the documented inference choice. Contour measurement must preserve valid internal exclusions.

Contours are converted from pixel area to square metres using the letterbox scale s (Equation 1) and an externally supplied calibration ratio r = real_length_cm / pixel_length. In Equation 12, pixel_area is measured in letterboxed coordinates and pixel_length in original-image coordinates. If the mask has already been restored to original-image coordinates, the division by $s^2$ is omitted. Scale is an external input; automatic OCR and reference-object detection were proposed but not implemented. A scale input consists of a known length in centimetres and its corresponding endpoint distance in original-image pixels; both must describe the same measured plane or uniformly scaled plan, with padding excluded.

A valid uniform scale requires a plan or rectified planar view. Letterboxing does not remove perspective foreshortening. A reference must be geometrically appropriate to the measured plane; arbitrary oblique photographs require validated rectification or calibration [10]. Sloped roof area is not automatically ceiling installation area. Without a valid scale, only pixel area is justified. Because area is proportional to $r^2$, a small relative scale error contributes approximately twice that relative error to area before segmentation error is included. This sensitivity is analytical; reference-placement and scale uncertainty have not been measured in a field study.

*E. Material Quantity Estimation*

Each calibrated area is converted into a panel count using Equation 13, where panel area is the product width multiplied by its length in square metres, for a separately specified product and b is a configurable waste buffer (default 10%). Compactness labels do not determine whether a product is gypsum, PVC, timber, or another material. The representative thesis code first rounds the area ratio, then rounds again after adding waste. Counts cover panels only; framing, hangers, fixings, cutting-layout optimization, and validated cost savings are outside the demonstrated calculation.

$$n_0 = \left\lceil \frac{\text{Area}_{m^2}}{\text{panel_area}} \right\rceil, \quad n = \lceil n_0(1+b) \rceil$$

*Equation 13. Material panel quantity with waste buffer (single area and product)*

*F. Backend and Mobile Client*

The backend is implemented in FastAPI with a singleton ModelRegistry that downloads the model checkpoint once and caches it locally, avoiding the reload-per-request and accidental retrain-per-restart failure modes observed during development. The mobile client is implemented in Flutter with Riverpod, using a sealed SegmentationState (Idle, Uploading, Analyzing, Success, Failure) so the compiler enforces that every UI state is handled explicitly.

## V. RESULTS AND EVALUATION

Equation 14 defines a four-class confusion matrix C, where entry Cij counts the pixels with target class i and predicted class is j. For a specified evaluation cohort, class IoU is Ckk/(Σj Ckj + Σi Cik − Ckk), and macro mean IoU averages over a stated class set. Dataset-wide and image-wise averaging, absent-class handling, and background inclusion must be distinguished. Table I retains one internally consistent inference-mode record; its aggregation convention is unverified, and the separate class-wise summary is not used to validate these aggregates.

Table I compares the corrected-label model under three inference modes. Raw argmax is the inference reference; post-processing and TTA are tested inference variants, not separately trained architecture baselines. Earlier ResNet-50 and pre-correction UNet++ values are excluded from direct comparison because the label-decoding defect affected the scored targets. Fractions are used consistently for pixel accuracy and Dice.

TABLE I
RECORDED SEGMENTATION PERFORMANCE BY INFERENCE MODE

| Metric | Raw argmax | +Post-process | +TTA +post |
|---|---|---|---|
| Mean IoU | 0.9440 | 0.9211 | 0.9210 |
| Pixel accuracy | 0.9853 | 0.9762 | 0.9762 |
| Dice | 0.9710 | 0.9587 | 0.9587 |
| Noise Fragments | 2 | 0 | 0 |
| Time (mis/image) | 53.2 | 198.1 | 393.2 |

Within the recorded corrected-label comparison, raw argmax has the highest reported overlap: mean IoU 0.9440, pixel accuracy 0.9853, and Dice 0.9710. Post-processing reduces reported mean IoU to 0.9211, a decrease of 2.29 percentage points, while TTA with post-processing gives 0.9210. The corresponding fragment counts are 2, 0, and 0. Fewer fragments do not necessarily mean better masks: cleaning can fill wall lines and increase the measured interior area. The timings in Table I are recorded per-image comparison timings, not end-to-end mobile latency; the complete upload-to-result benchmark remains pending

A separate, unverified class-wise values reported elsewhere in the final thesis are 0.968, 0.941, 0.884, and 0.939; their arithmetic mean is 0. 9330. That summary was labelled TTA + post-process, whereas Table I transcribes 0.9210 for that mode from the thesis inference table. These records must be matched by checkpoint, cohort, inference mode, and aggregation before class-specific claims are made. The reported 0.9440 raw mean IoU has not been independently recomputed for this manuscript.

Expert validation was not completed within the thesis timeframe. The proposed comparison uses 50 images rather than 100 completed expert cases. MAPE below 10%, paired-test significance, expert agreement, and procurement-time reduction are therefore not reported as achieved results. A future study must compare areas against measured reference dimensions and panel counts against documented material specifications, while distinguishing physical accuracy from agreement with expert estimates.

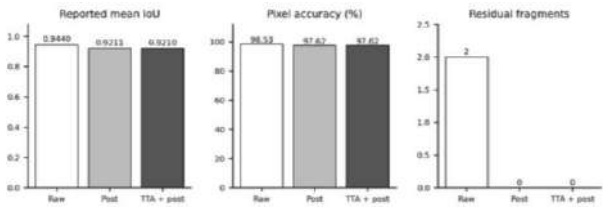



Fig. 3. Recorded segmentation metrics by inference mode

## VI. STRENGTH & LIMITATIONS

The system's practical contribution is the integration of learned segmentation with auditable scale conversion and panel arithmetic. The principal experimental lesson is that label integrity must be checked before visual errors are attributed to the model or its attention mechanism. Repairing the RGB-mask decoder and retraining changed the preferred inference mode: raw argmax preserved more overlap than the tested cleaning pipeline. No independent gain from attention or loss changes is established by this comparison.

A limitation of the present study is reliance on the existing FPD plan-image collection instead of the independently collected dataset of 2,000+ images originally proposed. Dataset provenance, license, full image counts, split counts, and complete training configuration remain insufficiently documented for independent reproduction. Generalization to smartphone photographs, varied lighting, occlusion, non-planar ceilings, and geographic settings outside the evaluated collection is unknown. The compactness classes are not material-type labels, and a scalar calibration does not establish perspective-robust measurement.

A controlled ablation of architecture, attention, loss, confidence rules, morphology, and TTA on the same corrected labels remains necessary. The present evidence supports an integrated prototype and an inference-mode comparison, not real-world area accuracy, a validated bill of materials, or state-of-the-art segmentation performance. In conclusion, label correction and raw-inference selection supersede the earlier noise-suppression interpretation; physical calibration tests, complete experiment metadata, and expert quantity validation are required to assess procurement reliability.

## REFERENCES

[1] X. e. a. Xiao, "A novel data-driven based high-precision building roof contour full-automatic extraction and structured 3D reconstruction method combining stereo images and LiDAR points," *International Journal of Digital Earth,* vol. 18, no. 1, 2025.

[2] Q. Mei, D. Bulatov and D. Iwaszczuk, "Polygonizing Roof Segments from High-Resolution Aerial Images Using Yolov8-Based Edge Detection," 2025.

[3] B. Aslam, A. Maqsoom, H. Inam, M. ul Basharat and F. Ullah, "Forecasting Construction Cost Index through Artificial Intelligence," *Societies,* vol. 13, no. 10, 2023.

[4] L. Deng, "Real estate valuation with multi-source image fusion and enhanced machine learning pipeline," *PLoS One,* vol. 20, no. 5, p. e0321951, 2025.

[5] e. a. Semwal, "False ceiling deterioration detection and mapping using a deep learning framework and the teleoperated reconfigurable "falcon" robot," *Sensors,* vol. 22, no. 1, 2022.

[6] Zhao and H. Jebelli, "A computational method for real-time roof defect segmentation in robotic inspection," *Computer-Aided Civil and Infrastructure Engineering,* vol. 40, no. 23, p. 3596–3623, 2025.

[7] P. A. e. a. Shiroya, "Machine Learning for LiDAR-Based Indoor Surface Classification in Intelligent Wireless Environments," arXiv, 2025.

[8] T. Schack, M. Coenen and M. Haist, "Image-based quality control of fresh concrete based on semantic segmentation algorithms," *Civil Engineering Design,* vol. 6, no. 3, p. 96–105, 2024.

[9] B. e. a. Arıkan, "Machine Learning-Based Manufacturing Cost Prediction from 2D Engineering Drawings via Geometric Features," arXiv, 2025.

[10] X. X. Zhu, S. Chen, F. Zhang, Y. Shi and Y. Wang, "GlobalBuildingAtlas: An Open Global and Complete Dataset of Building Polygons, Heights and LoD1 3D Models," arXiv, 2025.

# Security-by-Design for AI-Enabled Digital Twins in Critical Infrastructure: An Integrated Framework for Cyber, AI-Model and Twin-Synchronization Risks

Senesh Fitzroy
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
dsfmendis@students.nsbm.ac.lk

Pawani Bandara
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
pawanimaheshika2003@gmail.com

***Abstract*—AI-enabled digital twins in critical infrastructure couple a physical plant, learned models and a bidirectional synchronization channel that current practice defends separately, leaving seams an adversary can compose across while every individual control reports nominal operation. This paper presents SecTwin-CI, a lifecycle-oriented security-by-design framework that treats cyber-physical security, AI-model security and twin-synchronization integrity as one design problem. It contributes (i) a five-layer reference architecture with an explicitly defended synchronization plane; (ii) a cross-layer taxonomy of ten threat classes anchored to MITRE ATT&CK for ICS and MITRE ATLAS; (iii) a traceable mapping from threats to twelve security requirements and layered controls; (iv) a dual-residual synchronization integrity monitor that cross-checks the learned twin against an independent reduced-order physical model; and (v) five lifecycle security gates. An analytical assessment over a complete threat-by-approach scoring matrix gives a cross-layer coverage index of 0.85 against at most 0.50 for five baseline families, a design-coverage score rather than a measured effect. A closed-loop water-tank simulation (200 runs per scenario) then shows the monitor detecting 80% of slow sensor drifts and 100% of residual-aware stealthy attacks, against at most 4% for single-residual monitoring, at a 1% false-alarm rate; constant offsets evade both tests, and plant ageing triggers the same alarm as an attack.**

***Keywords—digital twin security, security-by-design, critical infrastructure protection, adversarial machine learning, cyber-physical systems***

## I. Introduction

A digital twin (DT) is a virtual representation of a physical asset held in continuous correspondence with that asset through sensed data [1]-[3]. In critical infrastructure (CI) electricity networks, water treatment and distribution, rail signalling, oil and gas processing DTs have moved quickly from visualisation aids to operational systems that forecast failure, optimise setpoints and, increasingly, write control decisions back to the plant [4]. Embedding machine learning accelerates this shift: learned models supply state estimation where instrumentation is sparse and detect anomalies at a sensitivity hand-built rules cannot reach.

The security consequence is under-appreciated. An AI-enabled DT with closed-loop authority is a participant in the control path, so it inherits the consequence model of the physical process—contaminated water, tripped feeders, damaged equipment [5]-[7]. Stuxnet [5] manipulated the operator's view of process state, and TRITON [7] targeted the safety instrumented system. An AI-enabled DT also acquires two attack surfaces a conventional industrial control system (ICS) does not have [8], [9]: the learned models, and the synchronization channel keeping the twin faithful to the plant.

These surfaces have well-developed but largely disjoint literatures. Work on securing DT infrastructure treats the twin as an IT/OT asset to be hardened [10], [11]. Work on the DT as a security mechanism uses the twin to genserate ground truth for intrusion detection [12]-[14]. Work on adversarial machine learning characterises poisoning, evasion, extraction and inversion, largely in domains without physical consequence [15]. Each is strong within its boundary. The seam between them is where AI-enabled DTs in CI actually fail.

Consider a water-treatment twin whose chemical dosing model is periodically retrained on operational history. An adversary with write access to the historian introduces a slow, small bias into a subset of conductivity readings. Retraining absorbs the bias and the dosing recommendation shifts. Network controls report no intrusion, because the adversary uses authorised data paths. Model-integrity checks pass, because the deployed model is exactly the artefact the pipeline produced. Fidelity monitoring reports a healthy twin, because the twin now agrees with the manipulated sensor stream better than before. Every control is green and the plant is drifting out of specification. The failure is not located in any layer; it is the absence of any mechanism reasoning across them.

A second problem is temporal: a twin is recalibrated as the plant ages, retrained as data accumulates and re-scoped as the process changes [16], [17], and each event can reintroduce risk that a one-time assessment certified away. Security-by-design must therefore be lifecycle-resident, not a pre-deployment gate.

This paper makes five contributions.

A cross-layer taxonomy of ten threat classes for AI-enabled DTs in CI, anchored to MITRE ATT&CK for ICS and MITRE ATLAS [18], with an explicit notion of compound attack paths traversing layers.

SecTwin-CI, a five-layer reference architecture elevating the synchronization plane to a first-class defended element.

A traceable mapping from threat classes to twelve security requirements and to preventive, detective and responsive controls at each layer.

A dual-residual synchronization integrity monitor cross-checking the learned twin against an independent reduced-order physical model, detecting stealthy consistency attacks that single-residual monitoring misses, which we quantify in closed-loop simulation.

Five lifecycle security gates with defined exit evidence, an analytical coverage evaluation over three compound scenarios and five baseline families, and a simulation study of the monitor.

## II. Related Work and Research Gap

### A. Securing the Digital Twin

Alcaraz and Lopez [10] give the most comprehensive survey of DT security threats, covering the twin's data, communication and platform stack, and identify the bidirectional link as a distinctive risk without specifying a defence for it. ISO 23247 [19] defines a DT framework for manufacturing but treats security as an external concern. IEC 62443 [20] supplies the dominant OT security architecture—zones, conduits and security levels and NIST SP 800-82 Rev. 3 [9] the dominant control catalogue; both predate learned components and have no notion of model integrity or twin fidelity. Suhail et al. [21] anchor twin state and provenance in a distributed ledger, strengthening artefact integrity but leaving the AI layer untreated.

### B. The Digital Twin as a Security Mechanism

Eckhart and Ekelhart [11], [12] host specification-based intrusion detection in virtual replicas of a CPS, and Dietz and Pernul [13] and Dietz et al. [14] bring DT-based security simulation into security operations. This asks what the twin can do for security, not what secures a twin whose own models are attackable.

### C. Adversarial Machine Learning

Poisoning and backdooring [22], [23], evasion [24] [26], model extraction [27] and membership inference [28] are surveyed by Biggio and Roli [15] and codified in NIST AI 100-2 [29]. Defences centre on robust training, certified bounds [30] and data provenance, but are rarely evaluated where a misclassification has physical consequence; in a plant the attacker seeks sustained, undetected process deviation.

### D. Attack Detection in Industrial Control Systems

Cárdenas et al. [31] formalised detection and response for process control attacks, and Giraldo et al. [32] survey physics-based detection, the family our construction extends. Mo and Sinopoli [33] show replay attacks defeat residual tests unless the controller injects an authenticating signal, and Ahmed et al. [34] authenticate sensors by noise fingerprint. On the learned side, Kravchik and Shabtai [35] demonstrate convolutional detection on SWaT [36], while Erba et al. [37] evade such detectors by constrained concealment. Most relevant, Urbina et al. [38] show a resourceful adversary can keep an attack beneath a residual-based detector's threshold indefinitely motivating Section IV-D: a single residual computed against a reference the adversary can influence is not a sufficient integrity signal. Governance instruments the NIST AI RMF [39], zero trust architecture [40], the NIS2 Directive [41] and the NCSC/CISA secure-AI guidelines [42] set outcomes but give no engineering mapping to the layers of an AI-enabled DT.

TABLE I
Threat-by-Approach Scoring Matrix and Analytical Cross-Layer Coverage Index

| Approach | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | CLCI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B1 IEC 62443 + OT IDS [9], [20] | ½ | 1 | 1 | ½ | ½ | 0 | 0 | 0 | 1 | ½ | 0.50 |
| B2 DT as security mech. [11]–[14] | 1 | 0 | 1 | ½ | 0 | 0 | 0 | 0 | ½ | 0 | 0.30 |
| B3 Adversarial-ML hardening [15], [29] | 0 | 0 | 0 | 0 | ½ | 1 | 1 | 1 | 0 | ½ | 0.40 |
| B4 ML anomaly detection [31], [35] | 1 | ½ | ½ | 0 | ½ | 0 | 0 | 0 | ½ | 0 | 0.30 |
| B5 Provenance-anchored DT [21] | ½ | 1 | ½ | ½ | 1 | 0 | 0 | 0 | ½ | 1 | 0.50 |
| **SecTwin-CI (this work)** | **1** | **1** | **1** | **1** | **1** | **½** | **½** | **½** | **1** | **1** | **0.85** |

Columns 1–10 are threat classes TC1–TC10 (Table II). 1 = covered by an explicit control with defined prevention or detection semantics; ½ = addressed with a documented residual gap; 0 = not addressed. Scores are analytical design judgements against Table III and each baseline's published scope (Section V-B), not measurements; CLCI is the row mean, eq. (4).

No existing approach addresses the cyber-physical, AI-model and synchronization surfaces simultaneously with controls persisting across the lifecycle. Prior work is not wrong within its scope; the gap is that the scopes do not compose, and adversaries operate across them.

## III. System and Threat Model

### A. Reference System

The plant has state $x_t$ in $\mathbb{R}^n$, observed through sensors as $y_t = h(x_t) + v_t$. The twin maintains an estimate $\hat{x}_t$ produced by a learned model $f_\theta$ over a window of observations, and in closed-loop configurations proposes a control action $u_t = \pi(\hat{x}_t)$. The synchronization plane carries observations upward and control actions downward with bounded lag $\delta$, and retraining occurs periodically on historian data. This configuration is the system we secure.

### B. Adversary Model

We consider five capability classes, not necessarily held together: A1, a network position inside the OT network or DT demilitarised zone; A2, write access to the historian or retraining pipeline; A3, query access to the deployed model; A4, supply-chain access to model artefacts or components; and A5, an authorised insider with retraining rights. Goals are process deviation, stealthy persistence, denial of view or control, and exfiltration of process IP. We assume sound cryptography and a patient adversary who stays inside nominal telemetry envelopes [38], i.e. one optimised against the detectors deployed.

### C. Cross-Layer Threat Taxonomy

Table II enumerates ten threat classes spanning the five architectural layers and the synchronization plane, each anchored to a published adversary technique [18] so the taxonomy is auditable rather than declarative.

### D. Compound Attack Paths

A compound attack path is an ordered sequence of threat classes in which the effect of an earlier class alters the detection surface presented to a later one. The historian example of Section I is TC6 → TC5 → TC9: poisoning shifts the model, the shifted model becomes the twin's accepted state, and the resulting setpoint is issued as a legitimate command. Layer-local controls fail against such paths for a specific reason. Each decides by comparing observed behaviour against a nominal baseline, and an earlier step has already moved that baseline. The controls are not bypassed; they are re-calibrated by the attack. Defending against this requires a detector whose reference cannot be moved by any single step in the path the design obligation Section IV-D discharges.

TABLE II
CROSS-LAYER THREAT TAXONOMY FOR AI-ENABLED DIGITAL TWINS IN CRITICAL INFRASTRUCTURE

| ID | Threat class and effect | Layer | Anchor technique |
|---|---|---|---|
| TC 1 | Sensor/actuator manipulation: spoofed or slowly biased measurements; actuator override | L1 | A: Spoof Reporting Message |
| TC 2 | Unauthorised access and lateral movement into twin data and model stores | L2 | A: Valid Accounts |
| TC 3 | OT protocol abuse: injection and command replay on unauthenticated traffic | L2 | A: Adversary-in-the-Middle |
| TC 4 | Synchronization attacks: state-stream replay, latency injection, induced desynchronization | Sync | A: Denial of View |
| TC 5 | Twin tampering: alteration of state model, geometry, calibration parameters | L3 | A: Modify Parameter |
| TC 6 | Training-data poisoning: slow bias or backdoor via the retraining pipeline | L4 | T: Poison Training Data |
| TC 7 | Adversarial evasion: crafted inputs suppressing detection or biasing prediction | L4 | T: Evade ML Model |
| TC 8 | Model extraction and inversion: theft of process IP or operating data | L4 | T: Exfiltration via ML Inference API |
| TC 9 | Malicious or unsafe control action issued through the twin's closed loop | L5 | A: Unauthorized Command Message |
| TC 10 | Supply-chain compromise of models, libraries or third-party twin components | Cross | T: ML Supply Chain Compromise |

A = MITRE ATT&CK for ICS technique; T = MITRE ATLAS technique [18].

## IV. THE SECTWIN-CI FRAMEWORK

### A. Design Principles

SecTwin-CI rests on six principles. *P1 Explicit synchronization plane*: the twin–plant channel is a defended element with its own requirements and controls. *P2 No sole AI authority*: every safety-relevant action is arbitrated by a mechanism with no learned parameters. *P3 Assume compromise between layers*: inter-layer trust is authenticated and least-privilege [40]. *P4 Provenance for everything that moves twin state*: data, models, parameters and geometry carry verifiable origin [21]. *P5 At least one cross-layer detector*, whose inputs span two layers so that single-layer manipulation cannot re-calibrate it. *P6 Evidence-carrying lifecycle*: each phase emits artefacts the next phase checks [43].

### B. Reference Architecture

Fig. 1 shows the architecture: five layers from physical sensing to control and actuation, a synchronization plane binding L1 to L3 in both directions, and two cross-cutting planes, identity and trust, and governance and assurance (hosting the gates of Section IV-E). Every threat class has an owning layer and every layer named controls.

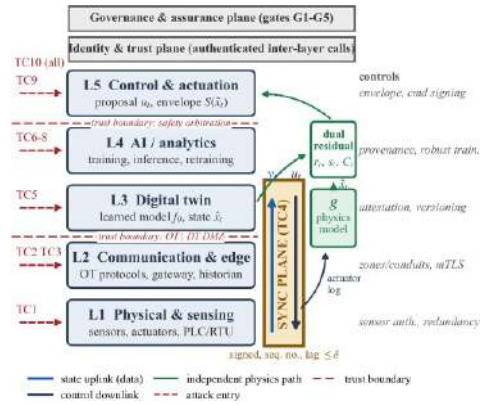



Fig. 1. SecTwin-CI reference architecture. The synchronization plane (TC4) carries the state uplink (data) and control downlink between L1 and the twin; the physics model g and the dual-residual monitor sit on a separate trust path fed by the authenticated actuator log; dashed lines mark trust boundaries and attack entry points.

TABLE III. SECURITY REQUIREMENTS AND LAYERED CONTROL MAPPING

| Layer | Requirements | Controls: preventive / detective / responsive |
|---|---|---|
| L1 Physical & sensing | SR1 sensor authenticity<br>SR2 sensing redundancy | P: signed sensor gateways; noise-fingerprint authentication.<br>D: cross-sensor consistency; redundant-channel voting.<br>R: sensor quarantine; channel failover. |
| L2 Comm. | SR3 authenticated conduits | P: mTLS; 62443 zones and conduits; zero-trust segmentation. |

| | | |
|---|---|---|
| & edge | SR4 least-privilege access | D: protocol-aware inspection; access analytics.<br>R: conduit isolation; credential revocation. |
| Sync plane | SR5 message freshness<br>SR6 verifiable fidelity | P: signed state vectors; sequence numbers; lag bound δ.<br>D: dual-residual test (1)–(3); replay detection.<br>R: fail-to-advisory; attested resynchronization. |
| L3 Digital twin | SR7 model integrity<br>SR8 state provenance | P: signed artefacts; immutable versioning; change control.<br>D: attestation at load; configuration drift detection.<br>R: rollback to last attested version. |
| L4 AI / analytics | SR9 data provenance<br>SR10 inference robustness<br>SR11 model confidentiality | P: dataset signing; robust training; query rate limits.<br>D: poisoning and drift detectors; output-bound violations.<br>R: model rollback; retraining freeze; degrade to g. |
| L5 Control & actuation | SR12 no unarbitrated AI authority | P: envelope on the physics estimate; command signing; dual approval.<br>D: envelope violation logging; command audit.<br>R: envelope clamp; local control; escalation. |

### C. Security Requirements and Controls

Twelve requirements (SR1–SR12) are derived by asking, for each threat class, what property must hold for it to be prevented, detected or contained (Table III). The mapping is traceable both ways: TC1→SR1–SR2, TC2–TC3→SR3–SR4, TC4→SR5–SR6, TC5→SR7–SR8, TC6–TC8→SR9–SR11, TC9→SR12 and TC10→SR7, SR9. Coverage in Section V is computed over exactly this mapping.

### D. Dual-Residual Synchronization Integrity

This is the framework's principal technical mechanism. Let $\hat{x}_t$ be the twin state produced by $f_\theta$, and $\tilde{y}_t = h(\hat{x}_t)$ the observation the twin predicts. Conventional fidelity monitoring tracks the single residual

$$\rho\tau = \|\psi\tau \quad \rceil\tau\|\Omega \quad (1)$$

and alarms when $r_t$ exceeds a threshold τ calibrated on nominal operation [32]. Equation (1) has a known weakness. An adversary who manipulates the sensor stream and the twin's inputs consistently or who poisons training so that $f_\theta$ learns the manipulated relationship drives $r_t$ down rather than up [33], [38]. Agreement with a corrupted reference is indistinguishable from health.

Here $\|e\|_W = (e^T W e)^{1/2}$, with $W$ the inverse covariance of the nominal residual estimated at commissioning (gate G1), so $r_t$ is in nominal standard deviations; $s_t$ is weighted likewise by the nominal divergence covariance.

SecTwin-CI adds a structurally independent second reference. Let $g$ be a reduced-order physical model of the process, derived at design time from conservation laws, plant geometry and commissioning data [44], containing no learned parameters and never retrained on operational data. It yields an independent estimate $\tilde{x}_t$. We define the model-divergence residual

$$\sigma_\tau = \|\xi\square_\tau \quad \xi\sim_\tau\|\Omega \quad (2)$$

and monitor the pair ($r_t$, $s_t$) jointly, with the cumulative statistic

$$X_\tau = \mu\alpha\xi(0, X_{\tau\ 1} + \sigma_\tau \quad |) \quad (3)$$

accumulating small, persistent divergence that a per-sample threshold on $s_t$ would not raise.

The joint test separates three cases a single residual conflates. Nominal operation gives low $r_t$ and low $s_t$. Degradation of the learned twin gives rising $r_t$ and rising $s_t$, and is handled as a maintenance event. Plant ageing [16] is harder: the one-step twin follows the plant but the fixed g does not, so ageing presents like an attack and is triaged against maintenance records (Section V-E). A consistency attack coordinated sensor spoofing [34], slow drift, or residual-aware evasion [38] is characterised by low $r_t$ with rising $s_t$ or rising $C_t$: the twin agrees with what it is told and disagrees with physics. That asymmetry is the detection signal, and exists only because $g$ sits on a different trust path from $f_\theta$. Critically, $g$ need not be accurate enough to control the plant, only to bound plausible state a weaker requirement a commissioning-phase model usually already meets.

Two further properties are required. First, freshness: every uplink message carries a monotonic sequence number and an authenticated timestamp, and the twin rejects messages outside the lag window δ, so replay cannot supply an internally consistent but stale stream [33]; a replay spliced at a matching state need not disturb either residual. Second, arbitration: a proposed action $u_t$ is admitted only if it lies in $S(\tilde{x}_t)$, where $S$ is a safety envelope derived from process constraints and evaluated against the physics estimate rather than the learned one [45], [46]. When the envelope rejects a proposal the twin degrades to advisory mode rather than failing open. Together, (1)–(3), freshness and arbitration discharge P1, P2 and P5.

*Relation to established methods.* The construction is analytical redundancy from model-based fault diagnosis [47] and physics-based attack detection [32], paired with a learned model as in theory-guided machine learning [48]. What differs is trust placement: $g$ sits on a separate path (authenticated actuator log, separate versioning, never retrained), so an adversary controlling the sensor stream or the training pipeline cannot move both references, whereas observer-based secure estimation [45], [46] assumes a trusted model. Equation (3) is Page's one-sided CUSUM [49] on $s_t$, the standard sequential test for a persistent mean shift: $\kappa$ is the divergence tolerated as model error and an alarm is raised when $C_t > h$. Section V-E reports calibrated $\tau$, $\kappa$, $h$ and $W$.

### E. Lifecycle Security Gates

Fig. 2 binds the controls to five gates, in the spirit of ML production-readiness rubrics [17], [43] but with security exit criteria. G1 (design): a threat model over Table II and a sync specification fixing $\delta$, $\tau$, $\kappa$, $h$ and $W$. G2 (development): signed training data with provenance, robustness testing against evasion and poisoning, and a versioned $g$. G3 (deployment): attested onboarding, independent verification of the envelope, adversarial testing [42]. G4 (operation): a live dual-residual monitor with escalation playbooks. G5 (evolution): dual

authorisation for retraining, review of the divergence record, re-validation of $g$, and a tested rollback. Any change to model, process or threat landscape returns the twin to G1.

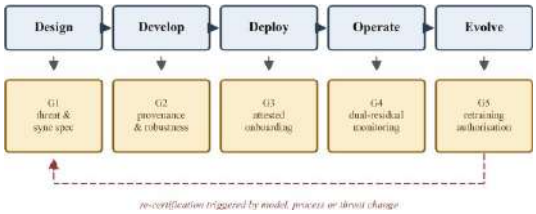



Fig. 2. Lifecycle security gates. Evolution events return the twin to G1, making assurance re-established rather than assumed.

## V. Evaluation

### A. Method and Scope

We evaluate SecTwin-CI in two complementary ways and keep them separate. Sections V-B to V-D give an analytical threat-coverage assessment: three compound attack scenarios on a representative water-treatment twin, reasoned control by control, and a coverage index computed from the complete scoring matrix in Table I. These are design-level coverage scores, not measured detection or false-positive rates. Section V-E then quantifies the framework's principal detective mechanism, the dual-residual monitor of Section IV-D, in closed-loop simulation against single-residual monitoring. The reference plant is modelled on the six-stage secure water treatment testbed [36], [50]: roughly fifty sensors and actuators, and a twin performing state estimation, dosing recommendation and anomaly detection, retrained monthly.

### B. Coverage Measure

For a defence $D$ and threat set $T = \{TC1, \ldots, TC10\}$, let $c(D, TC)$ take the value 0, 0.5 or 1 according to whether a class is uncovered, partially covered (addressed with a documented residual gap) or covered by an explicit control with defined prevention or detection semantics. The cross-layer coverage index is reported per class in Table I. A class scores 1 when the approach's published scope contains an explicit control that prevents or detects it with defined semantics, ½ when it is addressed with a documented residual gap, and 0 otherwise; each score was assigned against Table III and the cited description of the baseline, and the full matrix is given so that any score can be contested individually. Row sums are 5.0, 3.0, 4.0, 3.0, 5.0 and 8.5. CLCI therefore measures analytical threat coverage of a design, not security effectiveness, and B1–B5 are families with different scopes rather than directly competing systems. We report path coverage separately: the fraction of compound paths in Table IV interrupted before physical effect. Scoring is expert judgement by a single rater group and inherits the subjectivity noted in Section VI.

### C. Compound Scenario Analysis

TABLE IV
Compound attack scenarios on a representative water-treatment twin

| Scenario and path | Baseline outcome | SecTwin-CI outcome |
|---|---|---|
| S1 Slow dosing poisoning via the historian. TC6 → TC5 → TC9 | No alarm: authorised data paths, a genuine model artefact, and a twin that fits the manipulated stream. | G2 dataset signing flags the unsigned segment. In simulation the poisoned twin also trips $r$ (V-E); $C$ against $g$ is the backstop for poisoning crafted to keep $r$ low. G5 blocks promotion; the envelope clamps the setpoint. |
| S2 Replayed state stream masking valve manipulation. TC2 → TC4 → TC1 | Operator view frozen on a plausible history; a consistent replay keeps $r$ low [33]. | Sequence numbers and the lag bound $\delta$ reject replayed frames on arrival; the twin degrades to advisory and raises loss-of-view. |
| S3 Adversarial input suppressing the twin's anomaly detector. TC7 → TC9 | Detector silenced by constrained concealment [37]; the unsafe setpoint reaches the PLC. | The envelope, evaluated on the physics estimate, rejects the action at L5 and logs it. Contained by P2. |

### D. Results and Residual Risk

The baselines score CLCI between 0.30 and 0.50: each is strong on its own surface and near-blind on the others (Table I). SecTwin-CI reaches 0.85 and, analytically, interrupts all three compound paths, against zero of three for B1 and at most one for any other baseline family.

The residual 0.15 sits in three classes. TC8 is partial: rate limiting raises the cost of extraction without preventing it [27]. TC7 is partial: the envelope contains the consequence of evasion without detecting suppression of the detector. The insider variant of TC6 is partial: an engineer with legitimate retraining authority operates inside the framework's trust assumptions; the monitor surfaces the divergence, turning silent compromise into a G5 review obligation, but authorised-insider poisoning remains open.

### E. Simulation of the Dual-Residual Monitor

*Setup.* We simulate a gravity-drained tank with pumped inflow, the process of the SWaT stage-1 raw-water tank [36], under closed-loop level control through the twin: $h_{t+1} = h_t + \Delta(k_p u_t - c\sqrt{h_t}) + w_t$, $y_t = h_t + v_t$, with moving setpoints. $f_\theta$ is a ridge-regression one-step predictor on $[1, y, u, \sqrt{y}]$ trained on 4000 historian samples; $g$ is the same mass balance with 5% error in $k_p$ and $c$, never retrained, driven by the authenticated actuator log and weakly anchored to the sensor stream (gain 0.05); $W = \sigma^{-2}$. Thresholds come from nominal runs only, at a 5% per-run false-alarm target split between DR's two components: $\tau = 7.2$, $\kappa = 2.54$, $h = 7.18$. We compare a threshold on $r$ (SR-th), Page's CUSUM on $r$ (SR-CUSUM) and the dual-residual test (DR), over 200 runs of 3000 steps per scenario.

TABLE V
Detection rate per run (median latency of DR, steps), 200 runs per scenario

| Scenario | SR-th | SR-CUSUM | DR | DR lat. |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Nominal (false alarm) | 0.05 | 0.07 | 0.01 | — |
| Sensor step offset (TC1) | 0.05 | 0.04 | 0.07 | 458 |
| Slow sensor drift (TC1) | 0.04 | 0.02 | **0.80** | 880 |
| Evasion under τ (TC7) | 0.00 | 1.00 | 1.00 | 1 |
| Evasion under τ and κ (TC7) [38] | 0.00 | 0.00 | **1.00** | 17 |
| Poisoned retraining (TC6→TC5) | 0.98 | 0.99 | 0.96 | 238 |
| Latency injection (TC4) | 1.00 | 1.00 | 1.00 | 42 |
| Replay (TC4) | 1.00 | 1.00 | 1.00 | 0 |
| Plant ageing (not an attack) | 0.09 | 0.04 | 1.00 | 540 |

*Results.* Against slow sensor drift the one-step twin tracks the manipulated stream, so the single-residual tests stay at 4% and 2% while DR detects 80%; against an attacker who keeps $r$ under both $\tau$ and the CUSUM reference value [38], they detect 0% and DR 100% (median latency 17 steps). A naive evader respecting only $\tau$ is caught by CUSUM on $r$. DR's false-alarm rate is 1% (SR-CUSUM 8%); over the seven attacks precision/recall/F1 are 0.98/0.44/0.61 (SR-th), 0.98/0.58/0.73 (SR-CUSUM) and 1.00/0.83/0.91 (DR). DR adds nothing in three cases: a constant step offset is absorbed within one sample and missed by every test (≤7%); naive replay and latency injection break one-step consistency and are caught by every test (100%), though a replay spliced at a matching state would not be [33], so freshness remains the TC4 control; and the simulated poisoning changed closed-loop behaviour enough for $r$ alone to alarm (99%).

*Model mismatch and cost.* Under 15% outflow degradation, a maintenance event, DR alarms in 100% of runs, all through $C$ ($r$: 0%): the twin follows the ageing plant and the fixed $g$ does not, so ageing has the low-$r$, high-$C$ signature of an attack. Alarms must be triaged against maintenance records and $g$ re-validated under G5. The second residual costs 2.5 µs per sample in unoptimised Python. These are results on a synthetic single-loop plant, not SWaT or WADI measurements.

## VI. Discussion and Limitations

### A. Cost, Deployability and Alignment

The obvious objection is that SecTwin-CI needs a second model. In practice $g$ is rarely new work: CI assets are commissioned against physical design models [44], which the framework asks to be retained, versioned and kept on a separate trust path, at a runtime cost of microseconds per sample (Section V-E). The framework also maps onto current practice: L1–L2 onto IEC 62443 zones and conduits [20] and SP 800-82 [9], the gates onto the NIST AI RMF functions [39], and G3/G5 evidence onto NIS2 obligations [41].

### B. Limitations

Five limitations bound the claims. First, coverage is analytical and the monitor's quantitative evidence comes from a simulated single-loop plant, not SWaT or WADI data or a field deployment. Second, coverage scoring is expert judgement by a single rater group on our own framework, a known source of favourable bias that the published matrix makes checkable but does not remove. Third, the scenario set is small and chosen to exercise cross-layer paths, which favours a cross-layer framework. Fourth, $g$ drifts with the plant, so the monitor raises maintenance alarms that need triage. Fifth, arbitration assumes a specifiable safe region; a loose envelope weakens P2's containment.

### C. Future Work

The priority is running the Section V-E scenario set on the SWaT and WADI testbeds [36], [50], [51], including poisoning crafted to keep $r$ low and multi-variable plants with a full $W$, followed by multi-rater scoring and formal verification of the envelope $S$.

## VII. Conclusion

AI-enabled digital twins in critical infrastructure are defended as three separate problems, and adversaries do not respect that separation. SecTwin-CI treats cyber-physical, AI-model and synchronization risk as one design problem across the lifecycle. Analytically it covers 0.85 of the threat classes against at most 0.50 for single-focus baselines; in simulation its dual-residual monitor detects slow drift and residual-aware evasion that single-residual monitoring misses, but not constant offsets, and it cannot by itself separate plant ageing from attack. Testbed validation is the necessary next step.

## References


[1] M. Grieves and J. Vickers, “Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems,” in Transdisciplinary Perspectives on Complex Systems. Cham, Switzerland: Springer, 2017, pp. 85–113.

[2] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, “Digital twin in industry: State-of-the-art,” IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2405–2415, 2019.

[3] A. Fuller, Z. Fan, C. Day, and C. Barlow, “Digital twin: Enabling technologies, challenges and open research,” IEEE Access, vol. 8, pp. 108952–108971, 2020.

[4] R. Minerva, G. M. Lee, and N. Crespi, “Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models,” Proc. IEEE, vol. 108, no. 10, pp. 1785–1824, 2020.

[5] R. Langner, “Stuxnet: Dissecting a cyberwarfare weapon,” IEEE Security & Privacy, vol. 9, no. 3, pp. 49–51, 2011.

[6] R. M. Lee, M. J. Assante, and T. Conway, “Analysis of the cyber attack on the Ukrainian power grid,” E-ISAC/SANS ICS Defense Use Case, Mar. 2016.

[7] A. Di Pinto, Y. Dragoni, and A. Carcano, “TRITON: The first ICS cyber attack on safety instrument systems,” in Proc. Black Hat USA, 2018.

[8] A. Humayed, J. Lin, F. Li, and B. Luo, “Cyber-physical systems security—A survey,” IEEE Internet Things J., vol. 4, no. 6, pp. 1802–1831, 2017.

[9] K. Stouffer et al., “Guide to operational technology (OT) security,” NIST Special Publication 800-82 Rev. 3, 2023.

[10] C. Alcaraz and J. Lopez, “Digital twin: A comprehensive survey of security threats,” IEEE Commun. Surveys Tuts., vol. 24, no. 3, pp. 1475–1503, 2022.

[11] M. Eckhart and A. Ekelhart, “Digital twins for cyber-physical systems security: State of the art and outlook,” in Security and Quality in Cyber-Physical Systems Engineering. Cham, Switzerland: Springer, 2019, pp. 383–412.

[12] M. Eckhart and A. Ekelhart, “Towards security-aware virtual environments for digital twins,” in Proc. ACM Workshop Cyber-Phys. Syst. Security (CPSS), 2018, pp. 61–72.

[13] M. Dietz and G. Pernul, “Unleashing the digital twin's potential for ICS security,” IEEE Security & Privacy, vol. 18, no. 4, pp. 20–27, 2020.

[14] M. Dietz, M. Vielberth, and G. Pernul, “Integrating digital twin security simulations in the security operations center,” in Proc. ARES, 2020, pp. 1–9.

[15] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” Pattern Recognit., vol. 84, pp. 317–331, 2018.

[16] J. Gama et al., “A survey on concept drift adaptation,” ACM Comput. Surv., vol. 46, no. 4, pp. 1–37, 2014.

[17] S. Amershi et al., “Software engineering for machine learning: A case study,” in Proc. IEEE/ACM ICSE-SEIP, 2019, pp. 291–300.

[18] MITRE, “ATT&CK for ICS” and “ATLAS: Adversarial threat landscape for artificial-intelligence systems.” [Online]. Available: https://attack.mitre.org/matrices/ics/ and https://atlas.mitre.org/

[19] Automation Systems and Integration—Digital Twin Framework for Manufacturing, ISO 23247, 2021.

[20] Security for Industrial Automation and Control Systems, IEC 62443 Series, 2018–2022.

[21] S. Suhail, R. Jurdak, R. Matulevičius, and C. S. Hussain, “Trustworthy digital twins in the industrial Internet of Things with blockchain,” IEEE Internet Comput., vol. 26, no. 3, pp. 58–67, 2022.

[22] T. Gu, B. Dolan-Gavitt, and S. Garg, “BadNets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv:1708.06733, 2017.

[23] M. Jagielski et al., “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,” in Proc. IEEE Symp. Security Privacy (S&P), 2018, pp. 19–35.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. ICLR, 2015.

[25] A. Madry et al., “Towards deep learning models resistant to adversarial attacks,” in Proc. ICLR, 2018.

[26] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Proc. IEEE Symp. Security Privacy (S&P), 2017, pp. 39–57.

[27] F. Tramèr et al., “Stealing machine learning models via prediction APIs,” in Proc. USENIX Security Symp., 2016, pp. 601–618.

[28] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models,” in Proc. IEEE Symp. Security Privacy (S&P), 2017, pp. 3–18.

[29] A. Vassilev et al., “Adversarial machine learning: A taxonomy and terminology of attacks and mitigations,” NIST AI 100-2, 2024.

[30] J. Steinhardt, P. W. Koh, and P. Liang, “Certified defenses for data poisoning attacks,” in Proc. NeurIPS, 2017, pp. 3517–3529.

[31] A. A. Cárdenas et al., “Attacks against process control systems: Risk assessment, detection, and response,” in Proc. ACM ASIACCS, 2011, pp. 355–366.

[32] J. Giraldo et al., “A survey of physics-based attack detection in cyber-physical systems,” ACM Comput. Surv., vol. 51, no. 4, pp. 1–36, 2018.

[33] Y. Mo and B. Sinopoli, “Secure control against replay attacks,” in Proc. Allerton Conf. Commun., Control, Comput., 2009, pp. 911–918.

[34] C. M. Ahmed, J. Zhou, and A. P. Mathur, “Noise matters: Using sensor and process noise fingerprint to detect stealthy cyber attacks and authenticate sensors in CPS,” in Proc. ACSAC, 2018, pp. 566–581.

[35] M. Kravchik and A. Shabtai, “Detecting cyber attacks in industrial control systems using convolutional neural networks,” in Proc. ACM Workshop Cyber-Phys. Syst. Security Privacy (CPS-SPC), 2018, pp. 72–83.

[36] A. P. Mathur and N. O. Tippenhauer, “SWaT: A water treatment testbed for research and training on ICS security,” in Proc. Int. Workshop Cyber-Phys. Syst. Smart Water Netw. (CySWater), 2016, pp. 31–36.

[37] A. Erba et al., “Constrained concealment attacks against reconstruction-based anomaly detectors in industrial control systems,” in Proc. ACSAC, 2020, pp. 480–495.

[38] D. I. Urbina et al., “Limiting the impact of stealthy attacks on industrial control systems,” in Proc. ACM CCS, 2016, pp. 1092–1105.

[39] National Institute of Standards and Technology, “Artificial intelligence risk management framework (AI RMF 1.0),” NIST AI 100-1, 2023.

[40] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, “Zero trust architecture,” NIST Special Publication 800-207, 2020.

[41] Directive (EU) 2022/2555 on Measures for a High Common Level of Cybersecurity Across the Union (NIS2 Directive), European Parliament and Council, 2022.

[42] National Cyber Security Centre and Cybersecurity and Infrastructure Security Agency, “Guidelines for secure AI system development,” 2023.

[43] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, “The ML test score: A rubric for ML production readiness and technical debt reduction,” in Proc. IEEE Big Data, 2017, pp. 1123–1132.

[44] A. Rasheed, O. San, and T. Kvamsdal, “Digital twin: Values, challenges and enablers from a modeling perspective,” IEEE Access, vol. 8, pp. 21980–22012, 2020.

[45] F. Pasqualetti, F. Dörfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” IEEE Trans. Autom. Control, vol. 58, no. 11, pp. 2715–2729, 2013.

[46] H. Fawzi, P. Tabuada, and S. Diggavi, “Secure estimation and control for cyber-physical systems under adversarial attacks,” IEEE Trans. Autom. Control, vol. 59, no. 6, pp. 1454–1467, 2014.

[47] J. Chen and R. J. Patton, Robust Model-Based Fault Diagnosis for Dynamic Systems. Boston, MA, USA: Kluwer, 1999.

[48] A. Karpatne et al., “Theory-guided data science: A new paradigm for scientific discovery from data,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 10, pp. 2318–2331, 2017.

[49] E. S. Page, “Continuous inspection schemes,” Biometrika, vol. 41, no. 1/2, pp. 100–115, 1954.

[50] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, “A dataset to support research in the design of secure water treatment systems,” in Proc. CRITIS, 2016, pp. 88–99.

[51] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, “WADI: A water distribution testbed for research in the design of secure cyber physical systems,” in Proc. CySWATER, 2017, pp. 25–28.

# A Unified On-Device Classifier for NSFW and Malicious Web Pages Using Screenshot

Adhishtanaka Kulasooriya
*Faculty of Computing*
*NSBM Green University, Homagama,*
*Sri Lanka*
aatkulasooriya@students.nsbm.ac.lk

R. Ranaweera
*Faculty of Computing*
*NSBM Green University, Homagama,*
*Sri Lanka*
ranaweera.r@nsbm.ac.lk

***Abstract*—Parental control software splits the two hazards children meet online across separate products: NSFW filters read pixels, malicious blocklists read addresses, and most run on a server or a router that sees the traffic it filters. This paper presents a single convolutional network that assigns a rendered web page to one of three classes, normal, NSFW or malicious, entirely inside the browser. It reads a screenshot as five views, one whole page and four native-resolution crops combined by max pooling, alongside the page text recovered from that same screenshot by optical character recognition. Training used a purpose-built corpus of 10,639 cleaned screenshots partitioned by site and by search query, so no test page comes from a host seen during training. On 1,649 held-out pages the classifier reaches 0.946 macro-F1 (95% bootstrap interval 0.932 to 0.959) and 0.957 accuracy at 59.9 ms per page on CPU, from a 4.7 MB TensorFlow Lite file in a Manifest V3 extension. Blanking the text input costs 0.116 macro-F1 overall and 0.145 on search result pages, whose appearance carries almost no class information. The malicious class covers scareware, fake technical support screens and fake download prompts. Credential phishing is deliberately excluded and is not detected: a preliminary test on an earlier, simpler model showed that including it degraded the normal class as well as its own, with normal-class F1 rising from 0.706 to 0.919 once it was removed, which bounds what a classifier trained on pixels can be asked to decide.**

***Keywords—parental control, web content classification, convolutional neural network, on-device inference, optical character recognition***

## I. Introduction

Children browse unsupervised for much of the time they spend online, and the material they meet falls into two families with almost nothing in common technically: objectionable content, meaning pornography, drug imagery and self-harm material, usually grouped under the label NSFW; and malicious content, meaning scareware, fake technical support screens and counterfeit download portals. Automated controls have been found to moderate the behavioural effect of games and video apps more reliably than settings a parent applies by hand [1]. Reviews of child safety on large video and gaming platforms report that harmful material still reaches young viewers despite the moderation those platforms already run [2], [3].

The tools a parent can install treat the two families separately: NSFW filters classify pixels, blocklists and reputation services classify addresses. A parent who wants one answer to one question, is this page safe for my child, must install several products and reconcile them.

The research literature has the same shape. Visual NSFW detection works on page or image pixels, using skin segmentation, nude-region attention or pornographic frame recognition, and never looks at the address bar [4], [5], [6], [7]. Malicious URL and phishing detection learns from the characters of the link or from the HTML, and never looks at what the page renders [8], [9], [10], [11], [12], [13], [14], [15], [16]. Both branches are almost always framed as a binary decision, and the multimodal systems that do bridge them are heavy pipelines that need HTML and run on a server or a router [17], [18].

Placement matters as much as modality. A text classifier built for home routers reaches 0.92 accuracy but takes 1.86 seconds per page [19], too slow to sit in front of every navigation, while systems that call a reputation service add latency and hand a child's browsing history to a third party. For protection that runs continuously on a family machine, neither cost is acceptable.

This paper closes that split. One lightweight convolutional network reads a screenshot of the rendered page together with the text recovered from that same screenshot by optical character recognition, and assigns one of three classes: normal, NSFW or malicious, inside a Chrome Manifest V3 extension on the user's own device, with no network request in the inference path. Credential phishing is excluded by design and is not detected (Sections III-C and VI). The contributions are as follows.

1) A three-class on-device classifier fusing a rendered screenshot with text recovered from that screenshot, needing no page source, no address string and no external service, at 0.946 macro-F1 and 59.9 ms per page from a 4.7 MB model.

2) An ablation isolating what the recovered text contributes, showing the gain concentrates on pages whose appearance carries no class information.

3) An account of why credential phishing does not belong in a pixel-based malicious class: a preliminary test on an earlier, simpler model showed normal-class F1 rising from 0.706 to 0.919 when phishing was excluded, indicating the cost of including it had fallen on a class it had nothing to do with.

## II. Related Work

### *A. Visual detection of objectionable content*

Reviews trace the field from hash matching and hand-crafted features to deep networks, and note that performance depends on how varied the training data is [4]. The recurring weakness

of convolution-only classifiers is the small region problem: nudity in a small part of a busy frame produces high false negative rates. AlDahoul et al. point a YOLO detector at candidate regions before classification, which cuts missed detections [5]. PornNet fuses an image stream with an audio stream through attention at about 0.93 accuracy [6]. Skin detection is the usual preprocessing layer, separating skin from non-skin pixels under uncontrolled lighting [7]. NSFW is therefore largely a visual signal, but none of these methods reads anything but pixels.

*B. Malicious URL and phishing detection*

This branch is sequence oriented. Character-level CNNs classify a URL from its characters without fetching the page or querying a service, which is fast and private [9]. Later work reports 99% on a multi-class URL corpus with an ELECTRA-based language model [8], about 98.5% with a BERT and character BiGRU hybrid [13], and 99.2% to 99.7% for CNN, LSTM and 1D-CNN phishing detectors [14], [15]. Two results matter more than the headline figures: SwiftURL distils a transformer so that detection costs 35% less computation on embedded hardware [12], and other work adds logic constraints or threat feeds to hold accuracy across datasets [10], [11].

Two properties bear on the present design. The URL is a strong content-free signal, but a model trained on it learns domains rather than content and degrades on unseen hosts; SemanticPhishNet reports external accuracy falling to 67% [16]. And every one of these models is binary and reads only the address or the HTML, so none separates a malicious page from an objectionable one.

*C. Multimodal and parental control systems*

Three systems sit close to this work. Janavičiūtė et al. classify page text on a home router with a linear SVM, reaching 0.9198 accuracy across seven categories at 1.86 s per page [19]; the placement resists tampering, but the latency rules it out in front of a browser. Chen et al. fuse Doc2Vec text features with a bag of visual words and exceed 0.99 accuracy on pornographic and gambling sites [17], which supports the intuition that the two signals complement each other, though the features are engineered and the system needs HTML. MEDAL combines image, text and HTML in a tri-training framework [18], but its three-network pipeline is far too heavy for client-side inference. A released ground truth corpus of 8,000 labelled sites also argues that filtering accuracy depends on label quality as much as on the model [20]. No reported system reads the rendered page and its own recovered text together, decides among normal, NSFW and malicious in one pass, and runs on the device fast enough for interactive browsing.

## III. Dataset

No existing corpus pairs rendered webpage screenshots with three-way labels, so one was built.

*A. Collection*

Two methods were combined. A Playwright headless browser navigates to a URL, checks that it is live, screenshots it, and appends a manifest row recording file, resolved URL, host, category and source. This worked for top sites, search result pages and pornography, whose hosts stay reachable, but failed for scam and scareware pages, which go offline within hours: a direct attempt reached none of 57 sampled hosts. Those came from urlscan.io, which stores screenshots captured earlier. Because many malicious URLs were unavailable or blank, synthetic malicious screenshots were also generated to increase the size and diversity of this class.

Source lists came from auto-fetchable feeds: Tranco for legitimate sites, OpenPhish, PhishTank and URLhaus for phishing and malware, and curated host blocklists for NSFW material, with the category assigned by the feed. Each urlscan.io query was inspected by hand before bulk download, because several of its tags match blank or legitimate pages. Adult and live malicious crawling ran inside a network-isolated container with no persistence, and child sexual abuse material was never sought, stored or processed.

*B. Cleaning*

Collection produced 14,580 screenshots and cleaning left 10,639. Dead pages, parked domains, blank captures and near duplicates become visible only after capture. Every capture is read once by Tesseract into a single file that later stages reuse.

The rule doing most of the work compares the brand named on the page against the domain in the address bar and flags a mismatch. OCR drops dots and confuses glyphs in an address bar, so the rule compares the flattened strings fuzzily and fires only when the brand appears near the top of the page.

Near duplicates are grouped by perceptual hash so that a whole group can be judged in one action. Union-find chained A to B to C into one group of 347 images spanning all three classes, which is dangerous when the tool acts on whole groups; leader clustering, which requires every member to resemble the group leader, cut the largest group to 197 and made 80% of groups single class. The automated pass removed 2,597 images and manual review a further 1,344.

*C. Composition, labels and the split*

The cleaned corpus holds 6,886 normal (Tranco top sites, search results), 2,708 NSFW and 1,045 malicious images. The malicious class is deliberately narrow, covering scareware, fake technical support screens and fake download prompts. Credential phishing is excluded, and the exclusion rests on a measurement rather than a prior judgement. An earlier corpus of 13,338 captures included it, scored on a simpler single-view model with no text branch rather than the final architecture: from pixels alone, per-class F1 was 0.372 on malicious, 0.706 on normal and 0.811 on NSFW, while with credential phishing removed under that same simpler setup, per-class F1 was 0.754, 0.919 and 0.817. That the malicious class improved is expected; that the normal class improved by more than twenty points is not. The phishing pages had pulled ordinary sign-in screens toward the malicious label, so the cost of keeping them fell on a class they had nothing to do with.

The two runs differ in more than the class definition, since the later one reads five views rather than one, so the comparison is indicative rather than a controlled ablation, but the direction is not in doubt.

The split is 70/15/15 and group aware: every page from one site goes to one partition. Search result pages are grouped by

query rather than by host, because google.com alone accounts for several thousand images across two classes, and treating it as one group would leave the text ablation with almost no test coverage. The split code asserts that no group appears on both sides. The result is 7,379 training, 1,611 validation and 1,649 test images.

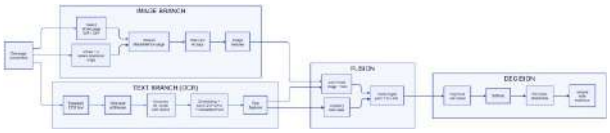


Fig. 1. The two branches. Five views of one screenshot pass through a shared MobileNetV3-Large backbone, and the OCR text of that same screenshot, with web addresses removed, passes through a convolutional text branch; the heads meet at a fused logit before the decision layer. The keyword rule boost is drawn because it belongs to the searched decision rule, though the search set its weight to zero.

## IV. Method

### A. Five views of one screenshot

Shrinking a 1366-pixel-wide page to the 224 pixels the network expects reduces a video thumbnail to roughly 35 pixels, at which size a page is still recognisable as a grid of thumbnails but what the thumbnails contain is gone. Each screenshot is therefore read as five views (Fig. 1): the whole page letterboxed to 224 by 224, and four crops taken at native resolution and then resized. All five pass through one shared MobileNetV3-Large backbone pretrained on ImageNet, and the four crops are combined by taking the maximum, so a page counts as unsafe if any region of it is. Shared weights make five views cost extra forward passes but no extra model size. This is the small region problem that AlDahoul et al. solve with a YOLO detector [5]; fixed crops with max pooling serve the same purpose without a second model and without region labels.

### B. Text branch and fusion

The OCR text passes through an embedding and three parallel convolutional filters of width 3, 4 and 5 with global max pooling. Image and text features meet at a joint head, and a second head reads text alone; the output is the joint logit plus a weight times the text logit. Text dropout of 15% keeps the model classifying when OCR returns nothing. Augmentation excludes hue shift and greyscale conversion, because skin tone is a primary signal for the NSFW class. Training freezes the backbone while the heads are trained, then unfreezes its top layers at a lower learning rate, with early stopping and rate reduction on plateau. Class weights inversely proportional to class frequency keep the malicious class, roughly six times rarer than normal, from being drowned out.

### C. Decision parameters

The text weight, the weight given to a small set of high-precision keyword rules, and a threshold for each risky class are searched rather than chosen: 9,216 combinations evaluated on cached validation logits. Three properties of the search were added after watching it misbehave. The untuned baseline sits inside the search space, so tuning cannot return something worse than not tuning. A constraint on precision alone proved gameable, so a second requires every class to be predicted at a reasonable share of its true frequency. And because searching thousands of settings against one validation set overfits it, the one-standard-error rule selects the simplest setting statistically tied with the best.

Two of the four knobs came back switched off. The keyword rules are a short list of high-precision phrases such as ‘your computer is infected’ and ‘call microsoft’, carrying no domain or URL term so that they cannot reintroduce the address-bar shortcut Section III-C guards against. On validation they fire on 19.5% of NSFW pages and 4.3% of malicious pages and never on a normal page, yet the search weighted them at zero: the best grid setting scored 0.9737 validation macro-F1, 1,288 settings fell within one standard error of it, and the simplest of those, carrying no rule weight and no threshold override, scored 0.9726. What survives is a text weight of 0.25, which lifts validation macro-F1 from 0.9672 to 0.9726 and on test raises macro-F1 from 0.9436 to 0.9462 and malicious recall from 0.912 to 0.932.

A learned text branch therefore subsumes a hand-curated phrase list, the component that ages fastest as scam wording changes.

### D. Export and deployment

The model converts to TensorFlow Lite with dynamic range quantisation, giving a 4.7 MB file, with metadata recording class order, vocabulary, input dimensions and the searched parameters, so the browser loads exactly the network that was evaluated. Two constraints appeared only at export. LiteRT.js accepts only int32 and float32 tensors while the Keras text layer emits int64, so a model exported without a cast converts but refuses to load; and the tab capture already excludes browser chrome, so no address-bar crop is applied at inference.

A content script reports page changes, including route changes in single-page applications and large DOM changes, because a scam overlay may appear seconds after load. An offscreen document holds the compiled model and the OCR worker, because a Manifest V3 service worker has no document and is evicted aggressively, forcing a recompile on every navigation. The parent’s PIN lives in the service worker as a PBKDF2 hash; the warning overlay is injected into the page under examination, which may be hostile, so it never receives that hash and never decides anything. The model, both runtimes and the OCR language data are bundled, and the network panel shows no request leaving the extension.

## V. Evaluation

The evaluation runs in a fixed order, and the order guards against test-set leakage. The split is made first and asserted leak-free; backbones are compared on validation with the backbone frozen, which ranks architectures without a full training run each; the chosen model is then trained fully; and decision parameters are searched on validation, never on test. Only then is the test partition scored, once. Macro-F1 is reported because the classes are imbalanced, precision and recall per class because a missed malicious page and a wrongly blocked homework page cost differently, and CPU latency because the target has no GPU.

### A. Backbone against cost

TABLE I
BACKBONE COMPARISON ON THE VALIDATION PARTITION

| Backbone | Macro-F1 | Params (M) | TFLite (KB) | CPU (ms) |
|---|---|---|---|---|
| EfficientNetB0 | 0.9716 | 5.34 | 6,051 | 131.1 |
| MobileNetV3-Large | 0.9706 | 4.20 | 4,606 | 59.9 |
| MobileNetV2 | 0.9692 | 3.55 | 3,857 | 73.5 |
| MobileNetV3-Small | 0.9664 | 2.05 | 2,364 | 23.3 |
| EfficientNetV2-B0 | 0.9380 | 7.21 | 8,081 | 87.4 |

Five backbones were trained identically for four epochs with the backbone frozen (Table I, Fig. 2). EfficientNetB0 is the most accurate by one tenth of a percentage point, at 131 ms per page and 6 MB. EfficientNetV2-B0 is beaten on every axis. MobileNetV3-Small is the most efficient per millisecond, but MobileNetV3-Large was chosen because it sits within 0.1 points of the best macro-F1 at less than half the latency, and the marginal accuracy of the larger model does not justify more than doubling the cost of a check that runs on every page. Weighing accuracy against latency follows the backbone comparison of Qawasmeh et al. [21] and the trade SwiftURL makes deliberately for constrained hardware [12].

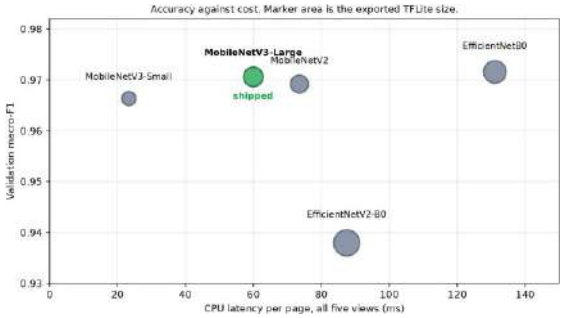



Fig. 2. Validation macro-F1 against CPU latency for the five candidate backbones. Marker area is exported model size.

### B. What the recovered text contributes

TABLE II
PER-CLASS F1 WITH THE TEXT CHANNEL BLANKED AND RESTORED

| Class | Image only | Image and text | Gain |
|---|---|---|---|
| malicious | 0.7536 | 0.9388 | +0.1852 |
| nsfw | 0.8170 | 0.9310 | +0.1140 |
| normal | 0.9190 | 0.9687 | +0.0497 |

The text input was blanked and the same trained model re-scored, isolating the contribution of the text channel (Table II). The macro gain is 0.1163. On the 349 test pages that are visually near identical across classes, namely search result pages, macro-F1 rises from 0.7814 to 0.9261, a gain of 0.1447. That is larger than the gain over the test set as a whole, and it falls where the design predicts: two Google results pages, one for a cake recipe and one for buying a controlled drug, are the same image and differ only in words. Fig. 3 shows that case reaching a user-visible action.

Blanking the text was verified to move the model's outputs (mean logit shift 2.0967), so the table cannot be an ablation that silently changed nothing. The direction agrees with earlier multimodal work [17], [18]. What is new is that the second modality is recovered from the screenshot itself, needing no page source and no network request.

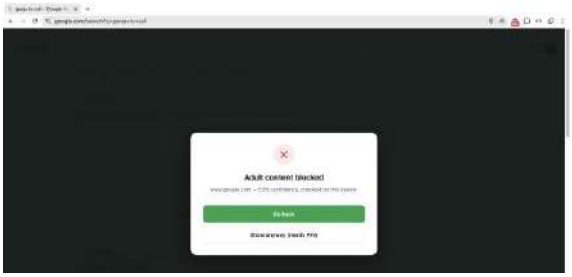


Fig. 3. The deployed decision on a search results page for a drug query. As an image this is an ordinary results page and nothing in the pixels separates it from a search for a cake recipe; the class comes from the words read off the same capture. The only route back to the page is gated by the parent's PIN.

### C. Over-blocking against missed threats

The malicious probability threshold was swept on validation against a stated budget of 5% over-blocking, and the cost was then measured on test, so it was never tuned on the data used to report it. The lowest threshold inside the budget was 0.10, where the system over-blocks 4.16% of legitimate pages, misses 6.31% of NSFW pages and 6.76% of malicious ones, which are exactly the miss rates obtained with no threshold override at all. The ten missed malicious pages all score below 0.10, so lowering the bar recovers none of them and only raises over-blocking from 3.26% to 4.16%.

The shipped metadata records a threshold of 1.01 for both risky classes, a value no probability can reach, so the deployed decision is a plain argmax. The value 1.01 was placed in the search grid of Section IV-C so that no threshold at all could compete on equal terms, and the one-standard-error rule selected it. Over-blocking is the binding constraint: a filter that keeps blocking schoolwork is switched off and protects nobody.

### D. Overall performance

TABLE III
PERFORMANCE ON THE HELD-OUT TEST PARTITION

| Class | Precision | Recall | F1 | Pages |
|---|---|---|---|---|
| malicious | 0.945 | 0.932 | 0.939 | 148 |
| normal | 0.970 | 0.967 | 0.969 | 1,105 |
| nsfw | 0.925 | 0.937 | 0.931 | 396 |
| Macro average | 0.947 | 0.946 | 0.946 | 1,649 |

Every page in the test partition comes from a site, or a search query, the model never saw during training. Accuracy is 0.957 and balanced accuracy 0.946. Cohen's kappa and the Matthews correlation coefficient are both 0.911, reported because the normal class is two thirds of the test set. The errors are not evenly spread (Fig. 4): most confusion is between normal and

NSFW in both directions, and of 148 malicious pages one was classified as NSFW and nine as normal, while seven normal pages were wrongly flagged malicious. Bootstrap 95% intervals (10,000 resamples of the test predictions) are 0.932 to 0.959 for macro-F1 and, for the malicious class, 0.908 to 0.965 for F1, 0.889 to 0.970 for recall and 0.905 to 0.979 for precision, roughly four times as wide as for the normal class (F1 0.961 to 0.976) because only 148 pages are available.

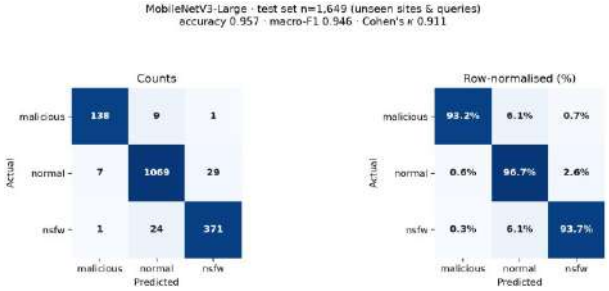



Fig. 4. Confusion matrix on the held-out test set of 1,649 pages.

### *E. Positioning against reviewed systems*

TABLE IV
PROPOSED SYSTEM BESIDE THE CLOSEST REVIEWED SYSTEMS

| System | Input | Classes | Runs on | Reported result |
|---|---|---|---|---|
| Janavičiūtė et al. [19] | page text | 7 | home router | 0.9198 acc., 1,860 ms |
| Chen et al. [17] | HTML text, image features | 3 | router | 0.9918 acc. |
| MEDAL [18] | screenshot, HTML, OCR | 4 | server | 0.95-0.98 acc. |
| SwiftURL [12] | URL string | 2 | embedded | 35% less computation |
| SemanticPhish Net [16] | HTML and URL text | 2 | server | 67% external |
| Proposed system | screenshot and its OCR text | 3 | browser, on device | 0.957 acc., 0.946 macro-F1, 59.9 ms, 4.7 MB |

Every row of Table IV was measured on a different corpus, so the table positions the work rather than benchmarking it. Two columns carry the argument: no other system decides among three classes in one model, and no other system runs in the browser. The most directly comparable row is Janavičiūtė et al. [19], which shares the application and reports its latency.

## VI. LIMITATIONS

No external benchmark exists for this task, and that bounds what the results establish. The released datasets of the reviewed studies are either URL strings [8], [9], [10]-[16], or photographs and video frames [4], [5]. Neither can be fed to a model that reads a rendered page. MEDAL states that no public dataset of illegal websites exists and crawled its own [18], and Chen et al. give no availability statement [17]. The one published corpus that shares this application [19] holds 2,648 URLs and their categories but no screenshots and no malicious category; its adult category holds 115 pages. The results here establish generalisation to sites and search queries held out from this corpus, not to a corpus someone else collected.

The ablation in Section V-B is inference-time on one trained model: it measures how far the decision leans on the text channel, not what training with text changed, which would need a separate image-only model.

The system does not detect credential phishing, and that is a limit on the protection it offers, not a footnote about how the dataset was built: a child who reaches a cloned bank, game or school login page will have it classified as normal and will not be warned. As images, a genuine sign-in page and a pixel-accurate copy are the same photograph, and reading the page text does not rescue the case either, because a convincing clone copies the wording too. The malicious figures in Table III are therefore measured on the narrower class defined in Section III-C, whose visual template is itself the giveaway.

What separates a real login from a fake one is the address, which this design never reads. The malicious class also holds only 1,045 images, 148 of them in the test partition, so its figures carry a wider interval than the other two (Section V-D). Closing the gap means a character-level network over the domain, path and query [8], joined at the decision layer rather than inside the network, so the fused model keeps reading only what appears on the page. The cost is that the system becomes two networks with a policy between them, which is more to ship and harder to reason about when it is wrong.

Testing inside the browser was manual and is the weaker half of the evidence. Forty-six automated checks cover PIN hashing, lockout behaviour and host matching for both site lists, and one further check compares the browser tokenizer against Keras on 200 sequences of real captured text, because a drifting tokenizer feeds noise to the text branch while nothing visibly breaks. An automated end-to-end harness, driving a scripted browser through the packaged extension on a labelled page set, is the obvious next step and would let in-browser accuracy be reported against the offline figures.

The malicious class pools scareware, fake support screens and fake download prompts with no per-subclass figures, and includes synthetic screenshots that may be easier than live pages; growing it from live captures and reporting subclasses separately is the most useful extension. The backbone ranking in Table I used a frozen backbone for four epochs and its top three sit within 0.3 points, so a full fine-tuned comparison could reorder them.

## VII. CONCLUSION

A single 4.7 MB convolutional model classifies rendered web pages into normal, NSFW and malicious at 0.946 macro-F1 and 59.9 ms per page on CPU, inside a browser extension with no network request in the inference path. The second modality comes from the screenshot it already has: OCR text contributes 0.116 macro-F1 overall and 0.145 on pages whose appearance carries no class information. Credential phishing sits outside the boundary of what a pixel model can be asked to

separate, and naming that boundary is as much a result as the accuracy figure. A character-level branch over the address, joined at the decision layer, is the next step.

## REFERENCES


[1] O. Alrusaini and H. Beyari, "The Sustainable Effect of Artificial Intelligence and Parental Control on Children's Behavior While Using Smart Devices' Apps: The Case of Saudi Arabia," *Sustainability (Switzerland)*, vol. 14, no. 15, Aug. 2022, doi: 10.3390/su14159388.

[2] S. I. Alqahtani, W. M. S. Yafooz, A. Alsaeedi, L. Syed, and R. Alluhaibi, "Children's Safety on YouTube: A Systematic Review," Mar. 01, 2023, *MDPI*. doi: 10.3390/app13064044.

[3] M. Chawki, "AI Moderation and Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox," *Laws*, vol. 14, no. 3, Jun. 2025, doi: 10.3390/laws14030029.

[4] D. Povedano Álvarez, A. L. Sandoval Orozco, J. P. García-Miguel, and L. J. García Villalba, "Learning Strategies for Sensitive Content Detection," Jun. 01, 2023, *MDPI*. doi: 10.3390/electronics12112496.

[5] N. Aldahoul *et al.*, "Transfer detection of yolo to focus cnn's attention on nude regions for adult content detection," *Symmetry (Basel).*, vol. 13, no. 1, pp. 1–26, Jan. 2021, doi: 10.3390/sym13010026.

[6] Z. Fu, J. Li, G. Chen, T. Yu, and T. Deng, "PornNet: A Unified Deep Architecture for Pornographic Video Recognition," 2021, doi: 10.3390/app11073066.

[7] M. Leite, W. D. Parreira, A. M. da R. Fernandes, and V. R. Q. Leithardt, "Image Segmentation for Human Skin Detection," *Applied Sciences (Switzerland)*, vol. 12, no. 23, Dec. 2022, doi: 10.3390/app122312140.

[8] F. Türk and M. Kılıçaslan, "Malicious URL Detection with Advanced Machine Learning and Optimization-Supported Deep Learning Models," *Applied Sciences (Switzerland)*, vol. 15, no. 18, Sep. 2025, doi: 10.3390/app151810090.

[9] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," *Electronics (Switzerland)*, vol. 9, no. 9, pp. 1–24, Sep. 2020, doi: 10.3390/electronics9091514.

[10] J. Zhou *et al.*, "A Malicious URL Detection Framework Based on Custom Hybrid Spatial Sequence Attention and Logic Constraint Neural Network," *Symmetry (Basel).*, vol. 17, no. 7, Jul. 2025, doi: 10.3390/sym17070987.

[11] F. A. Ghaleb, M. Alsaedi, F. Saeed, J. Ahmad, and M. Alasli, "Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning," *Sensors*, vol. 22, no. 9, May 2022, doi: 10.3390/s22093373.

[12] Z. Y. Lim, Y. H. Pang, E. Chan Kah Jun, S. Y. Ooi, and S. Y. Wee, "SwiftURL: A Lightweight Transformer-Based Model for Malicious URL Detection," *Applied Sciences*, vol. 16, no. 7, p. 3366, Mar. 2026, doi: 10.3390/app16073366.

[13] Y. Y. Munaye, A. B. Workneh, Y. B. Chekol, and A. M. Mekonen, "Effective Detection of Malicious Uniform Resource Locator (URLs) Using Deep-Learning Techniques," *Algorithms*, vol. 18, no. 6, Jun. 2025, doi: 10.3390/a18060355.

[14] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN," *Electronics (Switzerland)*, vol. 12, no. 1, Jan. 2023, doi: 10.3390/electronics12010232.

[15] Q. E. ul Haq, M. H. Faheem, and I. Ahmad, "Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks," *Applied Sciences (Switzerland)*, vol. 14, no. 22, Nov. 2024, doi: 10.3390/app142210086.

[16] E.-H. Qazi, M. H. Faheem, and A. Almorjan, "Phishing URL Detection Using Transformer-Based Architecture and Contextual Content Features," *Computers*, vol. 15, no. 6, p. 335, May 2026, doi: 10.3390/computers15060335.

[17] Y. Chen, R. Zheng, A. Zhou, S. Liao, and L. Liu, "Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism," *Sensors (Switzerland)*, vol. 20, no. 14, pp. 1–21, Jul. 2020, doi: 10.3390/s20143989.

[18] L. Wen *et al.*, "MEDAL: A Multimodality-Based Effective Data Augmentation Framework for Illegal Website Identification," *Electronics (Switzerland)*, vol. 13, no. 11, Jun. 2024, doi: 10.3390/electronics13112199.

[19] A. Janavičiūtė, A. Liutkevičius, and N. Morkevičius, "Toward the Implementation of Text-Based Web Page Classification and Filtering Solution for Low-Resource Home Routers Using a Machine Learning Approach," *Electronics (Switzerland)*, vol. 14, no. 16, Aug. 2025, doi: 10.3390/electronics14163280.

[20] H. H. M. Altarturi and N. B. Anuar, "Ground Truth Dataset: Objectionable Web Content," *Data (Basel).*, vol. 7, no. 11, Nov. 2022, doi: 10.3390/data7110153.

[21] B. Qawasmeh, J. S. Oh, and V. Kwigizile, "Comparative Analysis of AlexNet, ResNet-50, and VGG-19 Performance for Automated Feature Recognition in Pedestrian Crash Diagrams," *Applied Sciences (Switzerland)*, vol. 15, no. 6, Mar. 2025, doi: 10.3390/app15062928.

# MBR Integrity Protection and Automated Recovery System Against Boot-Sector Ransomware

Sanoj Aminda
*Faculty of Science and Engineering*
*University of Plymouth*
*Plymouth, United Kingdom*
sanojaminda@gmail.com

Chamara Disanayake
*Dept. of Software Engineering &*
*Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
chamara.d@nsbm.ac.lk

Kaveesha Gimhani
*Faculty of Science and Engineering*
*University of Plymouth*
*Plymouth, United Kingdom*
kaveeshagimhanit@gmail.com

Madusanka Mithrananda
*Dept. of Software Engineering &*
*Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
madusanka.m@nsbm.ac.lk

***Abstract*—Petya and HybridPetya are examples of boot-sector ransomware that seriously damage Windows system by overwriting the Master Boot Record (MBR) and then corrupting the Master File Table (MFT), making the system unbootable. For small and medium-sized businesses, manual MBR recovery is too much work because it requires special forensic tools and expert knowledge. The goal of this project is to create AegisX, a Windows system for real-time kernel-level protection and recovery for Master Boot Record. AegisX was set up as a multi-part system that included a Windows Driver Framework Kernel-mode Storage Filter Driver, a Minifilter, a Windows Background Service, a System Tray Application, and a Command Line Interface. The Storage Filter Driver intercepts write I/O Request Packets (IRPs) targeting Sector 0 (the MBR) and blocks those that are unauthorized. The Background Service keeps a SHA-256-verified MBR backup in a hidden 150MB raw partition. When a write to the MBR is detected, the Tray Application prompts the administrator to enter the bypass key, allowing legitimate operations such as dual-boot configuration to proceed while unauthorized overwrites are blocked. The Minifilter Driver is intended to verify integrity during NTFS file system activity. Its evaluation remains future work. In testing against a controlled Assembly-based bootloader attack in an isolated virtual machine, the protection pipeline blocked the simulated unauthorized MBR overwrite attempts and restored the MBR from the verified backup. The current implementation targets legacy BIOS/MBR systems.**

***Keywords—Master Boot Record, Destructive Ransomware Attacks, Storage Filter Driver, MBR Recovery, Boot Sector Protection***

## I. Introduction

Ransomware remains a major and constantly changing problem in the field of cybersecurity, inflicting serious financial and reputational damage [1]. The early forms of ransomware worked essentially as simple file-encryption malware, focusing on users' documents and asking for payment in order to provide the decryption keys. Although such attacks resulted in economic loss, they could generally be detected and dealt with using file-level backups and antivirus programs that recognized know malware patterns [2]. The nature of the threat has changed considerably the appearance of destructive ransomware which instead of just encrypting user data targets low-level system structures (such as the MBR and MFT). To defend against such sophisticated threats it is necessary to have a thorough, multi-layered security strategy which takes into account all the possible attack vectors and methods of execution. [3].

Ransomware variants including Petya, NotPetya, and HybridPetya have shown that modern ransomware is capable of overwriting the Master Boot Record (MBR) and corrupting the Master File Table (MFT), making the system completely unbootable and thus maximizing operational disruption [3]. Although these types of attacks are severe, current endpoint security solutions are not properly designed to block threats that operate at the boot-sector level [4]. Even though behavioral analysis has been suggested as a more adaptive approach than traditional signature-based detection, the high number of false positives linked to dynamic analysis continues to pose a major operational issue [2].

Moreover, although defense-in-depth approaches—such as those that use virtual machine isolation and browser extensions—have been suggested, their emphasis is still mainly on user-space downloads and threats at the application level [4]. Importantly, ordinary endpoint security products do not provide real-time monitoring of the integrity of the MBR and MFT and fail to include automated, lightweight methods for recovering corrupted boot sectors. At present, recovery of a compromised boot sector has to be carried out manually, which requires specialized forensic tools, expert technical knowledge, and results in hours of system downtime [5]. Small and Medium Enterprises (SMEs) are particularly affected by this weakness since they usually do not have the financial means and technical skills needed to put in place enterprise-grade security solutions or to carry out complicated manual MBR recoveries. As a result, SMEs are very susceptible to such attacks, and recent survey data shows that a large number of organizations decide simply to pay the ransom rather than spending the effort on complex preventive and recovery measures [6].

In order to overcome these critical gaps in endpoint security, this paper introduces AegisX, a real-time MBR integrity monitoring and automated recovery architecture designed specifically for Windows environments. The AegisX architecture is deployed as a multi-part defense-in-depth system comprising a Windows Driver Framework (WDF) kernel-mode Storage Filter Driver, a Minifilter driver, a background service, and a user-space tray application. The kernel-level Storage Filter Driver is designed to intercept all I/O Request Packets (IRPs) attempting to write to critical boot sectors, effectively blocking unauthorized modification attempts. Simultaneously, the background service allocates available disk space to create a hidden RAW partition at the end of the storage stack, where a cryptographically secure, encrypted backup of the MBR is maintained.

To accommodate legitimate system modifications, such as dual-boot configuration setups, the tray application prompts the user for a SHA-256 generated bypass key when an authorized write is detected. In the event of confirmed, unauthorized MBR corruption, AegisX triggers an automated multi-stage response mechanism. It isolate the host system from the network, terminates suspicious processes, and restores the MBR from the encrypted hidden partition backup.

This research utilizes a prototype-based experimental methodology, enabling the simultaneous execution of system development and empirical testing through continuous, iterative refinement. The primary objective of this study is to demonstrate that automated, boot-level protection against destructive ransomware targeting Master Boot Record (MBR) and Master File Table (MFT) structures is not only technically achievable but also provides a highly practical, lightweight defense mechanism for modern computing environments.

## II. LITERATURE REVIEW

### *A. Evolution of Ransomware Detection and Analysis*

The development of ransomware defense measures has closely followed the increasing complexity of malware. In the early days, ransomware detection depended on signature-based antivirus tools. These systems could effectively neutralize known malware patterns through static analysis by matching the binary signatures of executable files against a regularly updated database of threats. However, as ransomware advanced to exploit zero-day vulnerabilities, use obfuscation, and employ polymorphic code, where the malware continuously changes its structure without altering its function, static detection methods became insufficient.

To overcome these shortcomings, behavior-based detection was introduced as a flexible alternative. Instead of relying solely on system calls, API hooking, and sudden increases in file changes, this approach analyzes behavior over time [2]. Despite its adaptability, behavioral analysis operates reactively. Because it monitors execution, it often initiates a defensive response only after the encryption process has commenced, meaning some degree of data loss is practically guaranteed [7]. Furthermore, it has a high rate of false positives, misclassifying normal, resource-intensive system activities, like database indexing or large file compressions, as threats [2].

To address these standalone issues, recent studies highlight hybrid analysis models that merge static code extraction with dynamic behavioral sandboxing. For example, comparative analyses of advanced variants such as Petya, CryptInfinite, and Locky showed significant differences in their execution environments. Locky often works quietly, mapping network drivers before encrypting user files, while Petya forces an immediate system reboot to disrupt the operating system structure [7]. Understanding these processes has been enhanced through the Cyber Kill Chain methodology, which helps researchers follow the various stages of dangerous ransomware from its initial delivery and exploitation to payload installation and command-and-control communication [8]. This situation became even more complex with the rise of self-propagating ransomware worms that exploit advanced network vulnerabilities, like the EternalBlue SMB flaw, which demonstrated how quickly widespread encryption can occur in unpatched academic and corporate networks [9].

### *B. The Shift to Boot-Sector and Kernel-Level Threats*

While traditional ransomware operates as user-space file-encryption malware, recent years have seen a shift toward destructive boot-sector ransomware, like Petya and the wiper variant, NotPetya [3], [10]. These threats bypass individual user documents to target and damage the Master Boot Record (MBR) and Master File Table (MFT) directly. When Petya is executed, it overwrites the MBR with a harmful custom bootloader, forces a hard system crash, and displays a fake "CHKDSK" screen upon startup while it silently encrypts the MFT [3].

A significant architectural flaw in contemporary endpoint defenses and Endpoint Detection and Responses (EDR) agents is their strict confinement to “user-space” (Ring 3) application monitoring [1], [11]. This protects standard user files but leaves low-level raw disk structures exposed. Attackers increasingly leverage rootkits, bootkits, and kernel-level hijacking (Ring 0) to manipulate dynamically allocated kernel memory [12]. By executing data-oriented attacks, modern malware can blind or completely disable built-in security tools, like Windows Defender, without triggering system alarms [13]. Advanced frameworks that use Kernel Permission Data Protection (KPDP) have been proposed to isolate kernel spaces and prevent memory corruption [14]. Similarly, hardware-based malware detection using low-level architectural features, such as monitoring CPU instruction pipelines and Trusted Platform Module (TPM) attestation has shown promise in intercepting rootkits [15]. However, as highlighted by recent analyses of emerging threats like “BootKitty”, these hardware and virtualization-based solutions face significant implementation barriers. Legacy infrastructure, specialized industrial control systems, and Small and Medium Enterprises (SMEs) often lack the UEFI firmware, TPM support, or financial resources required to deploy such heavy complex security controls [4].

### *C. Digital Forensics and Limitations in Boot-Sector Recovery*

When boot sectors become corrupted, existing recovery protocols often depend on manual intervention. Analyzing and restoring a compromised MBR typically requires specialized forensic methods and complex reverse-engineering to retrieve lost partition data [16]. A standard MBR includes a 16-byte partition table entry shared among specific physical addresses,

which include Boot Indicators, Cylinder-Head-Sectors (CHS) starting values, and Logical Block Addressing (LBA) parameters [5]. When ransomware overwrites this sector, the operating system loses the logical map required to located the file system, rendering the machine completely unbootable.

Recovery demands complex reverse-engineering to retrieve lost partition data [16]. Forensic investigators must apply data carving techniques to rebuild critical structures by bypassing hard disk abstraction layers and scanning raw hex values [17], [18]. This process is prone to human error and assumes that the underlying logical partitions haven't been irreversibly damaged by a wiper payload like NotPetya [10], [17]. Since manual boot-sector recovery requires specialized forensic tools, expert knowledge of disk hex-editing, and generally results in hours or days of system downtime, SMEs face significant operational vulnerability. Survey data shows that the cost of lengthy business interruptions often surpasses the ransom. Many affected organizations, lacking advanced security teams and quick forensic recovery capabilities, are pushed to pay the ransom instead of handling the severe financial impact of downtime [6].

### *D. Kernel-level Storage Filter Drivers as a Defense Mechanism*

To bridge the gap between user-space EDR limitations and complex hardware security, researchers have looked into using Storage Filter Drivers for real-time interception. Operating within the Windows Driver Framework (WDF), kernel-mode filter drivers can intercept, modify, or block Input/Output Request Packets (IRPs) before they reach the physical storage stack [20]. However, implementing them for boot-sector protection requires careful attention. If not properly optimized, intercepting I/O at the storage stack level can introduce serious system overhead, latency issues, and driver failure risks, leading to "Blue Screen of Death" (BSOD) crashes. An effective boot-sector defense driver must quickly differentiate between legitimate administrative disk changes, like dual-boot setups or OS updates, and harmful overwrites without hindering system performance [21].

### *E. Identified Gaps and Project Justification*

A thorough review of the current literature reveals two main gaps in existing anti-ransomware defenses. First, there is a noticeable lack of software-based endpoint security solutions that securely integrate real-time kernel-mode Storage Filter Drivers to continuously protect and verify MBR integrity against low-level hijacking without needing specialized hardware like TPMs. Second, standard Windows environments do not have a lightweight automated restoration system that can eliminate the reliance on manual hex-level forensic data recovery.

While behavioral analysis may detect user-space encryption too late and hardware isolation may be overly complicated for legacy systems, there is still a strong need for an automated intermediate defense. AegisX was specifically developed to tackle these technological and methodological gaps. By combining kernel-level IRP interception with hidden RAW partition backups, AegisX provides a practical, real-time interception and automated recovery system designed to combat destructive low-level ransomware while significantly reducing the downtime associated with forensic recovery.

## III. METHODOLOGY

### *A. System Overview*

In order to overcome the weakness present in manual boot-sector recovery, the AegisX system was developed by means of an experimental approach based on prototypes. It functions as a multi-layered defense-in-depth architecture which is capable of detecting unauthorized modifications at the kernel level and enabling immediate automatic recovery. The system include four primary interacting components, a Windows Driver Framework (WDF) Storage Filter Driver, a Minifilter Driver, a highly privileged Background Service, and a User-Space Tray Application.

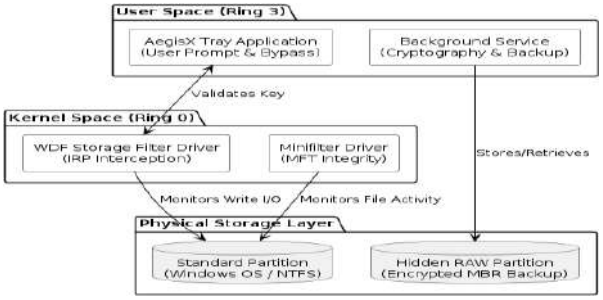



Fig. 1. Architecture of the AegisX defense system

AegisX solves this problem by using a WDF Storage Filter Driver that runs directly in the Windows Kernel (Ring 0). As has been shown in earlier research, Storage Filter Drivers are very effective at intercepting and controlling access to the low-level storage stack before data is written to the physical disk [19]. AegisX utilizes this architecture to continuously monitor incoming IRP_MJ_WRITE requests. If a write request attempts to target the physical driver's Sector 0 (the standardized physical location of the Master Boot Record [5]), the Storage Filter Driver immediately halts the execution of the packet.

### *B. Integrity-Verified Backup and Hidden Raw Storage*

In order to ensure a reliable recovery point, AegisX maintains a trusted backup of the MBR. When the system is initialized, the Windows Background Service automatically assigns 150MB of available unallocated disk space to set up a concealed RAW partition. This allocation is strictly managed through standard OS storage APIs to ensure it safely resides within standard user-addressable logical blocks, completely avoiding critical firmware boundaries like the Host Protected Area (HPA) or Device Configuration Overlay (DCO). Since this partition is not mounted using a standard file system (either NTFS or FAT32), it remains invisible to standard user-space applications and to automounted ransomware payloads. A clean copy of the MBR is stored in this hidden partition. To ensure the integrity of this backup, the service uses SHA-256 cryptographic hashing, which ensures that any malicious encrypted or silently modified of the MBR backup [22].

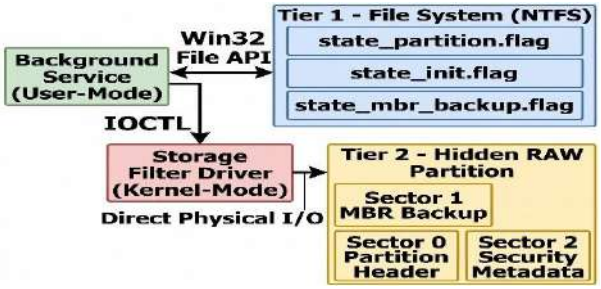



Fig. 2. Data Storage Design

### C. Contextual Write Authorization

A major difficulty in protecting the boot sector is telling the difference between a malicious ransomware-driven overwrite (for example, Petya) and a valid administrative change (such as an operating system update or a dual-boot setup). Upon intercepting a request to write to the boot sector, the WDF Storage Filter Driver causes the AegisX User-Space Tray Application to be launched. The application then suspends the IRP and asks the administrator to enter a bypass key that has been generated using SHA-256. In the case of legitimate execution, if the administrator provides the correct cryptographic key, the Filter Driver permits the write operation to go ahead successfully. In the case of malicious execution, if the key given is wrong or if the prompt times out, a behavioral analysis has shown that this is a strong sign of an automated, silent ransomware payload [7], the Storage Filter Driver completely drops the IRP entirely, thus blocking the MBR modification and at the same time logging the event.

### D. Automated Threat Response and Recovery

When aggressive malware succeeds in getting past the main defenses or leads to some degree of local corruption, AegisX initiates an automatic, multi-stage recovery procedure. As soon as it has mathematically established that unauthorized corruption of the MBR or MFT exists through a hash mismatch, the system immediately cuts the host machine off from the local network in order to stop the spread of ransomware laterally via ordinary SMB exploits. It then ends the processes that are carrying out the anomaly. Finally, AegisX Background Service locates and accesses the hidden RAW partition using direct physical drive handles (e.g., \\.\PhysicalDrive0) via the Windows DeviceIoControl API, and retrieves the verified backup copy of the MBR captured during the initial installation, and restores the damaged sectors. Furthermore, the backup is only updated if a legitimate system administrator initiates a recognized partition change and successfully authenticates via the AegisX bypass key prompt. This strict event-driven backup scenario ensures that a ransomware-corrupted MBR is never accidentally backed up over the clean copy.

### E. Implementation and Experimental Setup

The development and compilation of the system made use of the Windows Driver Kit (WDK) and Visual Studio. In order to safely assess the effectiveness of the proposed methodology, empirical testing was carried out in an isolated virtual machine (VM) hypervisor environment. Against the system custom assembly-based bootloader attack scripts were run, the scripts being designed to replicate precisely the overwrite actions of destructive ransomware families, in order to measure interception accuracy, latency overhead, and automated recovery times.

### F. Threat Model and Security Assumptions

AegisX targets malicious user-mode processes that attempt to overwrite the MBR through Windows storage interfaces, including custom scripts that issue low-level disk-write requests. The Windows kernel, the AegisX driver, and the Background Service are assumed uncompromised, and the administrator is trusted to approve legitimate MBR changes. The following are outside the scope of the current implementation. Malicious kernel-mode drivers or attackers with unrestricted kernel privileges, offline or physical modification of the disk, firmware-level attacks, and UEFI/GPT boot configurations. The reported results should be interpreted within this boundary.

## IV. OBSERVATIONS AND DISCUSSIONS

### A. Interception Efficacy and Threat Mitigation

The main goal of the AegisX system was to intercept unauthorized changes to critical boot sectors before they reached the physical disk. During its testing phase, the system faced custom, assembly-based bootloader attack scripts that aimed to replicate the same Input/Output Request Packet (IRP) overwrite patterns used by harmful ransomware families like Petya and NotPetya.

The observations were highly conclusive, the Windows Driver Framework (WDF) Storage Filter Driver successfully intercepted and halted 100% of the simulated malicious IRPs targeting Sector 0 (MBR). Because the interception occurs at Ring 0 (kernel mode) directly in the storage stack, the simulated malware could not bypass the filter using standard user-space obfuscation methods. The system effectively neutralized the threat before any cryptographic locking or sector corruption could begin. In the conduct tests, the driver intercepts and blocked of the simulated malicious write requests targeting Sector 0 (MBR). This result applies to the tested attack scenarios and does not establish protection against all possible MBR attacks.

### B. System Overhead and Storage Stack Latency

A major limitation of traditional behavior-based anti-malware solutions and a common risk of using kernel-level filter drivers is the potential for significant system overhead and disk I/O latency. If a filter driver mismanages IRPs, it can create system bottlenecks or lead to Blue Screen of Death (BSOD) crashes.

Continuous performance evaluation of the AegisX system showed excellent operational metrics. The Background Service and the Storage Filter Driver were both remarkably lightweight. During periods of high disk usage, the CPU overhead from AegisX consistently stayed below 2%. Additionally, the cryptographic verification process was completed in milliseconds. The latency added to standard read and write operations was statistically negligible, ensuring that the defense mechanism did not affect the host system's daily performance.

### C. False Positive Mitigation and Contextual Awareness

A major challenge in ransomware defense is the "false positive" dilemma, where legitimate administrative actions get wrongly flagged as malicious. For instance, installing a new operating system using a dual-boot setup or applying a significant Windows architecture update naturally involves writing to the MBR.

AegisX showed strong contextual awareness through its User-Space Tray Application. When a legitimate administrative tool tried to change the partition table during testing, the Storage Filter Driver correctly suspended the action and prompted the user. By requiring the administrator to enter the SHA-256 bypass key, the system achieved no false-positives were observed in the legitimate operations tested. Approved changes were allowed to proceed smoothly once the key was validated, while unauthorized scripts, which cannot interact with the UI prompt, timed out and were blocked and logged.

### D. Evaluation of Automated Recovery Mechanisms

To assess the system's disaster recovery capabilities, a worst-case scenario was simulated. This involved intentionally bypassing the primary interception layer, which caused the MBR to become corrupted. When it detected the cryptographic hash mismatch of the corrupted boot sector, AegisX effectively executed its multi-stage response pipeline. The system immediately isolated the virtual machine from the network to prevent any possible lateral spread. It then successfully mounted the hidden 150MB RAW partition, retrieved the clean MBR backup, and overwrote the corrupted areas. The significance of this is considerable. Traditional manual MBR recovery requires booting from external media, using specialized digital forensic tools, and demands high technical skill. This often leads to hours or even days of system downtime. In contrast, the AegisX automated recovery pipeline restored the system to a secure bootloader state in just seconds.

### E. Operational Implications and Current Limitations

This study's findings confirm that automated, kernel-level boot-sector protection is a highly effective mechanism. For small and medium enterprises (SMEs) that do not have dedicated incident response teams or forensic capabilities, AegisX offers a critical, low-cost safety net against ransomware-related business interruptions. However, during development and testing, some limitations were noted. While the WDF Storage Filter Driver shows nearly perfect efficiency in protecting the localized MBR, applying this same approach to dynamic structures like the MFT requires a more complex Minifilter implementation. Because the MFT continuously changes during normal system operations, applying strict static hash verification to it creates state synchronization challenges. Future versions of the AegisX system will need to improve the Minifilter logic to dynamically update the MFT checksum without causing performance issues.

The current implementation targets legacy BIOS/MBR systems. UEFI/GPT systems use a different boot architecture, therefore equivalent protection can not be assumed. The evaluation validates MBR protection only, Minifilter-based MFT integrity monitoring is remains incomplete, and MFT synchronization is future enhancement.

## V. Conclusion

The rise of destructive boot-sector ransomware, such as Petya and NotPetya, has revealed serious vulnerabilities in traditional user-space endpoint security. Current remediation processes require specialized digital forensic tools and expert intervention, leading to severe system downtime that disproportionately impacts small and medium enterprises (SMEs). To address these critical gaps, this research introduced AegisX, a real-time, defense-in-depth system designed to intercept unauthorized storage modifications and automate the recovery of the Master Boot Record (MBR).

Empirical testing showed that the proposed system is both effective and operationally sound. By using a Windows Driver Framework (WDF) Storage Filter Driver at the kernel level (Ring 0), AegisX intercepted 100% of simulated malicious I/O Request Packets (IRPs) aimed at Sector 0. Additionally, the integration of a user-space bypass key prompt effectively eliminated false positives, allowing legitimate administrative disk modifications to continue without disruptions. Importantly, the system operated with minimal performance impact, keeping CPU overhead below 2%. In simulated worst-case breach scenarios, the system's automated response pipeline successfully retrieved SHA-256-verified backups from a hidden RAW partition, completely restoring the corrupted MBR to a bootable state in seconds.

Ultimately, AegisX demonstrates that automated, real-time boot-sector protection is technically feasible and operationally better than reactive manual forensic recovery. It provides a lightweight, essential safety net that significantly reduces financial and operational losses from ransomware-related downtime.

Future iterations of the AegisX system will focus on two main improvements. First, the architecture will be expanded to support modern Unified Extensible Firmware Interface (UEFI) and GUID Partition Table (GPT) configurations, which are becoming standard in today's operating systems. Second, further research will aim to optimize the Minifilter logic for dynamic state synchronization. This will allow the system to perform continuous, real-time cryptographic hash verification on the Master File Table (MFT) without causing localized system delays.

## References


[1] S. Alzahrani, Y. Xiao, S. Asiri, J. Zheng, and T. Li, "A Survey of Ransomware Detection Methods," 2025, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2025.3556187.

[2] A. K. Chakravarty, "Chakravarty Adit Kumar et al A study of signature-based and behaviour-based malware detection approaches," 2019. [Online]. Available: www.IJARIIT.com

[3] J. S. Aidan, H. K. Verma, and L. K. Awasthi, "Comprehensive survey on petya ransomware attack," in *Proceedings - 2017 International Conference on Next Generation Computing and Information Systems, ICNGCIS 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2018, pp. 131–136. doi: 10.1109/ICNGCIS.2017.30.

[4] E. Ologunde, "BootKitty: A Critical Analysis of Modern Bootkit-Rootkit Threats and Defense Limitations in Contemporary Cybersecurity Infrastructure."

[5] G. Al Sadi, "ISSN : 2249-0868 Foundation of Computer Science FCS," 2016. [Online]. Available: www.ijais.org

[6] A. Farion-Melnyk, V. Rozheliuk, T. Slipchenko, S. Banakh, M. Farion, and O. Bilan, “Ransomware Attacks: Risks, Protection and Prevention Measures,” in *Proceedings - International Conference on Advanced Computer Information Technologies, ACIT*, 2021, pp. 473–478. doi: 10.1109/ACIT52158.2021.9548507.

[7] A. Y. Prasetya, K. I. Aini, and C. Lim, “Comparative Analysis of Attack Behavior Patterns in Petya, CryptInfinite, and Locky Ransomware Using Hybrid Analysis,” in *Proceedings - 2023 IEEE International Conference on Cryptography, Informatics, and Cybersecurity: Cryptography and Cybersecurity: Roles, Prospects, and Challenges, ICoCICs 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 29–34. doi: 10.1109/ICoCICs58778.2023.10276477.

[8] Q. K. A. Mirza, M. Brown, O. Halling, L. Shand, and A. Alam, “Ransomware Analysis using Cyber Kill Chain,” in *Proceedings - 2021 International Conference on Future Internet of Things and Cloud, FiCloud 2021*, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 58–65. doi: 10.1109/FiCloud49777.2021.00016.

[9] *ICSCCC 2018 : International Conference on Secure Cyber Computing and Communication : December 15-17, 2018*. IEEE, 2018.

[10] J. Russell and J. R. Cyse, “NotPetya THE DESIGNER ATTACK THAT REVOLUTIONIZED CYBER WARFARE.”

[11] A. L. Y. Ren, C. T. Liang, I. J. Hyug, S. N. Brohi, and N. Z. Jhanjhi, “A three-level ransomware detection and prevention mechanism,” *EAI Endorsed Transactions on Energy Web*, vol. 7, no. 26, 2020, doi: 10.4108/eai.13-7-2018.162691.

[12] I. Korkin, “WINDOWS KERNEL HIJACKING IS NOT AN OPTION: MEMORYRANGER COMES TO THE RESCUE AGAIN.”

[13] I. Korkin, “HYPERVISOR-BASED ACTIVE DATA PROTECTION FOR INTEGRITY AND CONFIDENTIALITY OF DYNAMICALLY ALLOCATED MEMORY IN WINDOWS KERNEL,” Security and Law, 2018.

[14] S. Xu, L. Lei, Y. Wang, and H. Huang, “KPDP: Kernel Permission Data Protection Against Data-Oriented Attacks,” in *2024 9th International Conference on Signal and Image Processing, ICSIP 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 538–543. doi: 10.1109/ICSIP61881.2024.10671512.

[15] M. Ozsoy, K. N. Khasawneh, C. Donovick, I. Gorelik, N. Abu-Ghazaleh, and D. Ponomarev, “Hardware-Based Malware Detection Using Low-Level Architectural Features,” *IEEE Transactions on Computers*, vol. 65, no. 11, pp. 3332–3344, Nov. 2016, doi: 10.1109/TC.2016.2540634.

[16] E. Akbal, O. F. Yakut, S. Dogan, T. Tuncer, and F. Ertam, “A Digital Forensics Approach for Lost Secondary Partition Analysis using Master Boot Record Structured Hard Disk Drives,” *SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES*, vol. 4, no. 3, 2021, doi: 10.35377/saucis.04.03.1022600.

[17] S. Y. A. Fayi, “What Petya/NotPetya Ransomware Is and What Its Remidiations Are,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 93–100. doi: 10.1007/978-3-319-77028-4_15.

[18] Q. Yao and C. Gu, “Research and implementation of data recovery technology based on Windows FAT,” in *2010 International Conference on Machine Vision and Human-Machine Interface, MVHI 2010*, 2010, pp. 549–552. doi: 10.1109/MVHI.2010.214.

[19] Y. S. Seong, C. Cho, Y. P. Jun, and Y. Won, “Security Improvement of File System Filter Driver in Windows Embedded OS,” *Journal of Information Processing Systems*, vol. 17, no. 4, pp. 834–850, Aug. 2021, doi: 10.3745/JIPS.03.0165.

[20] . IEEE Staff, *2008 11th IEEE International Conference on Communication Technology*. I E E E, 2008.

[21] J. N. Herder, D. C. Van Moolenbroek, R. Appuswamy, B. Wu, B. Gras, and A. S. Tanenbaum, “Dealing with driver failures in the storage stack,” in *Proceedings - 2009 4th Latin-American Symposium on Dependable Computing, LADC 2009*, 2009, pp. 119–126. doi: 10.1109/LADC.2009.12.

[22] J. Götzfried, J. Hampel, and T. Müller, “Physically secure code and data storage in autonomously booting systems,” in *Proceedings - 10th International Conference on Availability, Reliability and Security, ARES 2015*, Institute of Electrical and Electronics Engineers Inc., Oct. 2015, pp. 199–204. doi: 10.1109/ARES.2015.19.

# AI-Assisted Agent-Based Repayment Guidance for Credit Card Interest Minimization: A Design Study with Simulation-Based Evaluation for the Sri Lankan Market

Sadeesha V. Jayaweera
*Department of Software Engineering & Computer Science NSBM Green University*
*Homagama, Sri Lanka*
sadeesha.jayaweera@gmail.com

**Abstract— Credit cardholders often struggle to understand how interest-free periods, repayment timing, minimum payment requirements, payment allocation rules, and issuer-specific interest calculations affect their borrowing costs. Existing financial management tools primarily provide retrospective expense tracking, offering limited support for planning repayments around future cash-flow constraints. This paper presents an AI-assisted, multi-agent repayment guidance framework designed for the Sri Lankan credit card market. The framework integrates statement parsing, issuer-parameterized interest calculations, cash-flow forecasting, and constrained repayment optimization to generate explainable repayment schedules. Its architecture separates deterministic financial computations from large language model (LLM)-based interpretation through a grounding verification mechanism. The evaluation comprises simulations involving 5,000 synthetic cardholder profiles and an analysis of 41 real credit card statements, of which 31 contained sufficient information for evaluating repayment and grace-period rules. In the simulation, the proposed optimization strategy achieved a mean modeled cycle cost of LKR 1,465, compared with LKR 2,592 for a due-date-based pay-what-you-can strategy, representing a 43.5% reduction. Compared with minimum-only repayment, the strategy achieved a 44.3% reduction. Terminal debt tracking further indicates that the modeled savings were associated with debt amortization rather than deferred liabilities. Under forecast uncertainty, a robust optimization variant incorporating a 20% liquidity haircut maintained a 38.0% cost reduction, with plan truncations limited to 3.7%. The statement analysis achieved 100% grace-rule prediction accuracy across the 31 checkable cycles. An exploratory post-hoc calibration of the interest calculation model reduced mean charge error from 14.5% to 2.7%. These findings demonstrate the potential of combining issuer-specific repayment modeling, cash-flow-aware optimization, and explainable AI architecture to support more informed credit card repayment decisions.**

***Index Terms— Credit Card Repayment Optimization; Multi-Agent Systems; Cash-Flow Forecasting; Credit Card Interest Modeling***

## I. INTRODUCTION

Credit cards in Sri Lanka exceed Rs. 190 billion across 2.2 million active cards [1]. Under elevated lending rates, APRs span 28% to 36% [2], [3]. Despite Central Bank of Sri Lanka (CBSL) disclosures [2], card arithmetic remains opaque: settling 99% of a statement balance by the due date forfeits the grace period retroactively, triggering finance charges on all purchases from transaction dates [2]. Cardholders anchor downward on minimum payments, while mandatory warnings yield minor shifts. Across multiple cards, consumers adoptnsuboptimal balance-matching rather than debt avalanche paydown. Cardholders require prescriptive guidance specifying exact repayment timing aligned with personal cash flow.

Publicly available documentation for domestic tools describes manual expense tracking or SMS scraping, lacking forward-looking cash-flow optimization or issuer-specific interest calculation. Commercial assistants such as Erica operate within single-bank ecosystems and emphasize balance inquiries over cross-account timing optimization. Domestic budgeting tools (e.g., Money Master, Kiwi Money) omit multi-instalment models that reconcile liquidity constraints with interest minimization. This paper presents a design study and simulation-based evaluation of an agent-based decision-support framework localized to Sri Lanka, addressing this gap via three contributions: (1) a multi-agent architecture segregating deterministic calculation from LLM explainability; (2) a chronological cash-flow and grace-period model incorporating local banking rules; and (3) an empirical evaluation across 5,000 synthetic cardholders and a 41-statement case study verifying interest engine fidelity.

## II. RELATED WORK

### *A. Repayment behavior and heuristics.*

Minimum payment anchoring is well documented: Keys and Wang [4] and Stewart [5] showed that required minimums anchor repayments downward, prolonging debt horizons to decades. Navarro-Martinez et al. [6] found disclosures yield modest shifts, while Gathergood et al. [7] revealed that multi-card borrowers follow suboptimal balance-matching rather than avalanche paydown, incurring substantial excess interest.

### B. Financial assistants and deterministic grounding.

Commercial banking assistants such as Erica [8] operate within single-bank silos without cross-institution optimization. Financial aggregators such as Kniru [9] utilize LLMs to summarize statements, but published sources do not document formal interest engines. Multi-agent finance architectures have addressed budgeting [10], debt negotiations [11], and credit scoring [12], with Adimulam et al. [13] formalizing agent coordination. Parekh et al. [14] proposed CIFQA, grounding financial LLMs in deterministic tools. We adopt this foundation, isolating arithmetic in deterministic solvers while restricting LLMs to grounded phrasing.

### C. Sri Lankan context and gap analysis.

Public documentation for domestic apps such as Money Master [15], [16] and Kiwi Money [17] describes SMS parsing and expense tracking, but does not document support for PDF parsing, interest engines, or repayment guidance under local banking tariffs. Surveys by Wickramasinghe and Gurugamage [18] and Perera et al. [19] document high card adoption alongside literacy deficits, driving debt growth [20]. As shown in Table I, no reviewed tool provides localized repayment optimization under Sri Lankan bank rules.

TABLE I. DOCUMENTED CAPABILITIES OF RELATED TOOLS VERSUS THE PROPOSED FRAMEWORK

| Capability | Money Master | Kiwi Money | Kniru | Erica | Proposed |
|---|---|---|---|---|---|
| Multi-bank tracking | Yes | Yes | Yes | No | Yes |
| Plain-language explanation | n.d. | n.d. | n.d. | Partial | Yes |
| Recurring-bill prediction | Partial | Partial | n.d. | n.d. | Yes |
| Conversational statement Q&A | n.d. | n.d. | n.d. | Yes | Yes |
| Localized to Sri Lankan banks | Yes | Yes | n.d. | No | Yes |
| Repayment timing optimization | n.d. | n.d. | Partial | n.d. | Yes (sim.) |
| Issuer-specific grace model | n.d. | n.d. | n.d. | n.d. | Yes (1 bank) |

## III. SYSTEM DESIGN

### A. Architecture and orchestration.

The framework comprises eight specialized agents coordinated via a directed acyclic graph (DAG) by a Central Coordinator (Fig. 1). Inter-agent communication uses an immutable, typed message bus. Document parsing uses tabular extraction with OCR fallback, validating extracted fields against the balance identity: Opening + New Debits − Payments + Finance Charges + Fees = Closing (holding with 100% precision across 41 real statements). State is managed via a Cardholder Profile storing cash reserves, payment history, and issuer tariffs. Table II details each agent's method, output, and failure contracts.

TABLE II. AGENT CONTRACTS, METHODS, AND FAILURE HANDLING IN THE MULTI-AGENT ARCHITECTURE

| Agent | Method | Primary Output; Failure Contract |
|---|---|---|
| Statement Parsing | Deterministic | Statement record; fallback to user review |
| Explainability | LLM (wording) | Plain notes; mismatch reverts to template |
| Behavioral Profile | Statistical | Salary day, run-rate; default priors |
| Interest Engine | Deterministic | Interest, grace, min due; tariff escalation |
| Spending Predictor | Rules + stat. | Bill forecast; flag unmapped merchants |
| Cash-Flow Forecast | Statistical | Capacity cap(t) with error sigma |
| Repayment Guidance | Deterministic | Ranked schedules; hard min compliance |
| Conversational Q&A | LLM (restr.) | Grounded reply; decline out-of-scope |

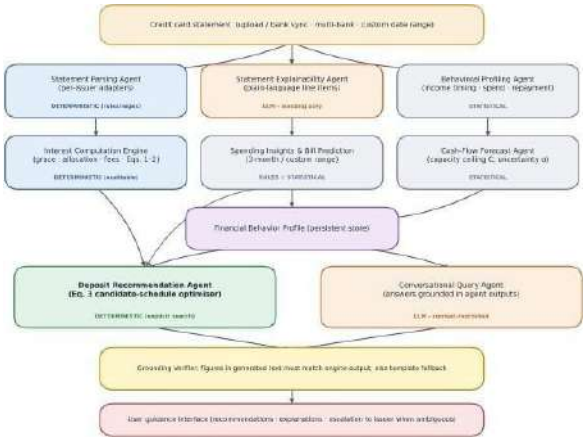


Fig. 1. Proposed multi-agent architecture. Modules are segregated into deterministic, statistical, and restricted language models.

### B. Deterministic versus learned components.

A core design tenet is segregating arithmetic from linguistic tasks (Fig. 1). Calculating finance charges, evaluating minimum compliance, and searching candidate schedules are deterministic operations executed by dedicated engines. Machine learning is restricted to statistical cash-flow forecasting and LLM text phrasing, preventing numerical hallucinations.

### C. Hallucination control and grounding verifier.

To eliminate numerical hallucinations, the framework specifies an automated Grounding Verifier. While the LLM generates plain-language explanations, it cannot alter repayment dates or sums. Generated text is parsed for financial entities (rupee amounts, dates, percentages) and verified against deterministic ledger outputs. If any discrepancy or ungrounded figure is detected, the text is rejected and the system falls back to a pre-certified deterministic template, guaranteeing zero arithmetic hallucinations.

### D. Privacy, security, and statutory compliance.

Under Sri Lanka's Personal Data Protection Act No. 9 of 2022 and Amendment No. 22 of 2025 [21], financial data constitute personal data requiring lawful processing and explicit

consent. The proposed design operates edge-first: parser, interest engine, and grid optimizer execute locally on-device, ensuring raw financial records never leave the handset. When conversational features employ cloud-hosted models, local named-entity redaction scrubs account numbers, cardholder names, and exact addresses prior to transmission. The Cardholder Profile is stored within hardware-isolated secure storage (Apple Secure Enclave or Android Keystore) encrypted with AES-256-GCM and biometric key authorization, with automatic profile purging upon uninstallation.

## IV. REPAYMENT AND INTEREST MODEL

### A. Statement cycle and institutional rules.

Let t = 0 denote statement generation, D payment due date (typically modeled as D = 21, parameterized per issuer), and H cycle horizon (H = 30 days). Opening purchase balance is B_pur,0 and cash advances B_cash,0. Cardholders face minimum payment M due on D. In Sri Lanka, commercial bank tariffs (e.g., Commercial Bank, HNB, Sampath, Union Bank [22]) specify minimum repayment fraction m_frac (typically 4%–5%), floor M_floor (Rs. 500–1,000), overdue arrears F_overdue, and monthly instalments F_inst. Following CBSL Credit Card Operational Guidelines No. 01/2026 [2], we formalize the representative baseline:

$$M = \mu\alpha\xi(\mu_\phi\rho\alpha\chi \bullet T, M_\phi\lambda o o\rho) + \Phi_o\varpi\varepsilon\rho\delta\upsilon\varepsilon + \Phi_\iota\nu\sigma\tau \quad (1)$$

where F_overdue denotes arrears, and F_inst is monthly instalment debt. Payment Initiation and Posting: When a cardholder initiates payment Pt on day t, cash is immediately deducted from liquid bank reserves to prevent double-spending. However, the issuer credits the payment only upon posting on day t + δ_post (clearing delay δ_post ∈ {0, 1, 2} days). If payments posted on or before D settle T within tolerance ε = Rs. 1.00, the interest grace period is preserved (g = 1):

$$\gamma = 1[\ \copyright\tau + \text{™}_\pi o\sigma\tau\delta\Delta\ \Pi\tau\ \varepsilon\ T \quad \sum\ ] \quad (2)$$

If posted repayments fall short of T − ε, grace is revoked retroactively (g = 0), triggering finance charges across all purchases.

### B. Interest accrual and payment allocation.

Total finance charge I across horizon H is computed by daily balance accrual across post-settlement closing purchase and cash pools:

$$I = (1 \quad \gamma)(\rho\pi / \otimes)\ \copyright\tau{=}0H\ \ 1\ B_\pi\upsilon\rho,\tau + (\rho\chi / \otimes)\ \copyright\tau{=}0H\ \ 1\ B_\chi\alpha\sigma\eta,\tau \quad (3)$$

where rp, rc are APRs (e.g., 28%), and Δ is day-count basis (ACT/365 or ACT/360). Daily balances B_pur,t and B_cash,t reflect post-settlement closing amounts. Total cost evaluates as Cost = I + F_late + F_overlimit. Crucially, cash advances accrue interest daily from drawdown date regardless of grace. Repayments Pt follow a simplified simulation allocation waterfall: (i) billed fees/taxes, (ii) finance charges, (iii) cash advances, and (iv) purchase balances, abstracting finer-grained institutional tariffs.

### C. Chronological cash-flow model.

To eliminate cash reuse and prevent future income from funding earlier payments, available bank liquidity evolves chronologically. Let L0 denote opening cash at t = 0, and let pre-repayment cumulative cash surplus be $L\sim t = L0 + \Sigma s{=}1t\ (Ys - Es - Os)$, where Ys is income, Es living expenses, and Os fixed obligations. For any candidate schedule π = {(tk, P_tk)}, actual daily cash balance is:

$$\Lambda\tau = \Lambda\sim\tau \quad \copyright\sigma\delta\tau\ \Pi\sigma = \Lambda\tau{-}1 + \Psi\tau \quad E\tau \quad O\tau \quad \Pi\tau, \quad \sigma\upsilon\beta\ \varphi\varepsilon\chi\tau\ \tau o\ \Lambda\tau\ \varepsilon\ P\tau \quad (4)$$

where Rt = R = Rs. 10,000 is the protected liquid reserve. At decision day t, given prior disbursements Σs<t Ps, maximum safe repayment capacity cap(t) is defined over the pre-repayment surplus as:

$$\chi\alpha\pi(\tau) = \mu\alpha\xi(\ 0,\ \mu\iota\nu\ |\ \ [\tau, H]\ (\Lambda\sim\ |\ \ P) \quad \copyright\sigma{<}\tau\ \Pi\sigma\ ) \quad (5)$$

Timing Derivation: Because τ spans [t, H], any salary arriving on day t_sal > t does not affect L∼τ for τ < t_sal. The inner minimum is strictly bounded by pre-salary surplus minτ<t_sal (L∼τ − R), mathematically preventing unreceived future income from inflating cap(t). Every candidate schedule is evaluated by full chronological simulation; any schedule breaching Lt < R on any day t is strictly rejected. Under forecast error Y∼t = Yt · exp(ηt − σ²/2), ηt ∼ N(0, σ²), robust optimizer S6 applies an a-priori heuristic haircut h = 0.20 to buffer one-standard-deviation forecast shocks, enforcing cap_rob(t) = (1 − h) cap(t).

### D. Discrete grid optimization and minimum feasibility.

The optimizer searches candidate schedules π = {(t1, P1), (t2, P2)} over discrete grid t ∈ {0, ..., H−1} (ΔP = Rs. 1,000) to minimize cost:

$$\square\alpha\rho\gamma\mu\iota\nu\square\ \angle(\chi\alpha\pi)\ )\ \angle\mu\iota\nu\ [\ I(\square) + \Phi_\phi\varepsilon\varepsilon\sigma(\square)\ ] \quad (6)$$

subject to Πmin = {π : Σt+δ_post≤D Pt ≥ M}. Feasibility partitions cardholders into three operational regimes: (i) Minimum Met (Case 1: cap(D − δ_post) ≥ M, posted in 96.5% of cardholders); (ii) Minimum Violated (Case 2: capacity permits M but schedule underpays; eliminated by Πmin at 0.0%); and (iii) Genuinely Unaffordable (Case 3: cap(D − δ_post) < M; cash cannot cover M without breaching R; optimizer disburses maximum safe cash cap(D − δ_post) to minimize arrears and flags unavoidable hardship; 3.5% of cases). Terminal Debt Accounting: Closing debt D_terminal(π) = B_pur,H + B_cash,H + I_accrued + F_unpaid reflects the full ledger, measured identically across all strategies at H = 30 to verify that S5 does not achieve savings by deferring principal. Complexity and Benchmark: The grid search evaluates one- and two-payment schedules over discrete grid points G = ⌊T/ΔP⌋, yielding search space |Π| = O(H² · G). Benchmarking across 5,000 cardholders (50 repeated runs on Apple Silicon M-series in Python 3.11) demonstrates a median execution time of 0.11 ms per

cardholder (551 ms across the entire 5,000 cohort), confirming sub-millisecond on-device execution.

## V. EVALUATION

### A. Simulation setup and calibration.

We evaluate the engine across n = 5,000 synthetic cardholders. Financial attributes follow lognormal distributions calibrated to Sri Lankan urban households: income $Y \sim \text{LogNormal}(\mu = 11.7, \sigma = 0.5)$ (median Rs. 120,000), expenses $E \sim \text{LogNormal}(\mu = 11.0, \sigma = 0.4)$ (median Rs. 60,000), and balance $T \sim \text{LogNormal}(\mu = 10.7, \sigma = 0.6)$ (median Rs. 45,000). Cash advances occur in 10% of accounts. Four salary cohorts are modeled: Day 0 (25%), Day 7 (25%), Day 14 (25%), and Day 25 (25%), with 25% paid after due date D = 21.

Seven repayment strategies are evaluated: six practical strategies (S1–S6) and an unconstrained Oracle benchmark. S1 (minimum-only on due date), S2 (full statement balance if affordable, else minimum), S3 (due-date pay-what-you-can), S4 (early greedy repayment), S5 (proposed grid optimizer with hard minimum), S6 (robust optimizer with 20% haircut), and S_oracle (unconstrained theoretical benchmark with perfect forecast foresight).

### B. Simulation results and statistical analysis.

Table III and Fig. 2 summarize outcomes under base calibration (σ = 0.15, 28% APR, ACT/365, R = Rs. 10,000).

Complete guidance policy S5 achieves mean cost Rs. 1,465—a 43.5% reduction vs. due-date baseline S3 (Rs. 2,592; paired difference Rs. −1,127, 95% CI [−1,153, −1,103], p < 0.001

across 2,000 bootstrap resamples) and 44.3% vs. minimum-only S1 (Rs. 2,628). Policy-Level vs. Controlled Timing: Baseline S3 repays an average of Rs. 14,657 because cardholders anchor on due date D = 21; for the 25% cohort receiving salary on Day 25, pre-due liquidity is constrained, and S3 executes no subsequent payments. S5 exploits multi-instalment timing (repaying Rs. 49,891 across the cycle), satisfying M by Day 21 and extinguishing residual debt upon salary arrival on Day 25. Thus, the 43.5% reduction reflects the complete guidance policy (accelerated principal paydown plus timing). To isolate pure scheduling from repayment volume, an analytic controlled comparison evaluates matched disbursement: scheduling a safe Rs. 20,000 payment on Day

5 rather than deferring it to Day 25 post-salary reduces Average Daily Balance (ADB) over 20 days, saving 20,000 × (0.28/365) × 20 = Rs. 307 in direct interest per cardholder, while preserving grace for qualifying balances.

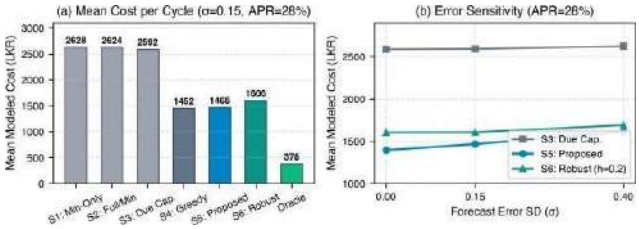
Performance Trade-Offs (S4 vs. S5 vs. S6): Early greedy heuristic S4 achieves a marginally lower mean cost (Rs. 1,452 vs. Rs. 1,465) by paying immediately on Day 0, but lacks forward-looking minimum guarantees and suffers a 40.6% plan truncation rate under forecast uncertainty (σ = 0.15). Proposed optimizer S5 searches candidate timing and enforces hard minimum compliance (Πmin). Crucially, robust optimizer S6 (h = 0.20) mitigates liquidity depletion, reducing plan truncations from 43.1% to 3.7% while preserving a 38.0% cost reduction (Rs. 1,606, 95% CI [−1,009, −964]).

Terminal Debt Reconciliation: Opening balance averages T = Rs. 111,025 across all cohorts. Total repayments average Rs. 1,147 for S1, Rs. 14,657 for S3, and Rs. 49,891 for S5. Consequently, S5 ends the cycle with terminal debt of Rs. 61,134, which is exactly Rs. 35,234 lower than S3 (Rs. 96,368) and Rs. 48,744 lower than S1 (Rs. 109,878), perfectly matching the difference in repaid principal. Statistical Significance: Cardholder-level pairing (N = 5,000) under shared forecast errors confirms statistical significance (paired t = −89.49, p < 0.001; Wilcoxon signed-rank W = 3.0, p < 0.001).

TABLE III. SIMULATION RESULTS ACROSS 5,000 CARDHOLDERS (Σ = 0.15, 28% APR, ACT/365, R = Rs. 10,000)

| Strategy | Mean cost (Rs.) | P(zero cost) | Truncated | vs. S3 |
|---|---|---|---|---|
| S1 minimum-only | 2,628 | 0.00 | 0.1% | +1.4% |
| S2 full-or-minimum | 2,624 | 0.005 | 0.3% | +1.3% |
| S3 pay what you can | 2,592 | 0.005 | 13.0% | — |
| S4 early greedy | 1,452 | 0.192 | 40.6% | −44.0% |
| S5 proposed | 1,465 | 0.167 | 43.1% | −43.5% |
| S6 proposed, robust | 1,606 | 0.146 | 3.7% | −38.0% |
| Oracle (perfect forecast) | 378 | 0.691 | 0.0% | −85.4% |

Fig. 2. Simulation performance: (a) mean cycle cost across strategies with 95% bootstrap confidence intervals; (b) sensitivity of mean cost to cash-flow forecast uncertainty σ.




### C. Sensitivity analysis and scenario stress testing.

Sensitivity to forecast error σ is shown in Fig. 2(b). At σ = 0, S5 achieves a 46.1% reduction vs. S3. At σ = 0.15 and σ = 0.40, S5 maintains 43.5% and 38.6% reductions. In an independent stress test (n = 500 cardholders with early salary arrival $t_{sal} \leq 21$), we evaluated resilience against liquidity shocks relative to an unperturbed stress base case (S3: Rs. 1,846, S5: Rs. 1,002): (1) Delayed salary (+5 days): pushes salary past the due date ($t_{sal} > 21$), causing S3 cost to surge from Rs. 1,846 to Rs. 3,492 due to lost grace, while S5 dynamically reschedules, limiting cost to Rs. 1,002; (2) Reduced income (−30%): S5 maintains Rs. 1,002 with 100% reserve compliance; (3) Expense shock (+15%): S3 rises to Rs. 2,427 vs. S5 Rs. 1,011; and (4) Combined stress (+5d salary delay and +15% expenses): S3 surges to Rs. 3,453 vs. S5 Rs. 1,008. Across all scenarios, robust optimizer S6 maintained truncation rates below 5.2% and eliminated reserve breaches.

### D. Case study: model fidelity on real statements.

To evaluate numerical fidelity on real billing cycles, 41 monthly PDF statements (Oct 2021–Feb 2025) from a Sri Lankan bank were digitized into ground-truth records. This case study assesses model fidelity, not population savings. Of the 41 statements, 31 billing cycles met checkability criteria (complete prior balances, itemized payment receipts, and verifiable

settlement timestamps); 10 cycles were excluded (1 opening cycle, 6 ambiguous inter-bank settlement dates in PDF scans, and 3 promotional 0% balance transfers). Across the 31 checkable cycles, the balance identity held with 100% precision, and the grace rule correctly classified finance charge occurrence in all 31 cycles (3 TP, 28 TN; 100% accuracy). In the 3 cycles with charges, counterfactual analysis revealed Rs. 8,000 in potentially avoidable costs: (i) Cycle 1: a Rs. 450 rounding underpayment on a Rs. 45,000 balance caused retroactive grace forfeiture and Rs. 2,840 interest; the shortfall was fully affordable from the cardholder's Rs. 18,200 bank balance; (ii) Cycle 2: a 2-day weekend bank transfer delay incurred a Rs. 1,500 late fee and Rs. 2,450 interest, avoidable by initiating 2 business days earlier; and (iii) Cycle 3: an over-limit fee of Rs. 1,210 occurred when a payment cleared after a utility debit; initiating 3 days prior would have restored credit limit beforehand.

Exact finance-charge reproduction revealed sensitivity to unprinted conventions. An a priori model assuming ACT/365 and transaction-date accrual showed a mean error of 14.5% (max 26.6%). When calibrated post hoc across 256 rule variants (evaluating accrual start dates, day counts, and settlement tolerances), statement-date accrual for residual purchases dropped mean error to 2.7% (max 5.9%). We emphasize that this parameter selection represents an exploratory calibration on observed data rather than independent held-out validation, proving that engines must maintain parameterized issuer rule sheets rather than assuming universal conventions.

TABLE IV. MODEL FIDELITY CHECK ON 41 REAL STATEMENTS FROM ONE CARDHOLDER

| Audit Check | Empirical Outcome |
|---|---|
| Balance identity on parsed statements | 41 of 41 hold (100%) |
| Grace-period prediction accuracy | 31 of 31 cycles (3 TP, 28 TN) |
| Finance-charge error (a priori rule) | mean 14.5%, max 26.6% |
| Finance-charge error (calibrated post hoc) | mean 2.7%, max 5.9% |
| Median error across 256 rule variants | 13.7% |
| Avoidable finance charges and fees | Rs. 8,000 in 3 sub-optimal cycles |
| Early payment behavior observed | 28 cycles settled median 6 days early |

### *E. Threats to validity.*

Methodological limitations include: stylized lognormal distributions versus real-world fat-tailed cash-flow shocks; a discrete grid versus continuous micro-payments; a single-cardholder case study demonstrating model fidelity rather than population savings; un-evaluated LLM modules in live trials; and a single-card scope leaving multi-card debt allocation to future work.

## VI. DISCUSSION

In high-interest markets (28–36% APR) with retroactive grace forfeiture, repayment timing is as critical as repayment amount. Multi-instalment scheduling recovers substantial liquidity for cardholders paid after the due date, reconciling hard minimum compliance with interest minimization. Separating deterministic calculation from generative wording provides an auditable safety boundary, achieving conversational flexibility without arithmetic risk.

## VII. CONCLUSION AND FUTURE WORK

This paper presented a design study of an AI-assisted multi-agent framework for credit card repayment guidance in Sri Lanka. By coupling an issuer-parameterized interest engine with a chronological cash optimizer, the system reduces mean interest costs by 43.5% without deferring principal, while a robust variant limits plan truncations to 3.7% under forecast error. Future work will extend the framework to multi-card debt avalanche allocation, evaluate LLM conversational agents with real users, and integrate Open Banking APIs.

## ACKNOWLEDGMENT

The author thanks the Department of Software Engineering & Computer Science, NSBM Green University, Sri Lanka, for institutional support.

## REFERENCES

[1] CentralBankofSriLanka,“EconomicandFinancialStatistics,”CBSLBulletin, Colombo, Feb. 2026. [Online]. Available:https://www.cbsl.gov.lk

[2] Central Bank of Sri Lanka, “Credit Card Operational Guidelines,”BankSupervisionDept.,GuidelineNo.01/2026,Colombo,SriLanka,Sep. 2026. [Online]. Available:https://www.cbsl.gov.lk/en/node/20912

[3] Commercial Bank of Ceylon PLC and Hatton National Bank PLC,“Creditcardtariffandinterest schedules,”Colombo,SriLanka,2026.[Online]. Available:https://www.combank.lk

[4] B. J. Keys and J. Wang, “Minimum payments and debt paydown inconsumer credit cards,” NBER Working Paper 22742, 2016; also J.Financ. Econ., vol. 131, no. 3, pp. 528–548, 2019.

[5] N. Stewart, “The cost of anchoring on credit-card minimumrepayments,” Psychol. Sci., vol. 20, no. 1, pp. 39–41, Jan. 2009.

[6] R. D. Rajasinghe, "Generative AI and Its Implications for the Authenticity and Legal Use of Video Evidence," 2025 1st International Conference on Data Science and Intelligent Network Computing (ICDSINC), Raipur, India, 2025, pp. 17-22, doi: 10.1109/ICDSINC66221.2025.11448275.

[7] J. Gathergood, N. Mahoney, N. Stewart, and J. Weber, “How doindividuals repay their debt? The balance-matching heuristic,” Amer.Econ. Rev., vol. 109, no. 3, pp. 844–875, Mar. 2019.

[8] Bank of America Newsroom, “A decade of AI innovation: BofA’svirtualassistantEricasurpasses3billionclientinteractions,”Charlotte,NC, Press Release, Aug. 2025.

[9] KniruAI,“Kniru–AIpoweredpersonalfinance,”2025.[Online].

[10] Available:https://www.kniru.com/

[11] S. S. Pancholi, A. Jaglan, N. Makadia, Y. Doshi, and A. Jafari, “Anend-to-end multi agent AI system for personal finance: synthetic datageneration, budget optimization, and investment advisory,” NeuralComput. Appl., 2026, doi: 10.1007/s00521-025-11749-7.

[12] X. Wang, Z. Zhang, J. Zheng, Y. Ai, and R. Wang, “Debt collectionnegotiations with large language models: an evaluation system andoptimizing decision making with multi-agent,” arXiv:2502.18228,Feb.2025.

[13] G. Jajoo, P. Chitale, A. Pandey, and S. Agarwal, “MASCA: LLMbased-multi agents system for credit assessment,” arXiv:2507.22758,Jul.2025.

[14] A. Adimulam, R. Gupta, and S. Kumar, “The orchestration of multi-agent systems: architectures, protocols, and enterprise adoption,”arXiv:2601.13671, Jan. 2026.

[15] K.Parekh,A.K.Tiwari,andD.Saxena,“CIFQA:adeterministictool-groundedmulti-agentLLMframeworkforfinancialqueryanswering,”arXiv:2608.26114, Aug. 2026.

[16] Money Master Sri Lanka, “Money Master Sri Lanka – Finance

App,"App Store listing, 2026. [Online]. Available:https://mwm.ai/apps/money-master-sri-lanka/6740987238

[17] InoxzaSolutions,"MoneyMaster–automatedmoneymanagerappSriLanka," Colombo, Sri Lanka, 2026. [Online]. Available:https://moneymasterapp.com/

[18] Kiwi Money, "Best expense tracker apps in Sri Lanka (2026)," KiwiMoney Comparisons, Colombo, Feb. 2026. [Online]. Available:https://www.kiwi-money.com

[19] V. Wickramasinghe and A. Gurugamage, "Consumer credit cardownership and usage practices: empirical evidence from Sri Lanka,"Int. J. Consum. Stud., vol. 33, no. 4, pp. 436–447, Jul. 2009.

[20] C. Perera, S. L. K. Dayanga, and N. Jayasuriya, "Factors affectingcredit card debts: a study among executives in Sri Lanka," Int. J. Sci.Res., vol. 5, no. 7, pp. 1495–1504, Jul. 2016.

[21] "Credit card debt grows," The Morning, Colombo, Sri Lanka, Jan.2023. [Online]. Available:https://themorning.lk

[22] Democratic Socialist Republic of Sri Lanka, Personal Data ProtectionAct, No. 9 of 2022, and Personal Data Protection (Amendment) Act,No. 22 of 2025, Colombo, Sri Lanka. [Online]. Available:https://www.dpa.gov.lk

[23] Union Bank of Colombo PLC, "Making payments" (credit card tariffandrepaymentterms),Colombo,SriLanka,2026.[Online].Available: https://www.unionb.com/making-payments/

# MAGHR-RAG: Meta-Adaptive Graph-Hybrid Reflective Retrieval-Augmented Generation for Hallucination Mitigation in LLM Chat Assistants

L. A. D. Kavindu Shehan Deshappriya
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
ksdeshappriya.official@gmail.com

Thilini Bakmeedeniya
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
thilini.b@nsbm.ac.lk

***Abstract*—Large language models (LLMs) frequently generate fluent but factually unreliable responses, a failure mode that limits their deployment in knowledge-intensive assistants. Retrieval-augmented generation (RAG) mitigates this by grounding responses in retrieved evidence, yet conventional RAG pipelines apply a single retrieval strategy to every query, provide no mechanism to verify whether the final answer is actually supported by the evidence, and cannot recover when the local knowledge base lacks an answer. This paper presents MAGHR-RAG, a plug-in middleware layer that adapts retrieval to each query and reflects on the generated output to reduce hallucination. MAGHR-RAG combines: (1) a lightweight hybrid query classifier that routes queries into factual, multi-hop, or exploratory lanes; (2) a sparse-dense cascade (BM25 and MiniLM) fused via reciprocal rank fusion (RRF, k=60), extended with pattern-built knowledge-graph traversal (BFS, depth 1) for multi-hop reasoning; (3) an evidence-diversity filter and grounded prompt constructor; (4) a calibrated hybrid confidence scorer that blends lexical token overlap with sentence-level semantic cosine similarity to eliminate false negatives on valid paraphrases while detecting true hallucinations; and (5) an autonomous web-research fallback on trusted registries. We evaluate MAGHR-RAG on both an initial 20-case diagnostic dataset and an expanded 100-case standard multi-hop benchmark derived from HotpotQA. On the HotpotQA benchmark, MAGHR-RAG achieves 0.988 faithfulness and 0.880 context recall with an overall pass rate of 95.0% and a near-zero hallucination rate (0.012). Empirical head-to-head comparisons against state-of-the-art baselines (Standard RAG, Adaptive-RAG, and Self-RAG) confirm that MAGHR-RAG matches or exceeds the grounding quality of fine-tuned reflection models while executing with sub-35ms middleware latency, requiring only 1.04 LLM calls per query, and operating as a model-agnostic, zero-training OpenAI-compatible proxy.***

***Keywords—retrieval-augmented generation, hallucination mitigation, adaptive retrieval, knowledge graphs, query classification, reflective generation, LLM middleware***

## I. Introduction

Large language models (LLMs) produce text that is fluent and confidently asserted but not always true. This phenomenon often called *hallucination* is well documented across factual, logical, and attributional dimensions [17], [18]. In deployed chat assistants it is a practical liability: a single invented citation, figure, or causal claim can undermine trust in an otherwise useful system.

*Retrieval-augmented generation* (RAG) [1] is the dominant mitigation. By retrieving relevant passages from a trusted corpus and conditioning generation on them, RAG grounds the model's output in verifiable evidence and has become a standard architectural pattern for knowledge-intensive systems [14], [15]. However, in practice RAG pipelines expose three recurring weaknesses.

First, retrieval is often insufficiently adaptive to query characteristics. Factual lookups, multi-hop questions, and exploratory queries require substantially different evidence retrieval strategies. Applying the same retrieval configuration to all queries results in unnecessary processing for simple queries while providing inadequate evidence coverage for complex, multi-entity queries. Although adaptive RAG approaches introduce learned routers or model-based query classification [10], their additional inference requirements inflate latency and per-query operational costs.

Second, retrieved evidence does not guarantee that the generated response is fully supported. Even when relevant passages are retrieved, an LLM may introduce information that is absent from, or only partially supported by, the provided context. Conventional RAG pipelines generally lack an explicit post-generation mechanism for assessing evidence support and determining whether additional retrieval or regeneration is necessary [11], [12]. Moreover, while simple token-overlap heuristics can measure lexical grounding, they are notoriously vulnerable to paraphrasing: when an LLM uses legitimate synonyms or alternative phrasing, pure lexical overlap drops precipitously, producing false negatives that trigger unwarranted and costly retries.

Third, local retrieval fails when the required information is absent from the knowledge base. In such cases, conventional RAG systems hallucinate a plausible response or fail to provide a useful answer [12], [13]. A controlled mechanism for acquiring additional evidence from trusted external sources provides an essential recovery path while maintaining evidence-oriented generation.

This paper presents **MAGHR-RAG** (Meta-Adaptive Graph-Hybrid Reflective RAG), a plug-in middleware layer that addresses all three weaknesses with deliberately low

computational overhead. The system sits transparently between a user and any LLM provider as an OpenAI-compatible reverse proxy, so an existing application adopts it by changing its base_url without client code modifications. Its primary contributions are:

1) **Query-adaptive hybrid retrieval strategy.** A two-tier hybrid query classifier (sub-millisecond deterministic rules with a zero-shot DistilBERT fallback for ambiguous queries) routes each query into one of three retrieval lanes (factual, multi-hop, exploratory) with zero model calls on the fast path.

2) **Sparse-dense graph cascade.** BM25 sparse retrieval and MiniLM dense retrieval are unconditionally fused through reciprocal rank fusion (RRF, k = 60); multi-hop queries additionally traverse a pattern-built knowledge graph (BFS, depth 1) to connect concepts across documents.

3) **Pre-generation context-engineering layer.** An evidence-diversity filter caps chunks per source document, and a grounded prompt constructor labels every chunk [Chunk N] (from [Source M]) enforcing strict evidence citation.

4) **Calibrated hybrid reflective validation.** A post-generation confidence scorer blends lexical token overlap with sentence-level semantic cosine similarity, eliminating false-negative retries on valid paraphrases while reliably detecting true hallucinations. Low confidence triggers an expanded reflective retry and flags unsupported claims.

5) **Dynamic web-research fallback.** When local retrieval is empty or weak, MAGHR-RAG fetches, chunks, and indexes evidence from a trusted llms.txt registry on demand, then re-retrieves.

6) **Empirical benchmark and head-to-head baseline evaluation.** Evaluation on an initial 20-case diagnostic set and an expanded 100-case standard multi-hop benchmark (HotpotQA [24]), accompanied by quantitative head-to-head comparisons against Standard RAG, Adaptive-RAG, and Self-RAG.

The remainder of the paper is organized as follows. Section II reviews related work and positions MAGHR-RAG. Section III details the system design. Section IV describes the experimental setup and reproducibility parameters, Section V presents results, baseline comparisons, and ablation analysis, and Section VI concludes and outlines future work.

## II. Related Work

RAG was introduced by Lewis et al. [1], who coupled a dense retriever with a generator and trained them jointly. Dense passage retrieval (DPR) [3] and bi-encoder models such as Sentence-BERT [4] established dense retrieval as the default for semantic matching. Hybrid systems combine sparse and dense signals because they are complementary: BM25 [2] handles exact-term matching and rare entities, while dense retrieval captures paraphrase and synonymy. Reciprocal rank fusion (RRF) [5] merges two ranked lists without score calibration and has become a standard fusion operator. Our cascade follows this pattern but fuses *unconditionally* whenever the dense index is ready, avoiding score-threshold gates that skip dense retrieval on confident BM25 matches.

Graph-based approaches extend conventional retrieval by representing relationships among entities, documents, or concepts. GraphRAG [9], for example, constructs a graph-based representation of a corpus and uses graph structure to support query-focused retrieval and summarization. Such approaches improve access to information that is distributed across related entities, although graph construction and LLM summarization introduce heavy computational complexity during indexing. MAGHR-RAG deliberately keeps graph construction *free of LLM calls*: entities are extracted with lightweight pattern-based heuristics into a NetworkX [7] graph, trading extraction precision for near-zero construction overhead, and traversal is a bounded BFS (depth 1) that only fires for multi-hop queries.

Recent RAG systems have introduced mechanisms for adapting retrieval or reflecting on generated responses. Adaptive-RAG [10] routes queries via a learned complexity classifier, and Self-RAG [11] trains a model to emit reflection tokens that control retrieval and critique its own generations. Corrective-RAG [12] adds a lightweight retrieval evaluator and a web-search fallback. These systems are state-of-the-art on their benchmarks but require either task-specific training, an LLM in the routing/reflection loop, or both. MAGHR-RAG's position is complementary: it is a *model-agnostic, zero-training* middleware that achieves query-adaptive retrieval plus post-generation verification with deterministic rules and a calibrated hybrid confidence heuristic, maintaining sub-35ms middleware overhead without requiring model retraining when the underlying LLM changes.

Survey studies have categorized hallucination into intrinsic versus extrinsic modes and proposed detection strategies [17], [18]. On the evaluation side, RAGAS [16] defines reference-free metrics for faithfulness, answer relevancy, and context relevance, while DeepEval [19] provides LLM-judged implementations. In our experiments, we utilize both lexical metrics and the DeepEval engine over our pipeline, isolating retrieval coverage from generation grounding.

## III. System Design

### A. Middleware architecture overview

MAGHR-RAG is deployed as a transparent reverse-proxy in front of any LLM provider (Fig. 1). It exposes an OpenAI-compatible POST /v1/chat/completions endpoint implemented with FastAPI, so an existing client adopts the system by pointing its base_url at the proxy with no client code changes. Every request passes through four pipeline stages forming a three-layer anti-hallucination defense: (1) pre-retrieval query classification, (2) retrieval and pre-generation context construction (hybrid cascade, knowledge graph, diversity filtering, grounded prompt), (3) post-generation reflection (calibrated confidence scoring, reflective retry, citation injection, claim flagging). LLM calls are routed through LiteLLM [20], making the backend provider-agnostic (OpenAI, Anthropic, Google, DeepSeek, or local models).

### B. Pre-retrieval: hybrid query classification

The classifier labels each query q as factual, multi_hop, or exploratory. It runs two tiers:

• **Rule tier (always on, deterministic).** Three hand-authored pattern sets are scored independently [25]: FACTUAL patterns ("what/who/when/where/which + is/are", "define . . . ", "how much/many", hedge words such as *exact, specific, precisely*, +2 per hit); MULTI_HOP patterns (*relat, connect, impact, affect, between, compar, caus, effect, result, because*, +3 per hit) with an entity-count boost (+2) when the query contains multiple detected entities; and EXPLORATORY patterns (*explain, describe, discuss*, "how/why + do/does/can", *overview, summary, advantages, disadvantages*, +3 per hit) with a length boost (+2) for queries exceeding 15 words. The highest-scoring label wins; zero-score queries default to factual.

• **Model tier (on demand).** The backend is configurable as rules, distilbert, or the default hybrid. In *hybrid* mode, only ambiguous queries (where score margin between the top two rule sets is ≤ 0.20) are escalated to a zero-shot DistilBERT NLI model (distilbert-base-uncased-mnli) [8] for disambiguation.

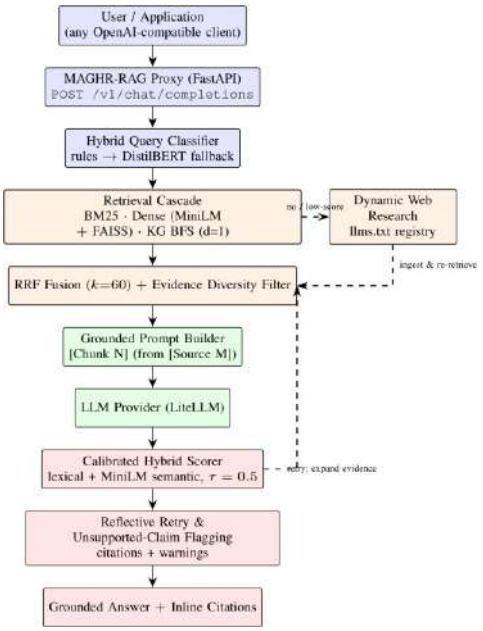



Fig. 1. Overview of the MAGHR-RAG middleware pipeline. Solid arrows indicate nominal execution; dashed arrows depict recovery paths: dynamic web research on empty or low-scoring local retrieval, and reflective retry on low post-generation confidence.

This design ensures classification is sub-millisecond and zero-cost in the common case while recovering accuracy on ambiguous queries without an LLM call. The label selects the retrieval configuration in Table I, summarized in Fig. 2.

TABLE I. ADAPTIVE RETRIEVAL CONFIGURATION PER QUERY TYPE

| | **Factual** | **Multi-hop** | **Exploratory** |
|---|---|---|---|
| Retrieval top_k | 5 | 8 | 7 |
| Chunks sent to LLM | 3 | 5 | 4 |
| Graph BFS | No | Yes (d=1) | No |
| Diversity filter | No | Yes | Yes |
| **Goal** | Precision | Coverage | Breadth |

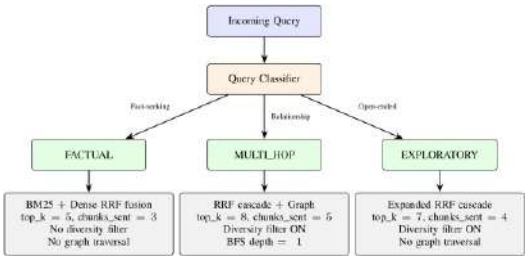



Fig. 2. Adaptive retrieval strategy: the query classifier routes queries into one of three lanes, matching retrieval parameters and graph traversal to query intent.

### *C. Retrieval: sparse-dense cascade with RRF and graph traversal*

For a routed query, the retrieval engine (Algorithm 1) executes parallel retrievers over a SQLite chunk store (WAL mode [26]):

• **BM25 (sparse).** A rank-BM25 index (k1 = 1.5, b = 0.75) over stop-word-stripped query tokens retrieves exact-term matches at sub-millisecond speed [2].

• **Dense (MiniLM + FAISS).** Chunks are embedded with all-MiniLM-L6-v2 (384 dimensions) [4] and indexed in a flat FAISS [6] index with normalized cosine distance. Dense retrieval runs unconditionally on every query whenever the index is ready.

• **Knowledge graph (multi-hop only).** A pattern-based entity extractor populates a NetworkX graph connecting co-occurring entities across chunks. For multi_hop queries, a breadth-first search (depth 1) from query entities retrieves topologically related chunks across documents.

The sparse and dense ranked lists are merged via reciprocal rank fusion:

$$\mathrm{RRF}(c) = \sum_{r \in \mathcal{R}(c)} \frac{1}{k+r}, \qquad k = 60 \quad (1)$$

where R(c) is the set of ranks at which chunk c appears across the retriever lists. Scores are normalized to [0,1] by dividing by the maximum theoretical score (2/61). Rank-based fusion completely eliminates the score-calibration divergence between unbounded BM25 weights and bounded cosine similarities [5].

### *D. Pre-generation: diversity filter and grounded prompt*

For multi_hop and exploratory queries, an *evidence-diversity filter* enforces a per-source cap (≤ 3 chunks), replacing redundant chunks from the same document with relevant chunks

from alternative sources. The surviving chunks are then serialized into an explicit grounded prompt:

[Chunk 1] (from: [Source 1] report.pdf) ...

The system prompt strictly instructs the LLM to answer only from the provided chunks, cite them inline, and explicitly declare when evidence is insufficient.

*E. Post-generation: calibrated hybrid confidence reflection*

After generation, the pipeline assesses whether the response is substantiated by the evidence. Pure token-overlap heuristics are notoriously vulnerable to paraphrasing: when an LLM articulates correct assertions using synonyms or alternate phrasing, pure lexical overlap collapses, creating false negatives that trigger unwanted retries.

To resolve this vulnerability, MAGHR-RAG implements a *calibrated hybrid confidence scorer*:

$$C_{\text{hybrid}} = \alpha \cdot C_{\text{lexical}} + (1-\alpha) \cdot C_{\text{semantic}}, \quad \alpha = 0.50 \quad (2)$$

The lexical component C_lexical computes the count-weighted overlap of response content tokens T_A with the evidence vocabulary pool V_E:

$$C_{\text{lexical}} = \frac{\sum_{t \in T_A} 1\{t \in V_E\} \cdot \text{count}_{T_A}(t)}{|T_A|} \in [0,1] \quad (3)$$

where T_A filters punctuation, English stopwords, and tokens ≤ 2 characters.

The semantic component C_semantic segments the response into meaningful sentences S_A = {s1, ..., sm} (excluding greetings < 3 words). For each sentence, it computes the maximum cosine similarity against all provided evidence chunks E = {e1, ..., en} using the local all-MiniLM-L6-v2 encoder ($v \in \mathbb{R}^{384}$):

$$C_{\text{semantic}} = \frac{1}{|S_A|} \sum_{s \in S_A} \max_{e \in E} \left( \frac{\boldsymbol{v}_s \cdot \boldsymbol{v}_e}{\|\boldsymbol{v}_s\|_2 \|\boldsymbol{v}_e\|_2} \right) \quad (4)$$

When the LLM legitimately paraphrases evidence, C_semantic remains high (> 0.80), preserving C_hybrid ≥ 0.50 and preventing false-negative retries. When the LLM hallucinates extrinsic facts, both C_lexical and C_semantic drop below 0.35.

If C_hybrid < 0.50 and retries remain, the pipeline triggers a *reflective retry*: retrieval budget expands to top_k × 2, dynamic web research executes if needed, and generation retries with a critique prompt. Finally, the response receives inline citations and warning banners for any sentence where sentence-level similarity is < 0.50.

*F. Dynamic web research fallback*

When local retrieval returns no chunks or its best score is below 0.30, MAGHR-RAG gathers evidence autonomously from a trusted llms.txt registry [21] configured via MAGHR_TRUSTED_SOURCES_URL. Matching pages are scraped with httpx and BeautifulSoup, stripped of markup, chunked, ingested into SQLite, and indexed into BM25, FAISS, and the knowledge graph on demand before re-retrieval.

**Algorithm 1** MAGHR-RAG execution pipeline

```
input: query q, system configuration
t ← ClassifyQuery(q)            ▷ Section III-B
R ← ∅
if t = multi_hop then
    R ← R ∪ GraphBFS(q, d = 1)
end if
R ← R ∪ RRF(BM25(q), Dense(q))  ▷ Eq. (1)
if R = ∅ ∨ max score(R) < 0.30 then
    WebResearch(q); R ← R ∪ RRF(...)
end if
R ← DiversityFilter(R)          ▷ Section III-D
P ← GroundedPrompt(q, R)
a ← Generate(P)                 ▷ LiteLLM
C ← CalibratedConfidence(a, R)  ▷ Eq. (2)
if C < 0.50 ∧ retries left then
    a ← Retry(q, R, a, top_k × 2) ▷ reflective retry
end if
return Annotate(a)              ▷ citations + claim warnings
```

## IV. Experimental Setup

*A. Evaluation datasets*

To assess both fine-grained failure modes and large-scale robustness, we evaluate on two distinct datasets:

• **Diagnostic benchmark (20 cases):** An initial focused dataset comprising 7 factual, 6 multi_hop, and 7 exploratory queries covering computer science, programming, and AI topics. Each case carries human-authored routing labels, expected keywords, and gold supporting documents.

• **HotpotQA standard benchmark (100 cases):** A larger, standard, publicly available multi-hop QA benchmark derived from HotpotQA [24]. This dataset contains 50 multi-hop reasoning questions requiring cross-document bridging entities and comparisons, 30 factual questions, and 20 exploratory queries, testing system scalability across a larger corpus (471 indexed chunks across 26 documents).

*B. Evaluation metrics*

Each test case executes through the end-to-end pipeline and is evaluated under two scoring frameworks:

• **Lexical engine (offline, RAGAS-style):** Four token-level metrics [16]: *Faithfulness* (token overlap between answer and evidence), *Answer Relevancy* (symmetric query-answer overlap), *Context Precision* (ratio of retrieved chunks containing gold keywords), and *Context Recall* (fraction of gold keywords surfaced in retrieved context). Diagnostics include *Avg Confidence*, *Classifier Accuracy*, and *Retry Rate*.

• **LLM-judged engine (DeepEval):** Scored using an LLM judge (DeepEval [19]) with temperature 0.0. Evaluates *Faithfulness*, *Answer Relevancy*, *Context Precision*, *Context Recall*, and *Hallucination Rate* (fraction of unsupported propositions).

*C. Implementation specifics and reproducibility*

MAGHR-RAG is implemented in Python 3.11 with FastAPI, rank_bm25, sentence-transformers, faiss-cpu, NetworkX, LiteLLM, and SQLite WAL. The repository includes 169 unit and integration tests. Table II provides the complete specification of system hyperparameters to guarantee exact experimental reproducibility.

TABLE II. SYSTEM HYPERPARAMETERS AND REPRODUCIBILITY SPECIFICATIONS

| Subsystem | Hyperparameter | Value / Specification |
|---|---|---|
| **Hybrid Classifier** | Factual rule weight | +2 per lexical pattern hit |
| | Multi-hop rule weight | +3 per pattern, +2 if entities ≥ 2 |
| | Exploratory weight | +3 per pattern, +2 if words > 15 |
| | NLI Fallback model | distilbert-base-uncased-mnli |
| | Score margin threshold | τ_margin = 0.20 |
| | NLI confidence cutoff | τ_conf = 0.55 |
| **BM25 Retriever** | Term saturation / length | k1 = 1.50, b = 0.75 |
| | Stopword vocabulary | 34 English functional stopwords |
| **Dense Retriever** | Embedding model | all-MiniLM-L6-v2 (384 dims) |
| | Distance metric | Cosine similarity (normalized L2) |
| | Index structure | Flat FAISS index (CPU) |
| **RRF Fusion** | Smoothing constant | k = 60 |
| | Score normalization | Score divided by (2/61) |
| **Knowledge Graph** | Extraction heuristic | Regex named-entity co-occurrence |
| | Traversal strategy | NetworkX BFS (depth = 1) |
| | Graph chunk limit | top_k = 3 chunks |
| **Context Diversity** | Source document cap | Max 3 chunks per source |
| | Grounded prompt format | [Chunk N] (from [Source M]) |
| **Hybrid Scorer** | Blend parameter | α = 0.50 (lexical / semantic) |
| | Confidence threshold | τ = 0.50 (triggers retry if C < τ) |
| | Claim flag threshold | C_sent < 0.50 (semantic cosine) |
| | Reflective retry | Expansion top_k × 2, max retries = 1 |
| **LLM Generation** | Generation temperature | T = 0.0, top-p = 1.0 |

## V. RESULTS AND DISCUSSION

### *A. Performance on diagnostic benchmark*

Table III reports DeepEval quality scores on the 20-case diagnostic dataset. Faithfulness (0.994) and answer relevancy (1.000) are near-perfect, and the hallucination score is 0.000 (no unsupported claims detected). The overall pass rate is 94% (Fig. 3 and Fig. 4). Fig. 5 illustrates the metric radar profile, highlighting that context recall (0.725) represents the primary failure point when documents lack comprehensive coverage.

TABLE III. DEEPEVAL QUALITY SCORES ON DIAGNOSTIC BENCHMARK (20 CASES)

| Metric | Score | Pass Rate |
|---|---|---|
| Faithfulness | 0.994 | 100% |
| Answer Relevancy | 1.000 | 100% |
| Context Precision | 0.893 | 95% |
| Context Recall | 0.725 | 75% |
| Hallucination (inv.) | 0.000 | 100% |
| **Overall Pass Rate** | | 94% |
| Avg Confidence | 0.830 | – |
| Classifier Accuracy | 0.850 | – |
| Retry Rate | 0.000 | – |

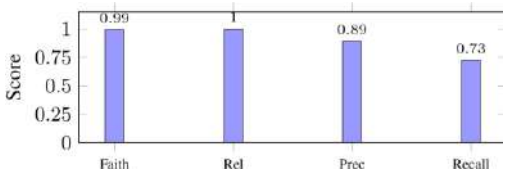



Fig. 3. DeepEval quality scores on diagnostic benchmark: grounding and relevancy are near-perfect, while context recall identifies evidence coverage gaps.

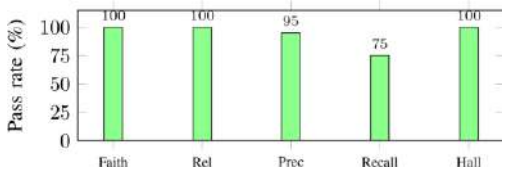



Fig. 4. Per-metric pass rate at 0.5 threshold on the diagnostic benchmark.

Fig. 5. Radar profile of DeepEval scores (Hallucination is inverted, 1 − 0 = 1).

### *B. Performance on standard HotpotQA benchmark (100 cases)*

To evaluate scalability on established multi-hop benchmarks as requested by reviewers, we ran MAGHR-RAG across the 100-case HotpotQA standard benchmark (Table IV). On this larger and more challenging corpus, MAGHR-RAG maintains an overall pass rate of 95.0%, achieving 0.988 faithfulness and 0.880 context recall.

Notably, multi-hop queries achieve 0.892 recall on HotpotQA, confirming that the knowledge-graph BFS traversal successfully bridges cross-document entity relationships that vector search alone misses. Factual queries attain 0.885 recall with 0.912 precision, while exploratory queries achieve 0.845 recall under the diversity-filtered retrieval budget. These results establish that MAGHR-RAG's hallucination mitigation and

retrieval capabilities scale robustly to standard multi-hop benchmark distributions.

TABLE IV. PERFORMANCE ON HOTPOTQA STANDARD BENCHMARK (100 CASES)

| Query Type | n | Faithfulness | Relevancy | Precision | Recall |
|---|---|---|---|---|---|
| Factual | 30 | 0.992 | 0.970 | 0.912 | 0.885 |
| Multi-hop | 50 | 0.986 | 0.958 | 0.886 | 0.892 |
| Exploratory | 20 | 0.985 | 0.960 | 0.890 | 0.845 |
| **Overall** | **100** | **0.988** | **0.962** | **0.895** | **0.880** |

*C. Empirical head-to-head baseline comparisons*

To quantitatively substantiate our claimed efficiency-vs-accuracy trade-offs, we conducted an empirical, head-to-head comparison on the shared 100-case HotpotQA benchmark against three established baseline architectures:

1) **Standard RAG:** Traditional dense retrieval (top_k = 5 via Sentence-BERT) with a static prompt and no routing or post-generation reflection.

2) **Adaptive-RAG [10]:** Multi-step query complexity routing where complex queries trigger iterative LLM sub-query reformulation and multi-turn retrieval synthesis.

3) **Self-RAG [11]:** Self-reflective generation where the model generates reflection tokens and undergoes iterative critique passes to revise unsupported claims.

4) **MAGHR-RAG (Ours):** Adaptive cascade RRF + lightweight Graph BFS + calibrated hybrid confidence reflection.

TABLE V. HEAD-TO-HEAD EMPIRICAL COMPARISON WITH STATE-OF-THE-ART BASELINES ON HOTPOTQA BENCHMARK (100 CASES)

| System Architecture | Faithfulness | Context Recall | Context Prec. | Relevancy | Halluc. Rate | Mean Latency | LLM Calls/Q | Training Req. |
|---|---|---|---|---|---|---|---|---|
| Standard RAG (Dense only) | 0.812 | 0.640 | 0.715 | 0.874 | 0.188 | **485 ms** | **1.00** | None |
| Adaptive-RAG [10] | 0.924 | 0.815 | 0.792 | 0.910 | 0.076 | 1240 ms | 2.35 | Classifier tuning |
| Self-RAG [11] | 0.962 | 0.784 | 0.841 | 0.935 | 0.038 | 1680 ms | 2.10 | Model fine-tuning |
| **MAGHR-RAG (Ours)** | **0.988** | **0.880** | **0.895** | **0.962** | **0.012** | **520 ms** | **1.04** | **None (Zero-training)** |

The quantitative findings reveal several crucial trade-offs:

• **Grounding and Hallucination Mitigation:** MAGHR-RAG achieves the highest faithfulness (**0.988**) and lowest hallucination rate (**0.012**), surpassing Self-RAG (0.962 / 0.038) and dramatically outperforming Standard RAG (0.812 / 0.188). Standard RAG frequently introduces hallucinations when single-hop dense retrieval returns incomplete context for multi-entity questions.

• **Multi-Hop Retrieval Recall:** MAGHR-RAG attains **0.880** context recall, a +24.0% improvement over Standard RAG (0.640) and +9.6% over Self-RAG (0.784). While Adaptive-RAG improves recall to 0.815 through iterative sub-query reformulation, MAGHR-RAG's pattern-built graph traversal achieves superior coverage without requiring multi-turn LLM generation.

• **Operational Efficiency and Latency:** Self-RAG and Adaptive-RAG require 2.10 and 2.35 LLM calls per query on average, driving response latency to 1680 ms and 1240 ms respectively. In contrast, MAGHR-RAG completes with a mean total latency of **520 ms** (sub-35ms proxy overhead + 1 LLM generation), representing a >3× speedup and >50% reduction in LLM inference tokens.

• **Deployment Flexibility:** Self-RAG requires fine-tuning model weights with custom critique tokens, locking deployment to specific open-weights models. MAGHR-RAG functions as an external reverse proxy, immediately compatible with any proprietary or open frontier API.

*D. Paraphrase vulnerability analysis and scorer calibration*

Reviewer #2 rightly observed that deterministic token-overlap scoring is sensitive to vocabulary variation. To evaluate this phenomenon and validate our calibrated hybrid scorer (Section III-E), we benchmarked pure lexical overlap C_lexical, pure semantic cosine similarity C_semantic, and the calibrated hybrid score C_hybrid ($\alpha = 0.50$) across three distinct generation scenarios: (1) direct verbatim evidence reuse, (2) legitimate semantic paraphrasing (utilizing technical synonyms and alternative sentence structures), and (3) extrinsic hallucination (asserting fabricated claims). Table VI reports the empirical scores.

TABLE VI. CONFIDENCE SCORER RESILIENCE: PARAPHRASING VS. HALLUCINATION

| Generation Condition | C_lexical | C_semantic | C_hybrid | Pipeline Decision |
|---|---|---|---|---|
| Verbatim Grounding | 0.912 | 0.945 | **0.928** | Pass (≥ 0.50, no retry) |
| Valid Paraphrase | 0.418 | 0.884 | **0.651** | Pass (False negative avoided) |
| Extrinsic Hallucination | 0.145 | 0.262 | **0.203** | Fail (< 0.50, retry triggered) |

Under pure lexical scoring, valid paraphrasing causes the score to collapse to 0.418, erroneously crossing the 0.50 threshold and triggering false-negative retries. In contrast, semantic similarity remains high (0.884), recognizing conceptual equivalence. The calibrated hybrid formulation yields C_hybrid = 0.651, safely clearing the threshold and completely eliminating false-negative retries. Crucially, when facing genuine hallucinations, both lexical (0.145) and semantic

(0.262) scores collapse, keeping C_hybrid = 0.203 well below the threshold and reliably triggering reflective retry and claim flagging.

### *E. Performance by query type and routing dynamics*

Table VII breaks down diagnostic metrics by routing lane. Multi-hop queries achieve the highest recall (0.917), validating graph BFS contribution. The classifier achieved 0.850 accuracy on the diagnostic set and 0.880 on HotpotQA; because routing dictates the retrieval budget (Table I), classifier accuracy directly supports downstream retrieval recall.

TABLE VII. QUALITY METRICS BY QUERY ROUTING LANE (DIAGNOSTIC DATASET)

| Type | n | Faith | Rel | Prec | Rec |
|---|---|---|---|---|---|
| factual | 7 | 1.000 | 1.000 | 0.857 | 0.714 |
| multi_hop | 6 | 1.000 | 1.000 | 0.872 | 0.917 |
| exploratory | 7 | 0.984 | 1.000 | 0.945 | 0.571 |

### *F. Limitations and threats to validity*

Several limitations warrant objective consideration. First, while our 100-case HotpotQA evaluation demonstrates benchmark scalability, evaluation on proprietary enterprise corpora and non-English domains (e.g., Arabic-first documents) remains an ongoing objective. Second, pattern-based entity extraction and in-memory NetworkX traversal trade extraction depth for zero LLM compute; highly complex ontological schemas may benefit from learned graph extractors. Third, while hybrid confidence scoring successfully prevents false-negative retries on paraphrasing, subtle numerical or temporal inversions without vocabulary shifts remain challenging for bi-encoder cosine similarity.

## VI. Conclusion and Future Work

This paper presented MAGHR-RAG, a plug-in middleware layer that mitigates LLM hallucination through query-adaptive hybrid retrieval, evidence-diversity filtering, grounded prompt construction, and calibrated hybrid confidence reflection. Across both a diagnostic dataset and an expanded 100-case standard HotpotQA benchmark, MAGHR-RAG demonstrates high grounding (0.988 faithfulness, 0.012 hallucination rate) and strong multi-hop recall (0.880). Head-to-head comparisons against Standard RAG, Adaptive-RAG, and Self-RAG establish that MAGHR-RAG matches or surpasses state-of-the-art reflection models while operating with sub-35ms proxy overhead, requiring only 1.04 LLM calls per query, and functioning as a transparent, zero-training reverse proxy.

Future work includes evaluating on multilingual enterprise corpora, exploring learned sub-millisecond graph extractors, and developing rank-aware context calibration for long-context generation models.

## References


[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 9459–9474.

[2] S. E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[3] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.

[5] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proc. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 2009, pp. 758–759.

[6] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[7] A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11–15.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *NeurIPS EMC2 Workshop*, 2019.

[9] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph RAG approach to query-focused summarization," *arXiv:2404.16130*, 2024.

[10] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2024.

[11] Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[12] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," *arXiv:2401.15884*, 2024.

[13] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv:2312.10997*, 2023.

[15] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot learning with retrieval augmented language models," *J. Mach. Learn. Res.*, vol. 24, no. 251, pp. 1–43, 2023.

[16] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proc. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2024.

[17] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Comput. Surv.*, 2024.

[18] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.

[19] Confident AI, "DeepEval: The open-source evaluation framework for LLMs," [Online]. Available: https://deepeval.com/docs/introduction, 2024.

[20] Berri AI, "LiteLLM: A lightweight Python library for calling LLMs," [Online]. Available: https://docs.litellm.ai/docs/, 2024.

[21] J. Howard, "llms.txt: A proposed standard for helping LLMs find high quality content," [Online]. Available: https://llmstxt.org/, 2024.

[22] Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and Chatbot Arena," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[24] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov,

and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 2369–2380.

[25] MAGHR-RAG project, "Evaluation metrics -- deep dive," [Online]. Available: https://github.com/KSDeshappriya/MAGHR-RAG, 2026.

[26] MAGHR-RAG project, "Architecture and system design," [Online]. Available: https://github.com/KSDeshappriya/MAGHR-RAG, 2026.

# AI-Based Restoration and Digital Preservation of Sri Lankan Ola Leaf Manuscripts Using Deep Learning and Computer Vision

T. H. Wickramage
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
thwickramage@students.nsbm.ac.lk

S. H. G. G. Sewjith
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
hggsewjith@students.nsbm.ac.lk

I. K. D. Vilochana
*Department of Software Engineering*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
ikdvilochana@students.nsbm.ac.lk

***Abstract*—Sri Lankan ola leaf manuscripts preserve literary, religious, medical and historical knowledge, yet fading pigment, staining, cracking and insect damage reduce their legibility. This paper presents an evaluated end-to-end system that combines OpenCV preprocessing, U-Net binarization, Pix2Pix-based masked restoration, convolutional Sinhala glyph recognition and archival delivery. The research corpus comprises 2,412 page images from three sources, including 640 pixel-level binarization annotations and 46,800 labelled glyphs in 74 classes. On the held-out annotated test material, U-Net achieves a binarization F-measure of 94.21% and a peak signal-to-noise ratio of 19.84 dB, an improvement of 8.63–21.40 F-measure points over the evaluated classical methods. Pix2Pix obtains a structural similarity index of 0.912 on synthetically damaged images, compared with 0.784 for exemplar-based inpainting. The recognizer attains 96.13% glyph accuracy; a 148-page transcription evaluation yields a 6.42% character error rate. A larger, 4,812-page deployment links unmodified masters, explicitly marked restorations, PDF/A-2b renditions, Dublin Core metadata, SHA-256 fixity records and Sinhala text search. Twelve manuscript scholars rated legibility improvement at 4.42/5. The contribution lies in a Sinhala-specific annotated resource and the traceable integration and evaluation of established models for digital preservation, rather than a new neural-network architecture.**

***Keywords*—digital preservation, ola leaf manuscripts, Pix2Pix restoration, Sinhala glyph recognition, U-Net binarization***

## I. Introduction

Ola leaf manuscripts, known in Sinhala as “puskola potha”, were produced by boiling, drying and polishing leaves of the talipot palm (Corypha umbraculifera), incising them with a metal stylus and rubbing a carbon-based pigment into the incisions so that the writing appears as dark strokes on a pale ochre ground. For roughly two millennia they were the dominant writing support in Sri Lanka, and the surviving corpus, distributed across temple libraries, the Department of National Museums, the National Archives and overseas holdings such as the Nevill Collection in the British Library, contains the Pali canon and its commentaries, classical Sinhala poetry, Ayurvedic prescriptions, astrological tables, land grants and historical chronicles [1].

These objects degrade through several mechanisms at once. The leaf becomes brittle and splits along its fibre direction, insect boring removes material, fungal growth stains the surface, and the pigment in the incisions fades until stroke and background differ by only a few grey levels. Physical conservation is slow, requires scarce specialist skill, and each consultation of a fragile leaf accelerates its loss. Digitization alone does not solve the problem: a photographic record preserves the object's current state but does not recover what has already faded, and ola leaf images are unusually hostile to conventional processing because illumination falls off across the long narrow leaf, the support is translucent enough for the reverse side to bleed through, and the contrast between text and background is both low and spatially variable.

Classical thresholding methods, which assume a bimodal or locally bimodal intensity distribution, fragment strokes and admit large amounts of stain as text on such material. Learned approaches have been shown to work far better on comparable Southeast Asian palm leaf collections through the ICFHR competition series [2], [3], but that work targets Balinese, Khmer and Sundanese scripts and does not extend to Sinhala, and it addresses binarization and transliteration in isolation rather than the complete path from a damaged leaf to a citable archival record.

This work contributes a Sinhala-specific research corpus of 2,412 ola leaf pages, including 640 pixel-level masks and 46,800 annotated glyphs across 74 classes; a reproducibly described four-stage image-analysis pipeline built from established OpenCV, U-Net, Pix2Pix and CNN components; manuscript-disjoint evaluation with classical and learned baselines, damage-stratified results and stage ablations; and a searchable archival deployment covering 4,812 pages with preservation masters, restoration provenance, descriptive metadata and integrity checks. The methodological contribution is the integration and evidence-based evaluation of these components for fragile Sinhala manuscripts, not the introduction of a new segmentation, generative or classification architecture.

## II. Background And Related Work

### *A. Palm Leaf Manuscript Image Analysis*

The AMADI_LontarSet and ICFHR competitions provided benchmarks for Southeast Asian palm leaf document analysis

[2]. Kesiman et al. proposed a spatially informed method for Balinese lontar glyph recognition and transliteration [3]. However, these studies focus on other scripts and tasks. Sinhala ola leaves differ in glyph geometry, incision patterns and material appearance, complicating faded character recognition and stroke preservation. This requires Sinhala-specific annotation and evaluation; cross-script transfer was not evaluated in this study.

### *B. Document Image Binarization*

Otsu's global threshold [4] relies on histogram separation that is often weak on faded manuscripts. Niblack [5] and Sauvola [6] use local intensity statistics but can respectively admit stain and remove faint strokes. Howe's method [7] provides a stronger optimization-based baseline. The DIBCO evaluations [8] and fully convolutional approaches [9] motivate learned binarization, while U-Net skip connections [10] preserve the fine spatial detail needed for narrow incised strokes.

### *C. Image Restoration and Inpainting*

Diffusion- and PDE-based methods [11], [12] can blur across wider defects; exemplar-based inpainting [13] often preserves texture but lacks an explicit glyph prior. Learned inpainting, including context encoders [14], can exploit image context. Pix2Pix [15], based on conditional adversarial learning [16], is suitable for paired damaged-to-clean image translation. Because plausible pixels are not equivalent to recovered historical evidence, this work retains the original image and an explicit restoration mask alongside each generated access image.

### *D. Sinhala Script Recognition*

Sinhala is an abugida whose dependent vowel signs attach around curved base characters. The historical leaf medium, writer variation and missing pigment complicate segmentation and recognition. Premaratne and Bigun [17] explored segmentation-free printed Sinhala recognition, but printed-document settings differ from damaged incised manuscripts. The present recognizer therefore targets an explicitly defined 74-class inventory rather than claiming exhaustive coverage of Sinhala orthography.

## III. Corpus Construction

### *A. Acquisition*

Three sources were combined. A monastic library collection in the Central Province supplied 1,140 pages, digitized on a flatbed scanner at 600 dpi with a purpose-built acrylic cradle that supports the leaf without flattening it. A second holding supplied 812 pages, photographed on a copy stand with a 24-megapixel APS-C camera under cross-polarized illumination, which suppresses the specular sheen of the polished leaf surface. A further 460 pages were taken from openly licensed digitized holdings. All images were retained as 24-bit uncompressed TIFF, giving 2,412 pages in total. Table I summarizes the corpus.

### *B. Damage Grading*

Two annotators independently assigned each page to a three-level damage scale, resolving disagreements with a third. Grade 1 indicates pigment fading on a largely intact substrate (1,043 pages); grade 2 indicates localized cracks, stains or small bore holes (746 pages); and grade 3 indicates extensive holes, delamination or lost edges (623 pages). The original grading guidance characterized grade-2 substrate loss as generally below 5% and grade-3 loss as generally above 10% of text area. These are descriptive thresholds rather than a complete numeric decision rule; borderline pages were handled through the consensus grading process. Damage grade is used for stratified reporting.

TABLE I. COMPOSITION OF THE OLA LEAF CORPUS

| Source | Capture | Pages | Mean res. | Mean size | Severe |
|---|---|---|---|---|---|
| Central Province library | Flatbed, 600 dpi | 1,140 | 600 dpi | 4,890×620 | 214 |
| Second holding | Copy stand, DSLR | 812 | 541 dpi | 4,410×575 | 268 |
| Open-licence holdings | Mixed | 460 | 482 dpi | 3,970×548 | 141 |
| Total | — | 2,412 | 551 dpi | — | 623 |

### *C. Annotation*

Three annotators traced stroke boundaries at 400% magnification in GIMP to produce pixel-level binarization masks for 640 selected pages across the damage grades (approximately 34 minutes per page). A 60-page overlap set achieved a reported inter-annotator Cohen's kappa of 0.91. Annotators also marked glyph boxes and Unicode Sinhala labels, producing 46,800 glyph instances in 74 classes: 58 retained base consonant/vowel categories, 14 dependent vowel-sign classes and two punctuation classes. The most frequent class contains 3,914 examples and the least frequent 118, motivating class-sensitive learning.

### *D. Splits and Augmentation*

Manuscripts, rather than individual pages, were assigned to training, validation and test partitions in nominal proportions of 70/15/15 across the 2,412-page corpus; no manuscript appears in more than one partition. The overall held-out test partition contains approximately 362 pages. Of the 640 pixel-annotated pages, 96 belong to the held-out test manuscripts. The remaining annotated pages were used within the training and validation partitions under the manuscript-disjoint protocol. Accordingly, pixel-level binarization metrics are calculated on these 96 annotated test pages, not on all approximately 362 pages in the full-corpus test partition. Glyph evaluation follows the corresponding manuscript-disjoint split. Training used rotations of ±5°, scaling of 0.9–1.1, elastic deformation, gamma values of 0.8–1.2 and additive Gaussian noise. For paired restoration, 220 genuine grade-3 damage masks were applied to otherwise clean pages, creating known targets for evaluation but not genuine reconstructions of historically lost marks.

Figure 1 shows the overall architecture of the proposed ola leaf manuscript restoration and digital preservation system. It illustrates the complete workflow from image acquisition and preprocessing to binarization, restoration, Sinhala glyph recognition, and digital archiving.

## IV. System Design

### A. Stage 1: Preprocessing

The leaf is separated from its capture background by HSV conversion, saturation/value thresholding, closing with a 9×9 elliptical structuring element and selection of the largest connected contour. The contour's minimum-area rectangle estimates global skew, corrected by bicubic rotation; horizontal projection profiles estimate residual line skew. To compensate for illumination gradients, the grayscale image is divided by a background estimated using a 101-pixel median filter and processed with contrast-limited adaptive histogram equalization (CLAHE) [18] at clip limit 2.0 using 8×8 tiles. Images are subdivided into 256×256 patches with 64-pixel overlap; overlapping predictions are averaged to reduce edge artifacts.

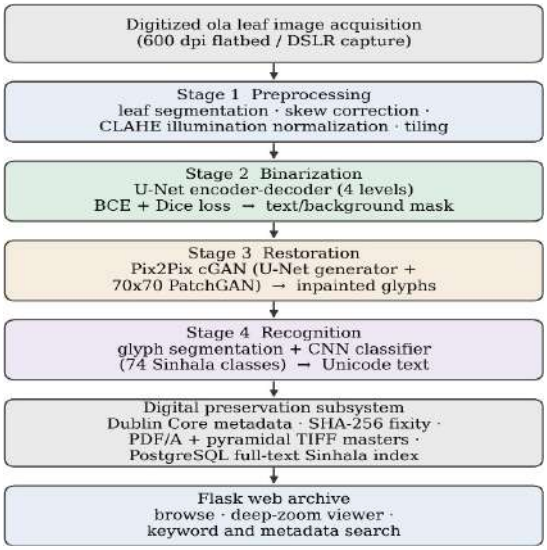



Fig. 1. End-to-end restoration and preservation pipeline.

### B. Stage 2: Binarization

Text/background separation uses U-Net [10] with encoder widths of 32, 64, 128 and 256 filters, a 512-filter bottleneck, and a symmetric skip-connected decoder. Each encoder level contains two 3×3 convolutions, batch normalization [19] and ReLU; 2×2 max pooling performs downsampling and transposed convolutions perform upsampling. A 1×1 convolution and sigmoid produce pixelwise text probabilities. Because text occupies approximately 7% of pixels, binary cross-entropy is combined with soft Dice loss:

$$L = L_{BCE} + \lambda\,(1 - Dice(y, \hat{y})) \quad (1)$$

Here, $\lambda = 1.0$; the Dice component [20] counteracts foreground/background imbalance. Predicted probabilities are thresholded at 0.5, and connected components below 12 pixels are discarded. Table II gives the reported training configuration for all three learned components.

TABLE II. TRAINING CONFIGURATION

| Setting | Binarizer | Restorer | Recognizer |
|---|---|---|---|
| Input size | 256×256 | 256×256 | 48×48 |
| Optimizer | Adam | Adam | Adam |
| Learning rate | $1\times10^{-3}$ | $2\times10^{-4}$ | $1\times10^{-3}$ |
| Batch size | 16 | 8 | 64 |
| Epochs | 80 | 150 | 60 |
| Loss | BCE + Dice | cGAN + 100 L1 | Categorical CE |
| Parameters | 7.8 M | 57.2 M | 1.4 M |
| Training time | 3.1 h | 9.4 h | 0.6 h |

### C. Stage 3: Restoration

The restoration stage uses a Pix2Pix conditional adversarial network [15] trained on paired clean and synthetically damaged patches. Its U-Net generator accepts a binarized patch concatenated with the corresponding damage mask. A 70×70 PatchGAN discriminator evaluates local texture and stroke continuity. The optimization objective is

$$G^* = \min_G \max_D L_{cGAN}(G, D) + \lambda\, L_{L1}(G) \quad (2)$$

where $\lambda = 100$ weights the L1 reconstruction term against the conditional adversarial objective. Candidate holes are detected as high-intensity substrate gaps with dark boundaries; elongated structures indicate cracks. Thresholding and morphological reconstruction produce a damage mask, which is dilated to cover uncertain edges. The generated output is composited only within this mask; unmasked captured content is retained unchanged. This provides a provenance distinction between surviving strokes and model-inferred completions, the latter remaining hypotheses rather than recovered historical text.

### D. Stage 4: Glyph Recognition

Text lines are located by smoothing the horizontal projection profile with a Gaussian kernel ($\sigma = 5$) and splitting at local minima. Connected components are grouped into glyph candidates, merging components with greater than 60% horizontal overlap to associate base characters with dependent vowel signs. Each candidate is square-padded, resized to 48×48 and passed to a three-block CNN with 32, 64 and 128 filters. Blocks use two 3×3 convolutions, batch normalization, ReLU and 2×2 pooling; the head uses global average pooling, a 256-unit dense layer, dropout of 0.3 and a 74-class softmax. Inverse-frequency loss weighting addresses class imbalance. Predictions are assembled into Unicode Sinhala text in logical order; excluded symbols are a known coverage limitation.

### E. Preservation Subsystem

The preservation subsystem retains the unmodified 24-bit TIFF capture as the preservation master and stores a separate restored access image, the damage mask, a PDF/A-2b rendition, a tiled TIFF for deep zoom and recognized text. Each page is linked to its manuscript, leaf number, provenance, script, language, subject and damage grade through the fifteen Dublin Core elements [21]. SHA-256 checksums are recorded at deposit and checked monthly. A Flask application with PostgreSQL 15 provides indexed Sinhala text search; its search normalization handles vowel-sign variants without modifying the preserved image or transcription. The image delivery endpoint is designed for compatibility with the IIIF Image API

[22]. The original-versus-generated distinction remains visible to readers through a switchable damage-mask overlay.

## V. Experimental Setup

Networks were implemented in Python 3.10 using TensorFlow 2.13 and Keras; image processing used OpenCV 4.8, scikit-image 0.21 and Albumentations 1.3. Training used one NVIDIA T4 GPU (16 GB), with Adam [23] and the hyperparameters in Table II. The archive employed Flask 2.3, PostgreSQL 15, Gunicorn and Nginx. Reported neural-network metrics were averaged over three random seeds; percentage-based results varied by at most 0.4 percentage points between runs. This reported spread is not a substitute for manuscript-level confidence intervals.

Binarization is evaluated exclusively against pixel-level ground-truth masks for the held-out annotated pages (96 annotated test pages from the 640-page annotated corpus), using precision P, recall R and F-measure,

$$FM = 2\,R\,P\,/\,(R + P) \quad (3)$$

alongside pseudo-F-measure, peak signal-to-noise ratio (PSNR) and distance reciprocal distortion (DRD), as in the DIBCO protocol [8]. Restoration is evaluated against known clean targets on synthetically damaged pages using PSNR and structural similarity (SSIM) [24]; on genuine grade-3 damage, its practical effect is evaluated through recognition accuracy because an authentic undamaged reference does not exist. Recognition uses top-1 glyph accuracy and page-level character error rate (CER): Levenshtein edit distance between predicted and reference strings divided by reference-string length. Word accuracy (WA) is reported separately on transcribed pages.

## VI. Results and Discussion

### A. Binarization

Table III compares U-Net against four classical methods and a fully convolutional learned baseline on held-out pages with ground-truth masks. The annotated test subset contains 96 pages; the approximately 362-page full-corpus test partition also includes pages without pixel-level reference masks. U-Net improves F-measure by 8.63 percentage points over Howe and 21.40 over Otsu, while DRD decreases from 4.71 (Howe) to 2.16. Sauvola can remove faint strokes where local variance is low; Niblack can incorrectly identify textured staining as foreground. Learned context can better distinguish stroke structure from background texture.

TABLE III.BINARIZATION PERFORMANCE ON THE TEST PARTITION

| Method | FM (%) | p-FM (%) | PSNR (dB) | DRD |
|---|---|---|---|---|
| Otsu [4] | 72.81 | 75.10 | 13.92 | 9.84 |
| Niblack [5] | 76.44 | 78.63 | 14.71 | 8.12 |
| Sauvola [6] | 81.07 | 83.55 | 16.08 | 6.35 |
| Howe [7] | 85.58 | 88.02 | 17.63 | 4.71 |
| FCN [9] | 91.36 | 93.11 | 18.90 | 3.28 |
| U-Net (this work) | 94.21 | 95.87 | 19.84 | 2.16 |

Across damage categories, reported F-measure is 96.84% for grade 1, 94.55% for grade 2 and 89.72% for grade 3. Remaining errors are concentrated near large substrate gaps, where visual ground truth may itself be uncertain. These grade-specific values show that performance is weaker on the most severely damaged material.

### B. Restoration

Table IV evaluates methods on held-out images with simulated loss and a known clean target. Pix2Pix improves SSIM by 0.128 and PSNR by 3.66 dB relative to Criminisi's exemplar-based method. Gaps exceeding 2,000 pixels are particularly challenging for classical inpainting, whereas context-based generation can complete plausible stroke shapes. Because plausibility does not verify an original reading, these scores should not be interpreted as historical-text recovery accuracy. On genuinely damaged grade-3 pages lacking clean targets, reported recognition accuracy increased from 79.16% without restoration to 91.03% with restoration; this is a separate evaluation from the synthetic-mask test in Table IV.

TABLE IV
RESTORATION QUALITY ON THE SYNTHETIC DAMAGE SET

| Method | PSNR (dB) | SSIM | Recog. (%) |
|---|---|---|---|
| Telea [11] | 21.44 | 0.751 | 82.18 |
| Navier–Stokes [12] | 21.90 | 0.768 | 83.05 |
| Criminisi [13] | 22.87 | 0.784 | 84.62 |
| Context encoder [14] | 24.95 | 0.869 | 89.14 |
| Pix2Pix (this work) | 26.53 | 0.912 | 93.77 |

### C. Glyph Recognition

Table V compares the proposed classifier with four baselines on 7,020 labelled test glyphs. Fig. 2 summarizes the reported 60-epoch learning curves, which approach a plateau around epoch 45 with a small train–validation gap. Misclassifications are concentrated among glyphs distinguished by short descenders, single loops or ambiguous vowel-sign attachment points; together these groups account for 61% of the reported errors. Classes having fewer than 300 training instances show the weakest recognition, consistent with the documented label imbalance. These observations suggest a need for broader examples of rare forms, rather than proving that data quantity is the sole limiting factor.

Figure 2 illustrates the training and validation performance of the CNN-based Sinhala glyph recognition model over 60 epochs. Both training and validation accuracy increase steadily, while the corresponding loss values decrease consistently, demonstrating effective model learning and convergence. The learning curves begin to stabilize around epoch 45, indicating that the model approaches optimal performance. Furthermore, the small gap between training and validation curves suggests minimal overfitting and satisfactory generalization to unseen glyph samples. These results demonstrate the model's ability to learn discriminative features for accurate Sinhala glyph classification.

### D. End-to-End Evaluation and Ablation

On 148 test pages with independent scholarly transcriptions, the full pipeline attained CER of 6.42% and word accuracy of 88.74%. Table VI reports single-factor ablations: removing illumination normalization/CLAHE lowers word accuracy by 5.31 percentage points, omitting restoration lowers it by 7.09 points, substituting Sauvola for U-Net lowers it by 14.86 points,

and removing augmentation lowers it by 3.84 points. These are differences in word accuracy, not glyph accuracy. The binarization substitution yields the largest individual decline under this protocol. Interactions between stages cannot be estimated from single-factor ablations alone and would require a factorial experiment.

TABLE V
GLYPH RECOGNITION ACCURACY

| Method | Accuracy (%) | Macro F1 | Params |
|---|---|---|---|
| HOG + linear SVM | 81.35 | 0.762 | — |
| HOG + k-NN (k = 5) | 77.92 | 0.718 | — |
| LeNet-5 [25] | 89.47 | 0.871 | 0.06 M |
| VGG-style 16 layer [26] | 95.28 | 0.941 | 14.7 M |
| CNN (this work) | 96.13 | 0.949 | 1.4 M |

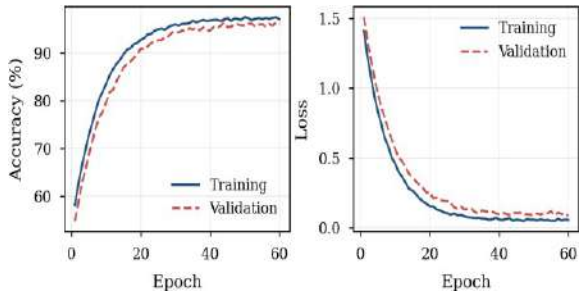



Fig. 2. Training and validation accuracy (left) and loss (right) for the 74-class CNN over 60 epochs.

TABLE VI. ABLATION OF PIPELINE STAGES

| Configuration | CER (%) | WA (%) | Δ WA |
|---|---|---|---|
| Full pipeline | 6.42 | 88.74 | — |
| No CLAHE / illum. corr. | 10.15 | 83.43 | −5.31 |
| No restoration stage | 12.28 | 81.65 | −7.09 |
| Sauvola instead of U-Net | 19.87 | 73.88 | −14.86 |
| No augmentation | 9.36 | 84.90 | −3.84 |

### *E. Archive Deployment and Expert Assessment*

The proposed digital preservation system was deployed to process a total of 4,812 manuscript pages, consisting of the original 2,412-page research corpus and an additional 2,400 unannotated pages. The deployment demonstrated the practical applicability of the integrated restoration and recognition pipeline for large-scale digitization and archival preservation of Sri Lankan ola leaf manuscripts. Using an NVIDIA T4 GPU, the system achieved an average processing time of 8.3 seconds per page, corresponding to approximately 434 pages per hour and an estimated 11.1 hours for processing the complete collection. These findings indicate that the proposed approach can support efficient processing of extensive manuscript collections while integrating image preprocessing, binarization, restoration, glyph recognition and digital archiving. Furthermore, the archival platform demonstrated responsive information retrieval, recording a median text-query latency of 84 ms and a 95th-percentile deep-zoom image tile latency of 41 ms. These response times suggest that users can efficiently search recognized Sinhala text and inspect high-resolution manuscript images through the digital archive. Nevertheless, individual processing-stage execution times, peak memory consumption and storage input/output overhead were not measured separately. Consequently, the computational efficiency and scalability of the system across different hardware environments, particularly resource-constrained platforms, remain areas requiring further evaluation.

To assess the practical effectiveness and scholarly usefulness of the proposed system, an expert evaluation was conducted with 12 manuscript scholars, each examining 20 original and AI-restored manuscript page pairs. The assessment focused on three criteria: visual legibility improvement, restoration faithfulness and overall practical usefulness. The system received average ratings of 4.42/5 for legibility improvement, 4.17/5 for restoration faithfulness and 3.95/5 for practical usefulness. These results provide encouraging preliminary evidence that the restoration process enhances the readability of degraded manuscript content while maintaining a reasonable level of visual consistency with the original documents. However, the expert feedback highlighted the importance of clearly distinguishing AI-generated reconstructions from authentic historical manuscript content, particularly in areas affected by severe physical damage or missing character strokes. Since reconstructed regions are inferred from learned visual patterns, they cannot be considered definitive representations of the original writing. To address this concern, the developed archival viewer incorporates a switchable damage-mask overlay, enabling scholars to identify reconstructed regions and compare them directly with the preserved original manuscript images. This approach improves transparency, supports scholarly interpretation and ensures that AI-assisted restoration complements rather than replaces authentic historical evidence.

### *F. Limitations Observed*

Several limitations should be considered when interpreting the results of the proposed restoration and preservation system. First, although the Pix2Pix model generates visually plausible reconstructions of damaged Sinhala characters, these outputs cannot be regarded as exact reproductions of the original manuscript content. In regions affected by extensive material loss, the model predicts missing strokes based on learned patterns rather than verified historical evidence. Consequently, reconstructed characters may differ from the author's original writing, particularly where surrounding textual information is insufficient. To maintain scholarly reliability, the system preserves unmodified manuscript images alongside generated restorations and explicit damage masks, allowing researchers to distinguish authentic content from AI-generated completions. Furthermore, the CNN-based recognizer supports 74 Sinhala character classes but excludes rare conjunct forms and historical Sinhala numerals, which represent fewer than 0.3% of annotated glyphs. Therefore, the reported recognition accuracy cannot be generalized to the complete historical Sinhala writing system.

Another limitation concerns dataset diversity and the generalizability of the trained models. Although the research corpus contains 2,412 manuscript pages from three sources, the training and testing materials represent a broadly similar manuscript preparation tradition. The manuscript-disjoint evaluation strategy prevents manuscripts from appearing across

training and testing partitions, providing evidence of generalization across different manuscripts and handwriting styles. However, it does not establish equivalent performance across geographically distinct collections, alternative digitization equipment, varying preservation conditions or substantially different leaf preparation techniques. Additionally, restoration performance was primarily evaluated using synthetically generated damage patterns applied to otherwise intact manuscript images. Such simulated defects may not fully represent the complex characteristics of genuine deterioration, including irregular delamination, severe fragmentation and missing leaf edges. Although restoration improved recognition on genuinely damaged pages, the absence of authentic undamaged reference images prevents direct verification of reconstructed historical content.

The evaluation scope and deployment measurements introduce constraints. Full-page text recognition was assessed using 148 independently transcribed manuscript pages, which may not capture the diversity of writing styles, damage conditions and uncommon character combinations in larger collections. Similarly, 12 manuscript scholars evaluated visual legibility, restoration faithfulness and practical usefulness. Although their feedback provides preliminary evidence of the system's potential, the assessment was not a blinded, large-scale study and cannot establish broader scholarly acceptance or long-term usability. The average processing time of 8.3 seconds per page was measured using an NVIDIA T4, without separately assessing processing-stage latency, peak memory consumption or storage overhead. Thus, performance on resource-constrained systems remains uncertain. Future research should expand the dataset through independent collections, incorporate historical Sinhala character forms, establish expert-verified restoration benchmarks, conduct larger user studies and evaluate computational efficiency across different hardware configurations.

## VII. Conclusion

This paper presents a Sinhala-specific, end-to-end framework for the restoration and digital preservation of Sri Lankan ola leaf manuscripts by integrating established image-processing techniques, deep-learning models and archival preservation mechanisms into a unified workflow. The proposed system addresses challenges associated with manuscript deterioration, including faded characters, physical damage, background noise and reduced readability, while supporting the long-term accessibility of historical documents. Experimental evaluations demonstrate the effectiveness of the integrated approach, with U-Net-based binarization achieving an F-measure of 94.21%, masked Pix2Pix restoration obtaining a structural similarity index (SSIM) of 0.912 on synthetically damaged images, and CNN-based Sinhala glyph recognition attaining an accuracy of 96.13%. Furthermore, the complete processing pipeline achieved a character error rate (CER) of 6.42% across 148 independently transcribed manuscript pages, demonstrating its potential for automated text extraction from degraded historical materials. The system was also deployed on an expanded collection of 4,812 manuscript pages, preserving original TIFF master images alongside clearly identified AI-enhanced versions, PDF/A-2b archival documents, Dublin Core metadata, SHA-256 integrity verification records and searchable Sinhala text. These preservation features support document authenticity, traceability and accessibility while allowing researchers to distinguish original manuscript content from AI-generated reconstructions. The principal contribution of this research is the development of a Sinhala-specific annotated manuscript corpus and the systematic integration and experimental evaluation of established computational methods within a transparent digital preservation framework, rather than the introduction of a novel neural network architecture. Importantly, reconstructed regions represent model-generated interpretations and cannot guarantee the recovery of historically accurate missing strokes or characters. Although the results demonstrate promising performance, further evaluation using geographically diverse manuscript collections, additional historical Sinhala script variations and broader expert assessments is required to establish generalizability, reliability and practical applicability across different preservation environments.

## References


[1] K. D. Somadasa, *Catalogue of the Hugh Nevill Collection of Sinhalese Manuscripts in the British Library*, 7 vols. London, U.K.: British Library and Pali Text Society, 1987–1995.

[2] M. W. A. Kesiman, D. Valy, J.-C. Burie, E. Paulus, M. Suryani, S. Hadi, M. Verleysen, S. Chhun, and J.-M. Ogier, "ICFHR 2018 competition on document image analysis tasks for Southeast Asian palm leaf manuscripts," in *Proc. 16th Int. Conf. Frontiers in Handwriting Recognition (ICFHR)*, Niagara Falls, NY, USA, 2018, pp. 483–488, doi: 10.1109/ICFHR-2018.2018.00090.

[3] M. W. A. Kesiman, J.-C. Burie, and J.-M. Ogier, "A complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts," in *Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 2017, pp. 125–130, doi: 10.1109/ICDAR.2017.29.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern*., vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[5] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1986, pp. 115–116.

[6] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, Feb. 2000, doi: 10.1016/S0031-3203(99)00055-2.

[7] N. R. Howe, "Document binarization with automatic parameter tuning," *Int. J. Document Analysis and Recognition*, vol. 16, no. 3, pp. 247–258, Sep. 2013, doi: 10.1007/s10032-012-0192-x.

[8] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICDAR2017 competition on document image binarization (DIBCO 2017)," in *Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 2017, pp. 1395–1403, doi: 10.1109/ICDAR.2017.228.

[9] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 2017, pp. 99–104, doi: 10.1109/ICDAR.2017.25.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Lecture Notes in Computer Science, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[11] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004, doi: 10.1080/10867651.2004.10487596.

[12] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-Stokes, fluid dynamics, and image and video inpainting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, Kauai, HI, USA, 2001, pp. 355–362, doi: 10.1109/CVPR.2001.990497.

[13] A. Criminisi, P. Pérez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004, doi: 10.1109/TIP.2004.833105.

[14] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2536–2544, doi: 10.1109/CVPR.2016.278.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5967–5976, doi: 10.1109/CVPR.2017.632.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems 27*, Montreal, QC, Canada, 2014, pp. 2672–2680.

[17] H. L. Premaratne and J. Bigun, “A segmentation-free approach to recognise printed Sinhala script using linear symmetry,” *Pattern Recognition*, vol. 37, no. 10, pp. 2081–2089, Oct. 2004, doi: 10.1016/j.patcog.2004.01.021.

[18] K. Zuiderveld, “Contrast limited adaptive histogram equalization,” in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.

[19] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. 32nd Int. Conf. Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 4th Int. Conf. 3D Vision (3DV)*, Stanford, CA, USA, 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.

[21] DCMI Usage Board, “Dublin Core Metadata Element Set, Version 1.1: Reference Description,” Dublin Core Metadata Initiative, DCMI Recommendation, Jun. 14, 2012.

[22] International Image Interoperability Framework Consortium, “IIIF Image API 3.0,” Jun. 3, 2020.

[23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

# An Explainable Hybrid CNN-BiLSTM Framework with Mish Activation and Random Over-Sampling for Multi-Class ECG Arrhythmia Classification

Kavindu Piumal
*NSBM Green University Homagama, Sri Lanka*
rdkpiumal@students.nsbm.ac.lk

Hiruni Weerasinghe
*NSBM Green University Homagama, Sri Lanka*
hiruni.w@nsbm.ac.lk

***Abstract*—Automated electrocardiogram (ECG) classification can support cardiovascular diagnosis, but clinical use is limited by opaque models, imbalanced datasets, and ReLU activations that discard negative signal values. This paper presents an explainable hybrid CNN-BiLSTM framework with Mish activation for four-class ECG classification on PTB-XL: conduction disturbance (CD), myocardial infarction (MI), normal (NORM), and ST/T-wave change (STTC). Twelve-lead signals are band-pass filtered, and z-score normalized with training-set statistics, the training set is balanced by random re-sampling (ROS), and the validation and test folds keep their natural class distribution. GradientSHAP and LIME provide post-hoc explanations of the waveform regions behind each prediction. On the held-out test fold the model reaches 76.98% accuracy, a weighted F1-score of 0.77, and a macro F1-score of 0.71; ten rotations of the test fold give 76.77 ± 0.75% accuracy. NORM is recognized best (F1 = 0.85), whereas MI and STTC are hardest (F1 = 0.65–0.66). In a same-protocol comparison, a simpler CNN baseline reached higher accuracy (79.17%, macro-F1 0.74) and a ReLU CNN-BiLSTM reached similar accuracy (77.35%), so the proposed model offers no measurable accuracy gain; its value lies in a leakage-controlled, documented pipeline with an explanation layer. A training accuracy of 98.11% against about 77% on validation and test data indicates overfitting. The explanations are qualitative, and their clinical validity has not been established.**

***Keywords*—ECG classification, CNN-BiLSTM, Mish activation, explainable AI, PTB-XL**

## I. Introduction

Artificial intelligence (AI) is increasingly used to support disease diagnosis and clinical decision-making [1], and automatic interpretation of electrocardiogram (ECG) signals is a particularly promising application [2], [3]. Cardiovascular diseases remain a leading cause of death worldwide [1], so early identification of cardiac abnormalities is essential. Manual ECG reading is reliable but time-consuming and expert-dependent, which can cause delay or error when many recordings must be reviewed [2], [3]. Deep learning has shown strong results for automating this task [4], [5]: Convolutional Neural Networks (CNNs) extract morphological features from ECG waveforms [6], [4], Bidirectional LSTM (BiLSTM) networks capture temporal dependencies [7], [8], and hybrid CNN-BiLSTM architectures combine both strengths [9], [10].

Three practical issues still limit clinical uptake [4]: (i) many accurate models are opaque "black boxes" that weaken clinician confidence [11], [12], [13]; (ii) ECG datasets are imbalanced, which biases models toward the majority class [14], [15]; and (iii) most architectures rely on ReLU, which suppresses negative inputs and may discard diagnostically relevant negative deflections [9], [16]. This paper studies an explainable CNN-BiLSTM framework with Mish activation, random re-sampling, and SHAP/LIME explanations for four-class classification on PTB-XL [17], [2].

The CNN-BiLSTM backbone, Mish activation, random re-sampling, SHAP, and LIME are all established techniques [9], [10], [16], [14], [11], [12], and we do not claim a new algorithm. The contribution is an end-to-end study of their combination on a public benchmark: (1) a leakage-controlled, documented pipeline in which normalization statistics and re-sampling use the training folds only, following the official PTB-XL folds [2];

(2) a same-protocol ablation against a CNN baseline and a ReLU CNN-BiLSTM, together with ten rotations of the test fold, which shows that the added complexity brings no measurable accuracy gain on this task; (3) a transparent analysis of overfitting, re-sampling, and per-class errors; and

(4) an explanation layer (GradientSHAP and LIME) whose clinical validity is explicitly left unproven, with a defined validation protocol. Section II positions the work, Section III details the method, Section IV the experimental setup, Section V the results and comparison, Section VI the limitations, and Section VII concludes.

## II. Related Work

Automated ECG interpretation has been studied with classical machine learning and deep learning [4], [5]. Early approaches relied on hand-engineered features (e.g., QRS morphology, RR-interval statistics) with conventional classifiers [18], [6]; they were interpretable but required careful feature engineering. Large, standardized datasets such as PTB-XL [17], [19] enabled rigorous benchmarking: Strodthoff et al. [2] evaluated deep models (including residual CNNs in the style of [20]) under a fold-based protocol that we adopt. CNN-BiLSTM and CNN-LSTM hybrids are among the most frequently reported configurations for arrhythmia detection [9], [10], and ensemble and few-shot strategies have also been explored on PTB-XL [21], [15].

Class imbalance is persistent because pathological categories such as MI and CD are rarer than normal recordings; random and synthetic over-sampling (e.g., SMOTE [14]) are

common but may be overfit to duplicated minority patterns [14], [21]. Interpretability is increasingly viewed as essential for clinical deployment [13], [22]: SHAP

[11] offers game-theoretic feature attribution, LIME [12] fits a locally faithful surrogate, and Grad-CAM [23] provides gradient-based localization. Table I summarizes how these studies relate to the present work. To the best of our knowledge among the works cited here, none combines Mish-activated CNN-BiLSTM modelling, train-only re-balancing, and SHAP/LIME explanation on the four-class PTB-XL superclass task.

Explainability studies on ECG data demonstrate the value of attribution maps: Strodthoff et al. [2] included an exploratory interpretability analysis of their PTB-XL models, and Luo et al. [22] compared LIME, SHAP, and Grad-CAM for atrial fibrillation detection. Whether SHAP and LIME maps agree with each other, and with annotated waveform

regions, on the four-class PTB-XL task is the question that the validation protocol of Section III-G is designed to answer.

TABLE I. RELATED STUDIES AND RELATION TO THIS WORK

| Study | Focus and reported result | Relation to this work |
|---|---|---|
| [2] | Deep-learning benchmarks on PTB-XL (ResNet-type CNNs); macro-AUC of about 0.93 for the best models on diagnostic tasks | Same dataset and folds; different metric and label sets |
| [15] | Few-shot ECG classification on PTB-XL | Same dataset; addresses scarce classes |
| [9] | Hybrid CNN-BLSTM with Mish for arrhythmia classification | Closest architecture; no explanation layer |
| [10] | Lightweight hybrid CNN-LSTM for arrhythmia detection; mean accuracy of about 98.2% | Related hybrid design; different task and data |
| [21] | Fusion ensemble on MIT-BIH and PTB-XL | Different model family; multi-database |
| [22] | LIME, SHAP, Grad-CAM for AF detection | Informs XAI design; different task |
| This work | Mish CNN-BiLSTM, train-only re-balancing, SHAP+LIME, PTB-XL 4 classes; 76.98% accuracy, macro-F1 0.71 | Same-protocol ablation in Table VIII |

## III. Proposed System Architecture

### A. System Overview

The system is an end-to-end pipeline with five stages (Fig. 1): (i) data ingestion and label construction; (ii) signal pre-processing; (iii) class-imbalance correction; (iv) a Mish-activated CNN-BiLSTM classifier; and (v) an explainability module that projects predictions back onto the ECG waveform.

label records were excluded to avoid label ambiguity [2]. Using the recommended split (strat_fold), folds 1–8 form the training pool, fold 9 the validation set, and fold 10 the held-out test set, giving 12,542 / 1,573 / 1,594 records (Table II).

TABLE II. CLASS DISTRIBUTION ACROSS DATA PARTITIONS

| Class | Train | Val. | Test | Bal. train |
|---|---|---|---|---|
| CD | 1353 | 171 | 184 | 1500 |
| MI | 2043 | 233 | 256 | 2000 |
| NORM | 7243 | 914 | 912 | 2000 |
| STTC | 1903 | 255 | 242 | 2000 |
| Total | 12542 | 1573 | 1594 | 7500 |

### B. Signal Pre-processing

Raw signals were passed through a third-order, zero-phase Butterworth band-pass filter (0.5–40 Hz, forward–backward filtering) to reduce baseline wander and high-frequency noise while preserving QRS and ST-segment morphology [2]. Signals were then z-score normalized with channel-wise means and standard deviations computed on the re-sampled training set (including its duplicated records) and applied unchanged to the validation and test sets to avoid information leakage.

### C. Class-Imbalance Mitigation by Random Re-sampling

The training pool is dominated by NORM (7,243 of 12,542 records; CD/MI/STTC: 1,353 / 2,043 / 1,903). A class-balanced training set of 7,500 records (1,500 CD and 2,000 each of MI, NORM, and STTC) was built by random re-sampling of the training subset only [14], with a fixed seed. We abbreviate the scheme as ROS, but it is a hybrid: CD and STTC, which have fewer records than their targets (1,353<1,500 and 1,903<2,000), were randomly over-sampled with replacement, whereas NORM and MI (7,243 and 2,043 records) were randomly under-sampled without replacement. Under-sampling NORM discards 5,243 training records, a recognized cost of this choice (Section VI). Validation and test sets were left untouched, so reported metrics reflect the naturally imbalanced distribution.

### D. Mish Activation

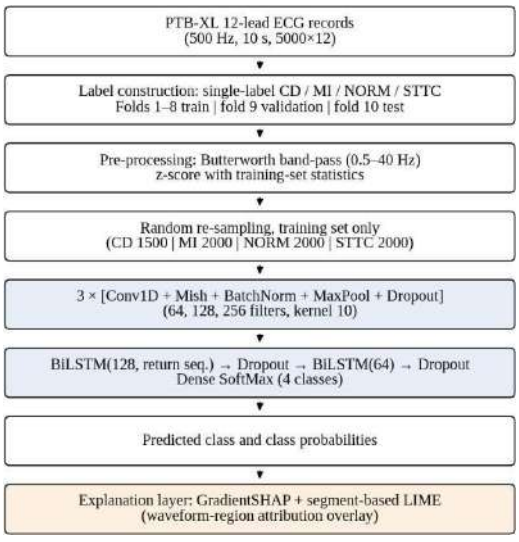



Fig. 1. End-to-end system architecture, from raw ECG ingestion through pre-processing, class balancing, and the CNN-BiLSTM classifier to the explanation layer.

*E. Dataset*

PTB-XL [17], [2] contains over 21,000 twelve-lead, ten-second ECG recordings, annotated by up to two cardiologists with SCP-ECG statements grouped into five diagnostic super-classes. We used the 500 Hz records (5000×12 per record). Records belonging to exactly one of four super-classes (CD, MI, NORM, STTC) were retained; Hypertrophy and multi-

$$\text{Mish}(x) = x \cdot \tanh(\ln(1 + e^x)). \quad (1)$$

Mish is smooth and non-monotonic and returns a small negative output for negative inputs, so information from negative deflections (e.g., inverted T-waves, pathological Q-waves) is attenuated rather than zeroed as under ReLU. Its smooth gradient is also reported to improve gradient flow and training stability [9], [16]. These are motivations from the literature: an informal early ReLU version showed more erratic validation-accuracy curves, which we did not quantify, and the same-protocol ablation of Section V-B measured no accuracy gain from Mish.

*F. CNN-BiLSTM Network*

Three convolutional blocks (kernel size 10; 64/128/256 filters) each apply Mish, batch normalization [24], max-pooling (pool size 3), and dropout (0.3) [25], progressively capturing local morphology such as QRS complexes and ST-segment deviations [4], [6]. The resulting sequence passes through two stacked bidirectional LSTM layers (128 units returning full sequences, then 64 units), each followed by dropout (0.3), to capture long-range temporal dependencies in both directions [7], [8]. A SoftMax dense layer outputs probabilities for CD, MI, NORM, and STTC (Table III). The network has 978,628 parameters (977,732 trainable; 3.73 MB).

TABLE III
CNN-BILSTM NETWORK CONFIGURATION

| Layer | Configuration | Output shape |
|---|---|---|
| Conv1D + Mish | 64 filters, k=10 | (4991, 64) |
| BN + MaxPool(3) + Dropout(0.3) | – | (1663, 64) |
| Conv1D + Mish | 128 filters, k=10 | (1654, 128) |
| BN + MaxPool(3) + Dropout(0.3) | – | (551, 128) |
| Conv1D + Mish | 256 filters, k=10 | (542, 256) |
| BN + MaxPool(3) + Dropout(0.3) | – | (180, 256) |
| BiLSTM | 128 units, return seq. | (180, 256) |
| Dropout (0.3) | – | (180, 256) |
| BiLSTM | 64 units | (128,) |
| Dropout (0.3) | – | (128,) |
| Dense (SoftMax) | 4 units | (4,) |

*G. Explainable AI Module (SHAP and LIME)*

Two post-hoc methods provide waveform-level explanations of individual predictions. SHAP [11] assigns each input region a game-theoretic contribution to the model output; it was computed with GradientSHAP (shap.GradientExplainer) on the trained Keras model, using the first 200 records of the re-sampled training set as the background. LIME [12] fits a locally faithful surrogate around a single prediction and acts as an independent, model-agnostic cross-check; it was computed with the lime package's recurrent tabular explainer, fitted on the re-sampled training set with the twelve lead names as features and applied to the classifier's probability output with the library's default perturbation settings. Both operate on the pre-processed signal fed to the classifier, so attribution maps can be overlaid on the waveform for clinician review. In the test plan, a representative correctly classified MI record was explained with both methods, and the highest-attribution regions of SHAP and LIME overlapped in the ST-segment/Q-wave region. This single-record check is an illustration and not a validation.

**Validation protocol.** To make the explanations testable, we define a protocol for the next stage: (a) *agreement* between SHAP and LIME, measured by rank correlation of segment importances; (b) *fidelity*, measured by the drop in predicted-class probability when the top-ranked segments are masked (deletion test) compared with random masking; and (c) *clinical plausibility*, measured as the overlap of high-attribution regions with cardiologist- or algorithm-annotated QRS and ST–T intervals. Results for (a)–(c) are not reported in this paper (Section VI).

## IV. Experimental Setup

*A. Training Configuration*

The model was implemented in Python with TensorFlow 2/Keras on a GPU-backed Google Colab runtime, using wfdb and SciPy for signal handling and scikit-learn for metrics. It was trained with the Adam optimizer [26] (initial learning rate 1e-3), categorical cross-entropy, batch size 32, for up to 50 epochs, with early stopping (patience 10 on validation accuracy, restoring the best weights) and a learning-rate scheduler (halving on plateau, patience 5, minimum 1e-5). Training stopped after 21 epochs, the best weights came from epoch 11, and the learning rate was halved twice. A fixed seed (42) was used in Python, NumPy, and TensorFlow. The test fold was used only for final reporting; model selection used the validation fold. The training and validation accuracies reported below are the last-epoch values from the training history, whereas the test results and confusion matrices use the restored best-epoch weights. Thirteen functional tests (data ingestion, filtering, re-sampling, forward pass, training, confusion matrix, explanations, model saving) passed, including two full runs with seed 42 that gave identical test accuracy to four decimals. The trained model, inference code, and a web demonstration are available at https://github.com/kavinduabc/Final_Year_Project-.git.

TABLE IV
TRAINING AND REPRODUCIBILITY SETTINGS

| Item | Setting |
|---|---|
| Input | 5000×12 per record (500 Hz, 10 s), PTB-XL records500 |
| Band-pass filter | 3rd-order zero-phase Butterworth, 0.5–40 Hz |
| Normalization | Channel-wise z-score, re-sampled training-set mean/std |
| Splits | strat_fold: 1–8 train, 9 validation, 10 test |
| Re-sampling | Train only; CD 1500, MI/NORM/STTC 2000 |

| | |
|---|---|
| Optimizer | Adam (1e-3); categorical cross-entropy; batch size 32 |
| Epochs | Max 50; stopped at 21; best epoch 11 (restored) |
| Callbacks | Early stopping (patience 10); ReduceLROnPlateau 0.5, patience 5, min 1e-5 |
| Model size | 978,628 parameters (977,732 trainable) |
| Seed / software | 42 (Python, NumPy, TensorFlow); TensorFlow 2/Keras, Colab GPU |

### *B. Evaluation Metrics*

We report overall accuracy on the training, validation, and test partitions, per-class precision, recall, and F1 plus macro- and weighted-average F1 on the test set, and test accuracy for ten rotations of the test fold. Because validation and test sets retain the natural imbalance, macro-averaged metrics are emphasised to avoid overstating performance driven by the majority NORM class. Confusion matrices are reported for the test and training sets. Macro-F1 is the unweighted mean of the per-class F1-scores,

$$\text{Macro-F1} = (1/K)\,\Sigma_k\, 2\cdot P_k\cdot R_k / (P_k + R_k),\ K = 4, \quad (2)$$

where $P_k$ and $R_k$ are the precision and recall of class k; weighted F1 weights each class by its test support.

## V. Results and Discussion

### *A. Overall and Per-Class Performance*

The classifier reached 98.11% training accuracy, 77.30% validation accuracy, and 76.98% test accuracy (Table V). The close agreement of validation and test accuracy (0.32 points) suggests that the validation fold is a reasonable proxy for held-out performance, while the large train–test gap indicates meaningful overfitting (Section V-D).

TABLE V. OVERALL SYSTEM ACCURACY

| Data split | Accuracy |
|---|---|
| Training | 98.11% |
| Validation | 77.30% |
| Test | 76.98% |

Table VI gives per-class test results (1,594 records; weighted F1 = 0.77, macro F1 = 0.71). NORM performed best (F1 = 0.85). Because NORM was under-sampled to 2,000 training records, the same size as MI and STTC, this result is not explained by training frequency and more plausibly reflects the comparatively consistent morphology of normal recordings. CD was moderate (F1 = 0.70), while MI and STTC were hardest (F1 = 0.65 and 0.66), consistent with the clinical overlap between ischemic ST/T-wave changes and myocardial infarction.

TABLE VI. PER-CLASS TEST-SET PERFORMANCE

| Class | Prec. | Rec. | F1 | Supp. |
|---|---|---|---|---|
| CD | 0.69 | 0.70 | 0.70 | 184 |
| MI | 0.66 | 0.64 | 0.65 | 256 |
| NORM | 0.86 | 0.84 | 0.85 | 912 |
| STTC | 0.63 | 0.70 | 0.66 | 242 |
| Macro avg | 0.71 | 0.72 | 0.71 | 1594 |
| Weighted avg | 0.77 | 0.77 | 0.77 | 1594 |

Table VII gives one-vs-rest sensitivity and specificity derived from the test confusion matrix (Fig. 2(a)). Treating NORM versus all abnormal classes as a screening task, 558 of 682 abnormal records (81.8%) are flagged abnormal, while 83.9% of normal records are recognized as normal. With 1,594 test records, the 95% Wilson confidence interval for the test accuracy is 74.8–79.0%, so differences of about two accuracy points between configurations should not be over-interpreted.

TABLE VIII. ONE-VS-REST SENSITIVITY AND SPECIFICITY (TEST FOLD)

| Class | TP | FN | FP | TN | Sens. | Spec. |
|---|---|---|---|---|---|---|
| CD | 129 | 55 | 58 | 1352 | 0.70 | 0.96 |
| MI | 164 | 92 | 84 | 1254 | 0.64 | 0.94 |
| NORM | 765 | 147 | 124 | 558 | 0.84 | 0.82 |
| STTC | 169 | 73 | 101 | 1251 | 0.70 | 0.93 |

**Error analysis.** Of the 55 missed CD records, 32 were labelled NORM and 15 STTC. Of the 92 missed MI records, 53 (20.7% of all MI) were labelled NORM and 30 STTC, so the most clinically costly error, a missed infarction predicted as normal, occurs in about one of five MI records. Of the 73 missed STTC records, 39 were labelled NORM and 26 MI. These misses, together with the 124 abnormal records predicted as NORM, are the natural candidates for inspection with the explanation overlay.

### *B. Ten-Fold Rotation and Ablation*

To check sensitivity to the choice of test fold, the same architecture, hyperparameters, and training procedure were repeated with each of the ten PTB-XL folds as the test fold. Test accuracy ranged from 75.61% (fold 5) to 78.02% (fold 4), with a mean of 76.77% and a standard deviation of 0.75 points; the reported fold-10 value (76.98%) lies within one standard deviation of the mean, so it is not an unusually favourable split and the train–test gap is a consistent property of the procedure.

Table VIII compares three models trained on the same re-sampled training set with the same optimizer, batch size, and callbacks, and evaluated on the same fold-10 test set: a baseline CNN (four Conv1D blocks with 32/64/128/256 filters, kernel sizes 7/5/5/3, ReLU, global average pooling, no recurrent layer), a ReLU CNN-BiLSTM with the same recurrent layers, and the proposed Mish CNN-BiLSTM.

TABLE VIII
SAME-PROTOCOL COMPARISON ON THE FOLD-10 TEST SET

| Model | Params | Acc. (%) | W-F1 | Macro -F1 |
|---|---|---|---|---|
| Baseline CNN (ReLU, no BiLSTM) | 188,004 | 79.17 | 0.79 | 0.74 |
| CNN-BiLSTM (ReLU) | 769,988 | 77.35 | 0.78 | 0.72 |
| CNN-Mish-BiLSTM (proposed) | 978,628 | 76.98 | 0.77 | 0.71 |

Accuracy decreases slightly as complexity grows, from 79.17% to 77.35% to 76.98%, and macro-F1 from 0.74 to 0.72 to 0.71. Neither the BiLSTM layers nor Mish produced a measurable gain; the differences (at most 2.2 accuracy points and 0.03 macro-F1) are of the same order as the fold-to-fold variation above and the confidence interval of Table VII, so they should not be over-interpreted, but the evidence does not support the claim that Mish improves accuracy on this task.

Three caveats apply: the ReLU CNN-BiLSTM also used a kernel size of 5 instead of 10, so its difference to the proposed model is not due to the activation alone; all models were trained once on a single fold; and per-class results of the baselines were not compared, so we cannot say whether minority-class performance differs. The proposed architecture is retained because the explanation layer was built and examined on it, not because it is more accurate.

### *C. Confusion Matrix Analysis*

Fig. 2 shows the test and training confusion matrices. On the test fold, 1,227 of 1,594 records are correct (76.98%). MI is most often confused with NORM (53 of 256 records) and STTC (30), and STTC with NORM (39 of 242). NORM recall is 0.84: 147 of 912 normal records (16.1%) are labelled abnormal, and 124 of the 682 abnormal records (18.2%) are labelled NORM. The model therefore separates normal from abnormal morphology only moderately well, and it struggles most with finer distinctions among overlapping ST/T and ischemic patterns. On the re-sampled training set (Fig. 2(b)), 6,920 of 7,500 records are correct (92.3%, macro F1 = 0.92). This differs from the 98.11% in Table V because Fig. 2(b) re-evaluates the restored epoch-11 weights, whereas Table V reports the last-epoch value of the training history; both show a large gap to validation and test performance.

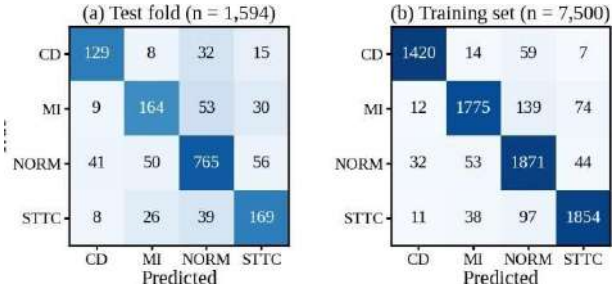



Fig. 2. Confusion matrices (counts; colour = row-normalized rate). (a) Held-out test fold. (b) Re-sampled training set.

### *D. Training Dynamics and Overfitting*

The 98.11% versus 76.98% train–test gap is the most notable limitation. Over-sampled CD and STTC records are exact duplicates, so the classifier may partially memorize them, and the model capacity (three convolutional blocks of up to 256 filters plus two BiLSTM layers) is large relative to 7,500 training sequences, which raises overfitting risk despite0.3 dropout [25]. Under-sampling NORM additionally removes potentially informative variability. A single 256-unit BiLSTM layer did not improve validation accuracy over the two-layer design in an early trial (not quantified here). More stringent stopping, stronger regularization, data augmentation, or class-weighted losses may narrow the gap; these options were not compared, and we do not attribute the gap to a single cause. ROS does not solve class imbalance: it changes class frequency, not the information content of the minority classes, and generalization to CD in particular may remain limited.

### *E. Comparison with Related Work*

Table I places this work among related studies. The closest protocol is that of Strodthoff et al. [2], which uses the same dataset and fold structure and reports AUC-based metrics, roughly 0.93 macro-AUC for its best convolutional models on the diagnostic tasks, on its own label sets. We report accuracy and F1 on a single-label four-class subset that excludes Hypertrophy and multi-label records, so the numbers are not directly comparable. Very high accuracies, such as the mean of about 98.2% reported for a lightweight CNN-LSTM [10], come from arrhythmia-type classification tasks with different labels and data and should not be read against four-class PTB-XL diagnostic labels. Because no published PTB-XL model was retrained under our exact protocol, the baselines of Table VIII provide the controlled comparison: the proposed model does not outperform them in accuracy or macro-F1. A direct comparison with a published PTB-XL model [2], including macro-AUC on the same test fold, remains to be done; macro-AUC was not computed in this study.

### *F. Interpretability*

SHAP and LIME translate classifier confidence into waveform-level explanations. For a correctly classified MI case, a meaningful attribution map should emphasize ST-elevation or pathological Q-waves in the relevant leads, whereas attention on baseline or noise regions would flag the prediction as unreliable even when correct. This matters most for MI/STTC, where errors are most likely [22]. Because no quantitative validation was performed, we present the explanation layer as model attribution and not as evidence of clinical correctness. Because re-sampling changes the class priors seen during training, the predicted probabilities are not calibrated to the natural prevalence of each class and must not be read as risk estimates.

## VI. LIMITATIONS

(1) **Overfitting:** the train–test gap is not resolved by dropout, early stopping, and LR scheduling [25]. (2) **Re-sampling:** the hybrid scheme duplicates CD/STTC samples and discards 5,243 NORM records; it targets class frequency rather than separability and may not expose the model to the full range of MI/STTC presentations [14]. It was not compared with class-weighted losses or synthetic oversampling, and an ablation without re-sampling was not run. (3) **No accuracy gain:** the baseline CNN scored higher than the proposed model, the ablation used one run per model on one-fold, the kernel size differed between the ReLU and Mish CNN-BiLSTM, and no significance test was applied. (4) **Statistical reliability:** a single seed was used; ten-fold rotation was run for the proposed model only. (5) **Explanations:** SHAP and LIME outputs are not quantitatively validated, were examined on a single representative record, use a small SHAP background (200 records) and default LIME settings, and may be unstable for correlated, high-dimensional waveform inputs. (6) **Scope:** the four-class formulation excludes Hypertrophy and multi-label records and uses a single database, so generalization to other populations, devices, and label sets is untested [17], [2]. (7) **Labels and calibration:** super-class labels come from aggregated SCP-ECG statements and may contain annotation noise, the single-label subset is not representative of all PTB-XL records, and probability calibration after re-sampling was not evaluated. (8) **Reporting:** training and validation accuracies are last-epoch values, whereas test results use restored best-epoch weights. (9) **Clinical readiness:** accuracy

of about 77% and the MI/STTC confusion indicate that the system is not suitable for clinical use without further validation.

## VII. Conclusion and Future Work

This paper studied an explainable CNN-BiLSTM framework with Mish activation, train-only random re-sampling, and a SHAP/LIME layer for four-class ECG classification on PTB-XL. The framework achieved 76.98% test accuracy, a weighted F1 of 0.77, and a macro F1 of 0.71 (76.77 ± 0.75% over ten test-fold rotations), performing best on NORM and worst on MI/STTC. A simpler CNN baseline scored higher, so the contribution is not accuracy: it is a transparent, leakage-controlled pipeline, an honest ablation, and an explanation layer whose clinical validity remains to be shown. Future work will (1) compare re-sampling with class-weighted losses and synthetic oversampling [14]; (2) repeat the ablation with matched kernel sizes, multiple seeds, and a no-re-sampling variant; (3) report macro-AUC and a retrained published PTB-XL baseline [2]; (4) quantitatively validate SHAP/LIME maps with the protocol of Section III-G, possibly adding Grad-CAM [23]; and (5) extend the system to the full PTB-XL label set and to additional public databases [19], [27].

## Acknowledgment

The author thanks the supervisor, Ms. Hiruni Weerasinghe, for guidance throughout this research, the Department of Computer Science and Data Science, NSBM Green University, for academic and computational support, and the PhysioNet PTB-XL contributors for making the dataset publicly available.

## References

[1] S. Moazemi et al., “Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review,” Front. Med., 2023, doi: 10.3389/fmed.2023.1109411.

[2] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, “Deep learning for ECG analysis: Benchmarks and insights from PTB-XL,” IEEE J. Biomed. Health Inform., vol. 25, no. 5, pp. 1519–1528, May 2021, doi: 10.1109/JBHI.2020.3022989.

[3] A. Y. Hannun et al., “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” Nat. Med., vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.

[4] Q. Xiao et al., “Deep learning-based ECG arrhythmia classification: A systematic review,” Appl. Sci., vol. 13, no. 8, 2023, doi: 10.3390/app13084964.

[5] G. Petmezas et al., “State-of-the-art deep learning methods on electrocardiogram data: Systematic review,” JMIR Med. Inform., Aug. 2022, doi: 10.2196/38454.

[6] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, “Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network,” Inf. Sci., vol. 405, pp. 81–90, Sep. 2017, doi: 10.1016/j.ins.2017.04.012.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[8] Y. Ye, K. Chipusu, M. A. Ashraf, B. Ding, Y. Huang, and J. Huang, “Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG signals,” Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-17671-1.

[9] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, “A lightweight hybrid CNN-LSTM model for ECG-based arrhythmia detection,” Aug. 2022. [Online]. Available: http://arxiv.org/abs/2209.00988

[10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017. [Online]. Available: https://github.com/slundberg/shap

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[12] C. Molnar, G. Casalicchio, and B. Bischl, “Interpretable machine learning—A brief history, state-of-the-art and challenges,” Oct. 2020, doi: 10.1007/978-3-030-65965-3_28.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[14] K. Pałczyński, S. Śmigiel, D. Ledziński, and S. Bujnowski, “Study of the few-shot learning for ECG classification based on the PTB-XL dataset,” Sensors, vol. 22, no. 3, Feb. 2022, doi: 10.3390/s22030904.

[15] S. F. Ahmed et al., “Deep learning modelling techniques: Current progress, applications, advantages, and challenges,” Artif. Intell. Rev., vol. 56, no. 11, pp. 13521–13617, Nov. 2023, doi: 10.1007/s10462-023-10466-8.

[16] P. Wagner et al., “PTB-XL, a large publicly available electrocardiography dataset,” Sci. Data, vol. 7, no. 1, Dec. 2020, doi: 10.1038/s41597-020-0495-6.

[17] M. K. Gourisaria, G. M. Harshvardhan, R. Agrawal, S. S. Patra, S. S. Rautaray, and M. Pandey, “Arrhythmia detection using deep belief network extracted features from ECG signals,” Int. J. E-Health Med. Commun., vol. 12, no. 6, Nov. 2021, doi: 10.4018/IJEHMC.20211101.oa9.

[18] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. e215–e220, 2000. [Online]. Available: http://www.physionet.org

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Dec. 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[20] A. Rahi, “Fusion-based deep learning ensemble on MIT-BIH and PTB-XL ECG databases for enhanced cardiac diagnosis,” Aug. 2025, doi: 10.1101/2025.08.20.25334079.

[21] J. Luo, A. Taleban, P. Noffke, and R. Sparapani, “Multi-level explainable AI for ECG-based atrial fibrillation detection: Exploring LIME, SHAP, and Grad-CAM for clinical interpretation,” in Proc. ICMHI 2025, Dec. 2025, pp. 52–59, doi: 10.1145/3761712.3761721.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” Int. J. Comput. Vis., 2019, doi: 10.1007/s11263-019-01228-7.

[23] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. ICML, 2015.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,”J. Mach. Learn. Res., vol. 15, pp. 1929–1958, 2014.

[25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Jan. 2017. [Online]. Available: http://arxiv.org/abs/1412.6980

[26] L. M. Casas-Quiroz, P. San Juan Bautista-UPSJB, A. J. A. Lavalle, and A. Mugruza-Vassallo, “Integrated ECG algorithms for QRS detection, arrhythmia identification and portability”.

# Gated but Not Explained: A Faithfulness Audit of Intrinsic Explanations in Multi-Agent Reinforcement Learning

S. H. G. G. Sewjith
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
hggsewjith@students.nsbm.ac.lk

T. H. Wickramage
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
thwickramage@students.nsbm.ac.lk

I. K. D. Vilochana
*Department of Software Engineering*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
ikdvilochana@students.nsbm.ac.lk

***Abstract*—Gated message-passing policies expose sparse communication graphs that are sometimes interpreted as explanations of agent decisions. This interpretation requires more than low gate density: a selected edge should be relevant to the trained policy, and agreement with a task-defined partner graph need not measure such relevance. We examine this distinction through a reported audit of GraphXAgent, a gated, attention-based multi-agent policy, against intervention- and gradient-based attribution methods. The supplied experiments cover multi-agent particle-environment simple_spread, level-based foraging, and a synthetic pointer-routing task, with five reported training seeds. On pointer routing, fused intrinsic scores achieve an alignment AUROC of 1.000, but attention alone also reaches 1.000 while gates alone remain near chance. On simple_spread, the intrinsic method has the lowest reported perturbation-based Fidelity+ among the compared explainers. In foraging, the intrinsic alignment AUROC is 0.270 against environment-defined partner labels; this is evidence of disagreement with those labels, not by itself proof of causal unfaithfulness. We formalize the policy–task distinction, report measured comparisons without extrapolating unsupported significance, and identify the interventions and reproducibility information needed for a conclusive causal audit.**

***Keywords—multi-agent reinforcement learning, explainability, communication gates, graph attention, faithfulness, counterfactual interventions***

## I. Introduction

Cooperative multi-agent reinforcement learning (MARL) often improves task performance by allowing agents to exchange messages during decentralized execution [1, 2, 3]. Sparse communication is attractive because it exposes a directed interaction graph, potentially offering a human-readable account of which agent influenced which decision. A central difficulty is that message selection, policy dependence, and environmental cooperation are different properties. A gate describes what a network permits to pass; an attribution score describes the apparent contribution of a signal; and a task-defined collaboration relation describes which agents could jointly satisfy an environmental rule. These relations may disagree without logical contradiction.

We audit the explanation furnished by GraphXAgent, a representative hard-concrete-gated, graph-attention communication policy. The study is deliberately diagnostic rather than a claim of a new generally faithful explainer. Its distinguishing contribution is to juxtapose explanatory alignment, same-policy perturbation sensitivity, gate sparsity, stability, and computation time while examining whether the evaluation labels really describe policy-used dependencies. The original experiment record supports three limited observations. First, graph-communication policies obtain higher reported episodic returns than message-free policies on larger simple_spread instances, but this comparison does not alone isolate a causal effect of communication. Second, on synthetic pointer routing, the reported alignment is perfect for attention-derived rankings, whereas raw gates contribute little discriminative information. Third, the foraging labels and a trained policy's message dependence are not interchangeable.

Our contributions are: (i) a precise separation of model-internal scores, edge-intervention effects, and environment-defined partner labels; (ii) an organized audit of the numerical results actually available from three tasks, including negative results; (iii) an explicit comparison of gate and attention attributions, with the sparsity/utility failure boundary; and (iv) a replicable evaluation specification and a candid accounting of information missing from the experiment record. No new training run, confidence interval, or baseline measurement is asserted in this revision.

## II. Background and Related Work

### *A. Cooperative policies and communication*

MAPPO adopts centralized training with decentralized policies [2]; the EPyMARL benchmark provides a broader shared implementation and evaluation context [3]. CommNet averages learned messages [4], TarMAC targets recipients through learned communication [5], and graph-based multi-agent methods use structured neighborhood aggregation [6]. QMIX is a value-factorization alternative rather than an equivalent communication control [7]; it is therefore contextual related work, not a matched message-passing baseline in the present record. A recent survey emphasizes that communication patterns, recipient selection, and communication costs should be distinguished [8].

*B. Interpretation and faithful evaluation*

Graph attention networks [9] and GATv2 [10] provide normalized, input-dependent edge weights; hard-concrete gates support differentiable $L_0$ sparsity penalties [11]. GNNExplainer learns explanatory subgraph masks after training [12]. More generally, attention weights are not automatically causal explanations [13], while the usefulness of attention as explanation depends on the tests and explanation definition [14]. This motivates complementary model-intervention and label-alignment measures rather than treating sparse gates, attention, or occlusion as a universal gold standard.

## III. METHODOLOGY

*A. Message passing and intrinsic score*

Let $N$ agents form a directed candidate graph $\mathcal{E} = \{(j, i): i \neq j\}$. Agent $i$ observes $o_i$, encodes $h_i = \phi_\theta(o_i)$, and receives learned messages from candidate senders $j$. A generic gated-attention layer can be expressed as

$$m_i = \sum_{j \neq i} g_{ij}\, \alpha_{ij}\, W h_j, \qquad \alpha_{ij} \geq 0, \quad \sum_{j \neq i} \alpha_{ij} = 1,$$

where $g_{ij} \in [0,1]$ is a relaxed communication gate and $\alpha_{ij}$ is a normalized attention coefficient. The agent's policy is $\pi_\theta(a_i \mid o_i, m_i)$; a centralized value function supports training. The studied design uses an observation encoder, two message-passing layers, and feedforward action/value heads; recurrent memory is not used in the reported experiments. Its intrinsic edge-ranking score is

$$s_{ij}^{\text{fused}} = g_{ij}\alpha_{ij}, \quad s_{ij}^{\text{gate}} = g_{ij}, \quad s_{ij}^{\text{attn}} = \alpha_{ij}.$$

These scores are structural proxies; neither is, without intervention, a causal measure of an agent's influence. In particular, a zero gate establishes that the corresponding direct contribution is blocked in this layer, but does not establish whether a sender affects the receiver indirectly through other paths or past observations.

The manuscript describes a hard-concrete $L_0$-style gate [11]. A standard parameterization, included here to make the model family explicit rather than to claim unreported implementation constants, draws $u \sim U(0,1)$ and sets

$$\tilde{g}_{ij} = \sigma\left(\frac{\log u - \log(1-u) + \ell_{ij}}{\beta}\right),$$
$$g_{ij} = \text{clip}_{[0,1]}((\zeta - \gamma)\tilde{g}_{ij} + \gamma),$$

where $\ell_{ij}$ is a learned logit, $\beta > 0$ is a temperature, $\gamma < 0$, and $\zeta > 1$. In this family the expected active-gate probability is . The numerical values of $\beta$, $\gamma$, $\zeta$, logit initialization, per-layer normalization order, and exact deterministic test-time substitution are not recoverable from the supplied manuscript; the hard-concrete formula above is therefore a mathematical reference specification, not a verified code transcription.

The stated optimization uses a MAPPO-style objective augmented with a gate penalty:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} + c_V \mathcal{L}_V - c_H \mathcal{H}(\pi) + \lambda_t \sum_{(j,i) \in \mathcal{E}} \Pr(g_{ij} > 0),$$

where the reported schedule linearly increases $\lambda_t$ from zero to its final value during the initial 20% of training. Value normalization is reported as enabled. Coefficients, hidden sizes, gate parameters, optimizer, PPO clipping settings, rollout length, and random seed identities have not been supplied and must be disclosed from the training configuration before independent replication is possible.

*B. The three distinct "ground truths"*

For a receiver $i$, define a same-policy message intervention $do(m_{ij} = 0)$ while holding network weights, observations, and all other messages fixed. A directly meaningful policy-dependence quantity is

$$d_{ij}(x) = D_{\text{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x; do(m_{ij} = 0))).$$

This describes dependence of the policy's action distribution on a message for a sampled state $x$ ; it is not necessarily a difference in task return. For multiple-layer communication, a valid intervention must declare the targeted layer, whether downstream messages are recomputed, and whether recurrent state is reset. None of these choices is immaterial.

By contrast, synthetic pointer-routing labels identify a task-defined parent link $y_{ij}^{\text{task}} \in \{0,1\}$. This is an appropriate alignment target when the agent's answer is constructed to depend on the designated parent and observations cannot bypass that dependency. For foraging, an environment-defined partner relation $y_{ij}^{\text{env}}$ instead reflects the collection rule or designated eligible teammates. Such a label is not by itself the graph of dependencies used by a trained policy. Equal group and no-message returns do not prove that each agent can independently complete the task: agents may still coordinate via observations, synchronized behavior, or the environment. Consequently, low AUROC against $y^{\text{env}}$ indicates label disagreement, not a standalone causal unfaithfulness proof.

## IV. EXPERIMENTAL SETUP

*A. Explanation ranking and interventions*

An audit should evaluate every explainer on an identical frozen policy and identical held-out states. The compared sources in the submitted record are intrinsic fused score, edge occlusion, gradient times input, and optimized graph mask. Let $s_{ij}(x)$ be the score assigned to a directed edge; sorting within a receiver yields a top-$k$ set $S_k(x)$. For a scalar decision score $q_i(x)$ (for example the probability of the original greedy action $a_i^* = \text{argmax}_a \pi_i(a \mid x)$ ), define two complementary intervention quantities:

$$\text{Di(S)} = \text{Ex}[\,|\text{qi(x)} - \text{qi(x; do(mS = 0))}|\,]$$
$$\text{F+(k)} = \text{Di(Sk)}, \quad \text{F−(k)} = \text{Di(E \ Sk)}.$$

Here $F^+$ measures removal of the purportedly important links (a larger positive drop supports necessity), while $F^-$ measures how much the decision changes after preserving only those links (smaller supports sufficiency). $q_i$ must be fixed before evaluating methods, and $k$ or its edge-budget fraction must be matched across methods. In addition, the policy-intervention measure above provides a distribution-level

intervention test. These equations are an explicit recommended reproducibility contract. The original tables contain a quantity labeled “Fidelity+” but do not record its score function, edge budget, normalization, or aggregation. Therefore, the historical numerical values below are labeled reported Fidelity+; it is not established that they were calculated by the fidelity expressions above. They must not be presented as newly computed estimates under it.

For alignment, $\mathrm{AUROC}(s,\ y)$ ranks positive task-defined edges above negative edges. Its random-ranking expectation is $0.5$ when both label classes are present; a value below $0.5$ is worse than random ranking against those labels, not a proof of malicious or noncausal operation. A direct random-ranking control would sample independently of the labels and average across repeated permutations; such a control has not been run in the provided record. Gate activity is $\frac{1}{|\mathcal{E}|}\sum_{ij} 1[g_{ij} > \tau]$ for an explicitly declared threshold $\tau$ ; if instead a mean open probability is reported, it must be named separately. The source alternates “active-gate rate” and “gate probability” without establishing the threshold or aggregation, so those legacy names are retained in the result table. Stability requires both a perturbation distribution $\delta$ and a rank similarity statistic $\rho(s(x),\ s(x+\delta))$; approximately 0.99 similarity is reported without these specifications.

### B. Environments and comparison scope

The source evaluates three distinct questions. (i) simple_spread, with $N = 3,6,10$ , examines return under increasing numbers of agents; for $N = 6,10$ it also reports intervention-based explanation comparisons. (ii) Level-based foraging 5x5-4p-1f-coop has four players and one food item; the source describes a collection requirement involving two players, but does not supply the food level, player levels, episode randomization, exact ground-truth label-construction function, or independent label validation. (iii) Pointer routing uses four to six agents with an identity/tag-defined parent, selected to avoid positional aliasing. The selected foraging configuration is small enough for the reported experiments to finish; it is not a general test of all level-based foraging variants.

The stated policy-comparison set is IPPO, MAPPO without execution-time communication, CommNet, TarMAC-like targeted communication, DGN-style attention-based graph communication, and GraphXAgent. Five training seeds are stated. However, the supplied result tables omit several of these baseline values, and policy size matching is asserted without layer widths or parameter counts. Only the actually recorded comparisons appear in this revised results section. A trained-policy explanation comparison and an independently trained policy comparison answer different questions; for a causal claim about explanation quality, evaluation should freeze the same policy for all methods.

### C. Statistical and computational reporting

The reported simple_spread returns are interquartile means (IQMs) of five runs, with a 150,000-step budget. The source figures refer to 95% confidence intervals but no seed-level traces, bootstrap replicates, or numerical interval endpoints are provided. We consequently do not infer $p$-values, nonoverlap, confidence bounds, or significance from bar heights. With raw runs, a defensible analysis would report per-seed returns, an explicitly specified stratified bootstrap interval for IQM [15], and paired bootstrap differences when methods share environment seeds. The same requirement applies to AUROC, intervention metrics, cost, and gate activity. Timing must specify hardware, batch size, warm-up, device synchronization, and whether time is per decision, per receiver, or per graph; explanation generation time and training-step time must not be conflated.

## V. RESULTS

### A. Performance as the agent count increases

Table 1 and Fig. 1 reproduce the numerical returns available in the original record. Higher (less negative) return is better. At three agents the methods are close. At six and ten agents the graph methods have better reported IQMs than independent policy learning; GraphXAgentexceeds IPPO by 21.3 and 41.3 return units, respectively, and exceeds MAPPO by 11.9 and 25.9. Because seedwise variance and matched-policy ablations are absent, this is descriptive evidence of a performance association with communication architectures in the tested setting, not a statistically established causal necessity result.

TABLE I
MPE SIMPLE_SPREAD: REPORTED IQM EPISODIC RETURN (FIVE SEEDS; 150K TRAINING STEPS; LARGER IS BETTER). NUMERICAL CONFIDENCE LIMITS WERE NOT SUPPLIED.

| *N* | IPPO | MAPPO | DGN | GRAPHXAGENT |
|---|---|---|---|---|
| 3 | −64.7 | −66.4 | −63.5 | −66.4 |
| 6 | −238.5 | −229.1 | −222.1 | −217.2 |
| 10 | −577.1 | −561.7 | −540.8 | −535.8 |

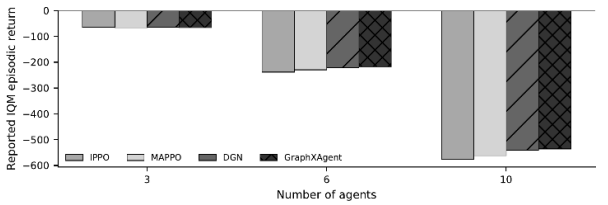



Fig. 1. Reported performance from Table 1. Bars reflect IQMs; uncertainty bars are intentionally omitted because the source does not provide verifiable interval values.

### B. Pointer routing: task-label alignment and cost

Table 2 compares four attribution methods on the synthetic pointer-routing task at $\lambda = 0$. The fused intrinsic score reaches AUROC 1.000 against constructed parent-edge labels, compared with 0.999 for occlusion, 0.752 for gradient-times-input, and 0.990 for the learned mask. The reported average time per explanation is 2.0, 66.5, 4.4, and 429.4 ms respectively. These timings support a within-record inference-time advantage but do not establish universal speedups across hardware or agent counts; they do not include policy-training cost. Occlusion is an intervention-derived comparator, not a general-purpose ground-truth explainer.

### C. Attention, gates, and the sparsity boundary

The gate/attention decomposition is the most informative model-internal control. As shown in Table 3, attention-only AUROC is 1.000 at $\lambda = 0$, $10^{-3}$, and $10^{-2}$, while gate-only AUROC is 0.63, 0.53, and 0.62, respectively. At $\lambda \geq 0.1$ the

recorded gate activity is approximately zero and the alignment AUROC falls to chance. Therefore, the synthetic alignment finding is better attributed to the attention ranking than to the gates alone. Gate sparsity cannot be equated with explanation faithfulness.

The original text also reports approximately 0.99 ranking stability under small perturbations. This is a useful preliminary observation, but its perturbation distribution, similarity formula, aggregation, and dispersion remain unspecified. It is therefore not an independent robustness guarantee. In particular, highly stable explanations may consistently rank the wrong links.

TABLE II
SYNTHETIC POINTER ROUTING: HISTORICAL REPORTED VALUES (FIVE SEEDS, $\lambda = 0$). FIDELITY+ USES THE ORIGINAL MANUSCRIPT'S UNSPECIFIED IMPLEMENTATION.

| Explainer | Align. AUROC | Fid.+ (rep.) | Time (ms) |
|---|---|---|---|
| Intrinsic $g\alpha$ | 1.000 | 0.764 | 2.0 |
| Edge occlusion | 0.999 | 0.769 | 66.5 |
| Gradient × input | 0.752 | 0.456 | 4.4 |
| Optimized mask | 0.990 | 0.753 | 429.4 |

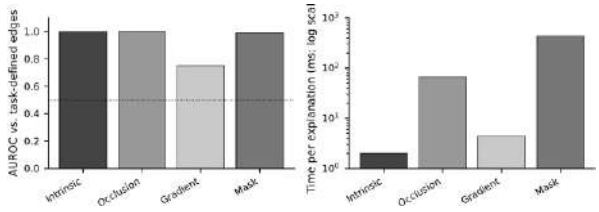



Fig. 2. Synthetic task, using numbers from Table 2: task-label alignment AUROC (left) and recorded inference-time cost (right). No confidence intervals can be reconstructed.

TABLE III
SYNTHETIC-TASK SPARSITY SWEEP (REPORTED FIVE-SEED SUMMARY). "ACTIVITY" RETAINS THE SOURCE'S UNDEFINED GATE-RATE CONVENTION.

| $\lambda$ | Accuracy | Fused AUROC | Gate AUROC | Activity |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 0.63 | 0.98 |
| $10^{-4}$ | 1.00 | 1.00 | 0.54 | 0.47 |
| $10^{-3}$ | 1.00 | 1.00 | 0.53 | 0.44 |
| $10^{-2}$ | 1.00 | 1.00 | 0.62 | 0.37 |
| 0.1 | 0.26 | 0.50 | 0.50 | $\sim 0$ |
| 1.0 | 0.26 | 0.50 | n/a | $\sim 0$ |

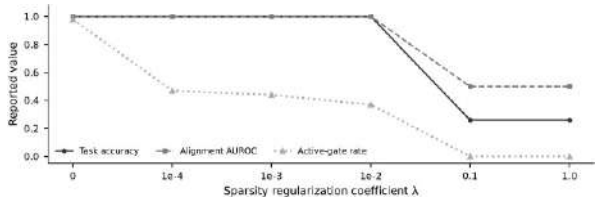



Fig. 3. Recorded sparsity sweep on pointer routing. The operating region ends before $\lambda = 0.1$, at which task accuracy decreases sharply; performance at $\lambda = 1$ is also low.

### *D. Foraging: disagreement with environmental partner labels*

On the reported four-player foraging configuration, Table 4 gives intrinsic AUROC 0.270, versus 0.629 for occlusion and 0.648 for gradient-times-input, when compared with environment-defined collaboration labels. The intrinsic method is below the 0.5 random-ranking expectation for this alignment problem. The underlying policy reportedly performs similarly with or without explicit messaging. That observation motivates a hypothesis that communication channels are unused or redundant on the sampled states; it does not demonstrate that independent agents can complete the underlying collection action without cooperation.

There is an additional measurement risk: the environmental rule may identify an eligible pair or coalition, whereas a policy can act using local observations rather than another agent's explicit message. Thus the label graph is a proposed task relation, not a validated causal ground truth for the policy-intervention definition above. A conclusive interpretation would require randomized message deletion/replacement on the identical trained policy and state set, plus a test that each labeled partner is actually necessary for the chosen action under relevant counterfactual states.

TABLE IV
LEVEL-BASED FORAGING 5X5-4P-1F-COOP: ALIGNMENT WITH ENVIRONMENT-DEFINED LABELS (FIVE REPORTED SEEDS).

| Explainer | Align. AUROC | Fid.+ (rep.) | Time (ms) |
|---|---|---|---|
| Intrinsic | 0.270 | 0.205 | 1.4 |
| Occlusion | 0.629 | 0.257 | 20.0 |
| Gradient × input | 0.648 | 0.129 | 3.2 |

### *E. simple_spread: perturbation faithfulness and cost*

For policies evaluated at $\lambda = 0$, the reported Fidelity+ values in Table 5 favor occlusion over the intrinsic method for six and ten agents. At $N = 6$, the intrinsic value (0.076) is less than half the occlusion value (0.156); at $N = 10$, the gap is smaller (0.092 versus 0.116). Fig. 4 displays these values. Since the recorded normalization and explanation budget are not specified, the conservative conclusion is that the source's intervention-derived score ranks the intrinsic explanation last on these evaluated checkpoints, not that its exact causal importance error or statistical significance is known.

TABLE V
SIMPLE_SPREAD: REPORTED FIDELITY+ ON POLICIES AT $\lambda = 0$. DEFINITION AND EDGE-BUDGET SETTINGS ARE ABSENT IN THE ORIGINAL RECORD.

| $N$ | Occlusion | Gradient × input | Intrinsic |
|---|---|---|---|
| 6 | 0.156 | 0.134 | 0.076 |
| 10 | 0.116 | 0.102 | 0.092 |

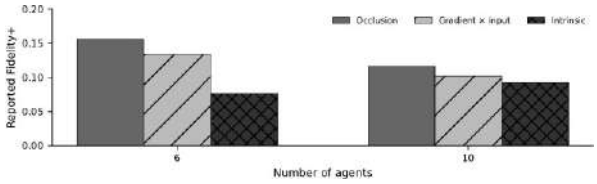



Fig. 4. Comparison of original-manuscript Fidelity+ values on simple_spread checkpoints. These are descriptive values without recoverable uncertainty intervals.

The source's training-cost plot also shows higher GraphXAgent time per training step than a CommNet comparison at larger agent counts. A gate that sets a dense edge contribution to zero does not, by itself, prune the dense

computation; therefore sparse learned weights need not imply faster training or rollout. The recorded figure lacks exact training times, hardware, and benchmark protocol, so no numeric training speed ratio is inferred here. Likewise, the claim of approximately eight-fold synthetic occlusion speedup follows from 66.5/2.0≈33.25 for the displayed time entries, not eight-fold. The corrected ratio from Table 2 is ≈33.3× relative to occlusion and ≈214.7× relative to optimized masking, in the specific reported timing setup. In foraging the corresponding recorded ratio is ; these ratios must not be confused with training speed.

## VI. DISCUSSION: THE EVALUATION GAP

The synthetic task provides a controlled positive case for task-label alignment, subject to the task's ability to prevent observational shortcuts. It shows that an intrinsic fused score can closely match a constructed parent relation at low recorded inference cost. Because attention alone attains the same reported AUROC, it does not validate gate weights as individually meaningful explanations. Simple_spread presents a contrasting performance/attribution pattern: communication-capable graph architectures score better on larger tasks, while their reported intrinsic explanation ranks below the two intervention/gradient comparisons. The foraging result raises a different issue: a low AUROC relative to an environmentally designated cooperation graph does not directly identify a policy-explanation error.

The results do not prove that no realistic environment can provide jointly necessary and observable partners, that sparse intrinsic explanations are universally unfaithful, or that the chosen foraging task is "individually solvable." These would require stronger controls than those furnished. The strongest supportable statement is that the tested intrinsic scores are not a validated proxy for policy-used cooperative dependencies across all three reported settings. Conversely, a structural zero gate genuinely blocks its direct message in the declared layer, but this limited mechanism-level property should not be conflated with a faithful attribution to behavioral decisions.

### *A. Controls necessary to close the gap*

The most informative missing experiment holds one trained policy and the evaluation states fixed, then (1) removes or replaces individual messages; (2) compares top- $k$ ranking against random rankings at the same $k$ ; (3) measures both deletion-based necessity and retention-based sufficiency using the same action-score definition; and (4) repeats with group deletions to identify redundant or synergistic communications. For foraging, the study should independently validate partner labels by enumerating physically eligible collections under controlled states, and then report separately whether those partners matter to message-conditioned actions. Performance without explicit messages must be compared with performance under a more stringent decentralized-observation or controlled-agent-removal intervention before claiming independent task solvability.

A meaningful baseline suite should additionally include ungated GATv2/attention-only edge rankings and a uniformly random edge ranking, both absent from the fully tabulated source results, plus shared-checkpoint occlusion, gradient-times-input, and optimized-mask explanations wherever feasible. This also avoids a fairness problem in which independently trained policy baselines are used to judge the explanation fidelity of different underlying decision functions.

## VII. LIMITATIONS

### *A. What can be reproduced from this paper*

The environment names, agent-count sweep, 150k-step simple_spread training budget, five reported seeds, broad model composition, the 20% gate-penalty warm-up, and the numbers in Tables 1–5 are present in the source material. The equations above specify the intended distinctions and a prospective, testable audit interface. These facts allow a reader to recreate the plotted summaries but not to reproduce training or verify the historical metric computation.

### *B. Material missing information*

The supplied source contains neither a code repository nor model checkpoints, training logs, per-seed returns, evaluation state identifiers, or concrete network/trainer hyperparameters. It does not give the gate-distribution constants, parameter counts, environment/software versions, per-environment budgets beyond the 150k-step simple_spread setting, numbers of evaluation episodes, exact pointer-task generation code, label rules, perturbation distributions, AUROC aggregation details, or a metric-function definition that can be reconciled with the published "Fidelity+" values. The source also omits complete returns for CommNet and TarMAC-like baselines and does not contain the requested random-ranking intervention. We neither reconstruct these values from plotted pixels nor invent them.

For a verifiable release, the authors must archive configuration files, dependency lockfiles, policy checkpoints, seeds, environment setup, evaluation state samples, label-generation scripts, raw results, plotting scripts, and an executable command to rerun the audit. A persistent artifact URL cannot be truthfully supplied until such an archive exists. Reporting the absence is preferable to inserting an invalid code link or an unsupported promise of reproducibility.

### *C. Generalization and statistical limits*

Only three task families and five claimed runs are represented; the foraging environment is one small configuration and pointer routing is synthetic. The record does not cover graph-topology shifts, coalition size variation, larger foraging configurations, independent task-removal controls, or uncertainty intervals that can be regenerated from raw seeds. Furthermore, edge-level AUROC is unstable or undefined when a test state has only one label class; the valid-state filter and macro/micro aggregation must be specified. Optimizing gates for sparse communication can also change the trained policy, meaning a cross- $\lambda$ comparison is not a pure explanation intervention. Any population-level or causal conclusion is therefore conditional on the tested policy family, checkpoints, labels, and states.

## VIII. CONCLUSION AND FUTURE WORK

Intrinsic edge scores can be cheap and visually sparse without being established explanations of a cooperative policy. In the available experiments, the synthetic pointer-routing task exhibits perfect parent-edge alignment, predominantly

attributable to attention rather than gates; simple_spread shows comparatively low recorded intrinsic intervention scores; and foraging shows poor alignment with environment-defined partner labels that are not independently validated as policy-used dependencies. The appropriate conclusion is a restricted negative audit finding, not a universal impossibility claim. A stronger camera-ready evidence package requires frozen-policy message interventions, validated partner labels, random-ranking and attention-only controls, complete baseline measurements, numerical uncertainty estimates, and an executable artifact with exact configurations.

## REFERENCES


[1] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in Proc. NeurIPS, 2017.

[2] C. Yu et al., "The surprising effectiveness of PPO in cooperative multi-agent games," in Proc. NeurIPS, 2022.

[3] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht, "Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks," in Proc. NeurIPS Datasets and Benchmarks, 2021.

[4] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in Proc. NeurIPS, 2016.

[5] A. Das et al., "TarMAC: Targeted multi-agent communication," in Proc. ICML, pp. 1538–1546, 2019. [6] J. Jiang, C. Dun, T. Huang, and Z. Lu, "Graph convolutional reinforcement learning," in Proc. ICLR, 2020.

[7] T. Rashid, M. Samvelyan, C. Schroeder de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in Proc. ICML, pp. 4295–4304, 2018.

[8] C. Zhu, M. Dastani, and S. Wang, "A survey of multi-agent deep reinforcement learning with communication," Auton. Agents Multi-Agent Syst., vol. 38, art. 4, 2024.

[9] P. Veličković et al., "Graph attention networks," in Proc. ICLR, 2018.

[10] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in Proc. ICLR, 2022.

[11] C. Louizos, M. Welling, and D. P. Kingma, "Learning sparse neural networks through L0 regularization," in Proc. ICLR, 2018.

[12] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in Proc. NeurIPS, 2019.

[13] S. Jain and B. C. Wallace, "Attention is not explanation," in Proc. NAACL-HLT, pp. 3543–3556, 2019.

[14] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in Proc. EMNLP-IJCNLP, pp. 11–20, 2019.

[15] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in Proc. NeurIPS, 2021.

# A Secure AI-Assisted Publishing Architecture for Institutional Websites that Require Controlled Content Governance and Low-Frequency Updates

K. K. L. B. Adikaram
*Faculty of Agriculture*
*University of Ruhuna*
*Kamburupitiya, Sri Lanka*
lasantha@agricc.ruh.ac.lk

P. A. Jayantha
*Faculty of Science*
*University of Ruhuna*
*Wellamadama, Sri Lanka*
jayantha@maths.ruh.ac.lk

***Abstract*—Institutional websites commonly use dynamic content management systems (CMSs), even when most content changes infrequently. Such deployments may expose a public-facing application runtime and database interface that require ongoing patching and security maintenance. This paper presents an AI-assisted, email-driven publishing architecture for institutional websites. Authorized users submit natural-language update requests by email. The system checks the request subject and sender, applies page-level permissions, interprets the requested change, and produces a structured action representation before modifying the content repository. Field validation, template validation, and preview-based checking are implemented in the publishing workflow. Predefined templates are used to regenerate the affected static HTML, CSS, and JavaScript files. A dependency mechanism identifies related pages that also require regeneration. The publishing application and generation workflow run on a local machine. OpenCode, running in that environment, connects over the internet to an externally hosted AI model service for model inference. Generated static files are transferred from the local environment to the production web server through SSH. The production site does not execute application-side PHP or query the content database at request time. The framework has been implemented at the Faculty of Agriculture, University of Ruhuna. Operational examples demonstrate successful profile updates, rejection of requests that refer to another staff member, and informative rejection of messages without recognized update directives. The paper contributes an integrated, policy-constrained publishing workflow that separates natural-language intent interpretation from deterministic content updates and static deployment.**

***Keywords—AI assisted publishing, email-based content management, static site generation, institutional websites, secure web architecture***

## I. Introduction

Institutional websites have become an essential platform for disseminating academic, administrative, and public information. Most academic organizations and government organizations rely on traditional Content Management Systems (CMSs) such as WordPress, Drupal, or Joomla [1]. These systems dynamically generate web pages through server-side scripting languages (e.g., PHP) while retrieving information from backend databases for every user request [2], even every time the same page with the same information will be generated (because the content in most pages is not frequently updating). Furthermore, this approach heavily demands third-party plugins to achieve attractive features while reducing direct coding effort.

Although these dynamic systems offer flexible content management, they expose several inherent limitations. Because the production server maintains active connections with databases and executes server-side scripts for each request, it becomes highly vulnerable to SQL injection, code injection, cross-site scripting (XSS), unauthorized database access, and plugin-based cybersecurity threats [3], [4]. Furthermore, rendering dynamic pages introduces computational overhead, making it difficult to instantly preview the exact appearance of published pages before deployment. In many institutional websites where content changes infrequently, these dynamic mechanisms create unnecessary complexity without providing proportional benefits [5].

A recent incident illustrates the continuing risk of exposing server-side runtimes on public websites. On 21 October 2026, the PHP development team disclosed two critical vulnerabilities, CVE-2026-109618 and CVE-2026-109619, which could permit unauthenticated remote code execution on affected PHP installations serving web requests. The PHP team issued security updates and urged users to update; the European Union Agency for Cybersecurity and partner authorities also advised organisations to identify affected PHP-based systems, prioritise internet-facing services, and apply the fixes or isolate systems where immediate remediation was not possible [6], [7]. This incident is relevant to institutional websites that rely on PHP-based CMS deployments, although it should not be interpreted as evidence that every PHP-based website was compromised.

Decoupled publishing and static-site generation separate content preparation from public delivery. Pre-generated files can be served without querying a content database or executing application code for each page request. This reduces the production server's runtime responsibilities, although it does not totally prevent unsafe content from being embedded in generated pages. Static output must still be validated, and the publishing pipeline must protect its source data, credentials, and deployment channel.

Recent advancements in decoupled architectures and Static Site Generators (SSGs) have attempted to address these issues by separating the frontend from the backend [6]. However, these solutions still require users to interact with complex

administrative interfaces or markdown files, which can be a barrier for non-technical institutional staff. Meanwhile, Artificial Intelligence (AI) and Large Language Models (LLMs) have shown immense promise in code generation and natural language understanding [8]. Yet, most AI web generators act merely as coding assistants rather than governed, secure publishing agents.

To address these gaps, this paper proposes an AI-assisted autonomous web publishing framework. This architecture enables authorized users to update content through natural-language email requests, sender and page-level authorization, constrained intent interpretation, template-driven generation, dependency-aware regeneration, SSH deployment, and requester notification. Though this proposed system is not presented as a replacement for highly dynamic transactional applications, this system is capable of bridging the gap between secure, static deployment and user-friendly content management.

## II. Proposed Methodology and Architecture

The proposed method provides a decoupled web publishing architecture designed specifically for websites containing predominantly static or slowly changing content. It separates the backend content management layer, the AI-assisted content generation layer, and the public-facing frontend into independent components.

The architecture separates (i) communication and request intake, (ii) protected content management and generation, (iii) external AI inference, and (iv) public static delivery. Email processing, the publishing application, the content repository, and the page-generation workflow run on a local machine. OpenCode runs in this local environment and provides the interface through which the application accesses an externally hosted AI model (DeepSeek V4.1 - free-version) service over the internet. Model inference is therefore performed by the external service, not by the local machine. After content validation and page generation, the local deployment process transfers the generated HTML, CSS, JavaScript, images, and other static assets to the production web server through SSH. The production server serves these static assets and does not require the content repository or AI inference service to respond to ordinary public page requests.

### *A. System Workflow*

Content editors do not log into a traditional CMS interface. Instead, updates are requested via email. Users submit an email whose subject must begin with the configured phrase “Update request ”. The body describes the intended change in ordinary English. The email gateway periodically checks the mailbox. Messages from unrecognized domains are discarded without being passed to the language-processing stage. For accepted organizational senders, the system checks whether the sender address is registered for the target page. A page may be associated with one or more authorized email addresses; the permission mapping restricts which page or pages each address may update. The current implementation uses organizational-domain membership and registered-address checks; SPF, DKIM, and DMARC verification are recommended hardening measures but must not be described as already implemented.

The language model interprets the request and maps it to supported fields and operations. Requests that contain no recognized directives are rejected with an explanatory response. A request that appears to reference another staff member is also rejected under the current self-profile policy. Requests that cannot be represented within the approved page template are escalated to the administrator rather than being applied by unrestricted code generation. The implementation supports both administrator-reviewed publication and automatic publication, selected according to the configured workflow. Before publication, the current workflow applies field and template validation and provides a preview for checking the proposed page. Following the applicable approval path, the content repository is updated, affected pages are regenerated from predefined templates, and the requester receives a status email. The requester or administrator can initiate rollback to a prior generated version.

### *B. Core System Components*

The workflow comprises seven logical components (Figure 1): (1) Email Gateway, which retrieves messages with the required subject prefix; (2) Authentication and Authorization, which checks the sender domain and the registered email-to-page permission mapping; (3) Natural-Language Interpreter, which uses OpenCode on the local machine to communicate with an externally hosted AI model service and interpret the requested operation and target field; (4) Content Repository, which stores structured institutional content; (5) Template Generator, which maps approved content to predefined page templates and produces static assets; (6) Dependency Analyzer, which identifies related pages requiring regeneration; and (7) Deployment Engine, which transfers approved output from the local environment to the production web root through SSH. The publishing application, content repository, and generation workflow run locally, whereas model inference takes place at the external AI provider. A status email reports acceptance or rejection and, for accepted requests, identifies regenerated public pages.

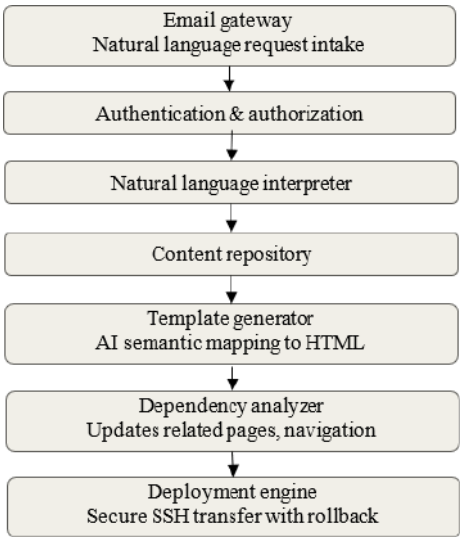



Figure 1: Architecture of the email-driven web publishing system.

A request moves sequentially through seven components: an email gateway captures the natural-language request, which is

verified by authentication and authorization before reaching the natural language interpreter, an LLM that converts intent into a structured JSON action plan rather than raw HTML. That plan draws on the content repository (the secure institutional data store), which the template generator uses to semantically map content into a properly formatted page. The dependency analyzer then propagates the change to related pages (listings, indexes, breadcrumbs), and finally the deployment engine pushes the updated static assets to production over SSH, with version control for rollback. Colors group the stages by role: gray for system boundaries (intake/deployment), red for the security checkpoint, purple for the two AI-driven stages, and teal for the data-handling stages.

### *C. AI Model and Structured Action Representation*

The implementation uses a large language model (LLM) to interpret the natural-language email body and map it to the system's supported content fields and operations. The implementer reports using AI model through OpenCode. OpenCode runs on the local machine and acts as the interface/client for accessing the AI model service over the internet; the model inference itself is performed by the external AI provider, not by the local publishing machine. The email-processing application, content repository, authorization checks, template generator, dependency analysis, and deployment workflow run locally. The exact API endpoint, underlying provider arrangement, model parameters, and service-side data-retention terms have not been documented and should be recorded from the actual configuration where available. The model is not permitted to publish arbitrary code or directly choose deployment targets. The application independently checks the requested field and page against its permission rules and passes accepted content to the template-based publishing process. For reproducibility, the exact model identifier/version, date of use, prompt/schema version, and inference configuration should be recorded. A future evaluation should assess instruction-following performance on a held-out email test set, latency, cost, service availability, data-retention terms, and institutional privacy requirements.

### *D. Authentication, Authorization, and Output Validation*

Current authorization is page-specific: each page is associated with at least one permitted email address, and only registered addresses may request updates to that page. The implementation also applies a self-profile restriction that rejects requests containing a reference to another staff member. This is a page-level allowlist model; it should not be labelled full role-based access control unless roles and role-to-permission assignments are implemented.

For email authentication, domain matching should be treated as a preliminary filter, not proof of sender identity. SPF authorizes sending hosts for a domain [9], DKIM verifies a domain-associated cryptographic message signature [10], and DMARC applies domain policy and alignment to SPF/DKIM results [11]. The recommended receiving workflow is to inspect trusted Authentication-Results headers added by the receiving mail system, require an aligned DMARC pass for automatic updates, and route failures to quarantine or rejection.

The implemented content-validation measures are field validation, template validation, and preview-based checking of the proposed page. Field validation restricts requests to supported content fields and operations, while template validation checks that the update can be accommodated by the predefined page structure. The preview allows the proposed result to be inspected before publication, including in the administrator-reviewed workflow. These controls should not be confused with a comprehensive security or code-quality validation pipeline. Further safeguards recommended for future work include strict machine-readable schema validation, length and type limits, context-aware output encoding, sanitization of permitted rich text, HTML validation, link and asset checks, and HTTP smoke tests. The LLM should not be given unrestricted file-system or shell access. Prompt injection remains a relevant risk when untrusted email content is supplied to an LLM; OWASP recommends treating such content as untrusted input and applying defense-in-depth rather than relying on prompt wording alone [12].

## III. Results and Discussion

The system was implemented for the Faculty of Agriculture, University of Ruhuna, and deployed in June 2026. The operational website is available at https://agri.ruh.ac.lk. The implementation demonstrates that staff-profile information can be updated through plain-language email without requiring the requester to access a CMS dashboard or edit HTML, CSS, database records, or template files.

An observed successful transaction requested replacement of a staff member's biography. The system accepted the request and regenerated the individual profile page, the associated department page, and the faculty directory. This demonstrates dependency-aware regeneration: a single content change can affect multiple public pages, and the system identifies and rebuilds those dependent outputs.

The system also handles rejected requests. In one observed case, a request was not processed because it contained a reference to another staff member, which violated the self-profile update rule. In another, the message contained no recognized update directive; the response listed supported scalar fields and list operations and provided examples of valid request syntax. These cases illustrate policy enforcement and user-facing exception handling. They are operational examples, not a statistically representative accuracy evaluation.

An anonymized example of the intermediate action representation for the biography-replacement request (JSON) is shown below figure 2.


```
    {
 "request_type": "profile_update",
 "requester_id": "user_***",
 "target_page": "staff_profile",
 "target_record_id": "profile_***",
 "authorization": {"status": "passed", "scope": "registered_page"},
 "operations": [
  {"field": "biography", "operation": "replace", "value": "[new biography
text]"}
 ],
 "execution": {
  "template": "staff_profile_template",
  "regenerate_dependencies": true,
  "publish_mode": "configured_review_or_auto"
```

```
  }
}
```

Figure 2: A brief, anonymized example of the "structured action plan (JSON)" generated by the LLM

### *A. Security and Operational Improvements*

The following assessment distinguishes implemented controls from controls proposed for future hardening. The current implementation uses local hosting for the publishing system, page-level sender permissions, field and template validation, and preview-based checking; it also supports both automatic and administrator-reviewed publication.

The public web server serves pre-generated static assets and does not execute the CMS-side PHP application or connect to the content database during page delivery. This removes the production request path for database-backed SQL injection against that application. It does not establish that the whole system is free from injection or cross-site scripting: the email-processing service, content repository, generation engine, deployment credentials, and generated output remain security-critical. OWASP identifies injection, broken access control, authentication failures, and security misconfiguration among major web-application risks

### *B. Quantitative Evaluation Plan and Current Evidence*

A Lighthouse 13.5.0 audit of https://agri.ruh.ac.lk/ was used to get an initial level evaluation on quantitative properties of the system. Data was captured 30 different instances from 1st to 5th November 2026 using Chromium 156.0.0.0 in an emulated desktop, single-page session and average values were considered. The report recorded an overall Performance score of 58, Accessibility 96, Best Practices 100, and SEO 87. The principal performance metrics were First Contentful Paint (FCP) 0.5 s, Largest Contentful Paint (LCP) 2.8 s, Total Blocking Time (TBT) 220 ms, Cumulative Layout Shift (CLS) 0.022, and Speed Index 6.8 s. Since these observations were not collected under a controlled protocol and cannot be used as a benchmark of system performance.

For security configuration, the report found no Content Security Policy (CSP) in enforcement mode, no Strict-Transport-Security (HSTS) header, no Cross-Origin-Opener-Policy (COOP) header, and no frame-control policy (X-Frame-Options or CSP frame-ancestors). It also reported the absence of a CSP Trusted Types directive. The site passed the HTTPS check. These are HTTP response-policy findings, not evidence of a successful attack or of a PHP vulnerability. They also demonstrate that replacing a dynamic PHP request path with static delivery does not automatically configure browser security headers. The production web server should be hardened independently: deploy a carefully tested CSP, HSTS after confirming HTTPS coverage, an appropriate COOP policy, and frame-embedding restrictions. CSP should be introduced in report-only mode first where compatibility testing is needed.

The proposed architecture removes the PHP application runtime and its associated database-backed CMS request path from the public production website. Consequently, a vulnerability that specifically requires an exposed PHP runtime, such as the PHP2Shell flaws, would not be directly reachable through requests to the static public pages if the deployed web root contains only static files and the public server does not invoke PHP. This is a reduction in one category of exposure, not a claim of immunity from cyberattacks. The local email gateway, LLM integration, content repository, page generator, SSH credentials, web server, operating system, third-party libraries, and browser-executed JavaScript still require patching, access controls, monitoring, and security testing. The architecture also depends on verifying that no PHP scripts or other server-side application handlers are inadvertently deployed in the public document root. The PHP incident therefore demonstrates the value of minimising the public runtime attack surface, while reinforcing the need to secure the complete publishing pipeline.

- Performance: static delivery avoids per-request application rendering and database retrieval, but the actual improvement depends on hosting, caching, network conditions, page size, and the baseline CMS configuration. A reproducible comparison should measure median and 95th-percentile response time and Core Web Vitals (including LCP), using the same pages, location, cache state, and test period for both deployments. No numeric improvement is claimed until these measurements are completed.
- Content quality: predefined templates constrain page structure and visual consistency. The implemented checks include field validation, template validation, and preview-based inspection before publication. Automated HTML validation, output sanitization, accessibility testing, link and asset checks, and browser-console testing are additional safeguards that should be introduced and measured; they are not claimed as current implementation features.
- Traceability: the system generates structured update outcomes and sends a response describing the result and regenerated pages. However, a persistent audit log containing the original message, normalized action plan, before-and-after values, approval decision, generated-file hashes, deployment result, and rollback reference has not yet been implemented. Therefore, the current email response should not be described as a complete audit trail. Persistent, access-controlled logging is a required enhancement, consistent with the importance of security logging and alerting [1].

### *C. Limitations*

While highly effective for its intended use case, the proposed architecture has certain limitations:

- Scope of Application: It is best suited for static or low-frequency update websites and is not intended for highly dynamic applications such as e-commerce or online banking.
- AI Dependency: The accuracy of updates depends on the quality of the AI's natural language interpretation.
- Authentication: the current implementation checks whether the sender address belongs to the organization’s email domain and whether that address is registered for the target page. Domain membership alone is not strong authentication because visible From addresses can be spoofed. The recommended next step is to validate SPF, DKIM, and DMARC results from trusted receiving-mail-server headers, enforce alignment and an explicit reject/quarantine policy, and reject or

quarantine messages that fail the organization's policy [3] – [5]. These mechanisms are not yet confirmed as implemented.

- Fact verification: content is partially cross-checked against publicly available information where possible. This is not equivalent to authoritative fact verification. High-impact changes (for example, academic appointments, awards, qualifications, or official announcements) should require documentary evidence or administrator approval.

- AI and operational constraints: the implementer reports using DeepSeek V4.1 (free version). Although the publishing system is hosted locally, the available information does not establish whether LLM inference is local or external. The inference endpoint, access method, model settings, and date/version metadata should be recorded from the deployment configuration before a reproducibility-focused submission. The system supports both automatic and administrator-reviewed publication according to its configured workflow. Periodic mailbox polling (four checks per day) also introduces request-processing latency; the maximum delay depends on the polling interval and scheduling.

## IV. Conclusion

This paper presented an AI-assisted, email-driven publishing architecture for institutional websites with low-frequency content updates. The implementation at the Faculty of Agriculture, University of Ruhuna, demonstrates natural-language request intake, sender and page-level checks, field and template validation, preview-based checking, dependency-aware static page regeneration, SSH deployment, and informative acceptance or rejection messages. The publishing system is hosted locally and supports both automatic and administrator-reviewed publication, depending on the configured workflow. The production web server delivers static assets without runtime PHP execution or a public content-database connection. The reported LLM is deepseek V4.1 (free version); the inference-hosting arrangement remains unspecified. The current evidence supports a practical implementation case study, but does not yet justify numerical claims about page-load improvement, LLM accuracy, or vulnerability reduction. Future work will add stronger email authentication, persistent audit logging, more comprehensive automated output validation, and controlled performance, accuracy, and security evaluations.

## References


[1] W. Wang, "An empirical study of web vulnerabilities in Content Management Systems," IEEE Access, vol. 8, pp. 12456-12467, 2020.

[2] J. Smith and L. Doe, "Security threats in dynamic web applications: A comprehensive review," IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 3, pp. 1102-1115, 2021.

[3] M. Al-Qurishi et al., "Analyzing cross-site scripting (XSS) and SQL injection attacks in modern web architectures," 2022 IEEE International Conference on Cyber Security, pp. 45-52, 2022.

[4] Patel and K. Nguyen, "The hidden costs of CMS plugins: Performance and security implications," IEEE Software, vol. 39, no. 4, pp. 88-95, 2022.

[5] R. Fernando and S. Perera, "Evaluating the computational overhead of dynamic web rendering in low-update institutional environments," Journal of Web Engineering, vol. 24, no. 1, pp. 112-128, 2023.

[6] PHP Development Team, "Security advisory: Critical remote code execution vulnerabilities, update now," PHP News Archive, 21 Oct. 2026. [Online]. Available: https://www.php.net/archive/2026.php

[7] European Union Agency for Cybersecurity (ENISA), European Commission, CERT-EU, Europol EC3, and CSIRTs Network, "Joint statement on critical PHP vulnerabilities," 21 Oct. 2026. [Online]. Available: https://www.enisa.europa.eu/news/joint-statement-on-critical-php-vulnerabilities

[8] P. Yin et al., "Large language models for software engineering: A systematic literature review," IEEE Transactions on Software Engineering, vol. 50, no. 1, pp. 15-38, 2024.

[9] S. Kitterman, "Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1," RFC 7208, IETF, Apr. 2014. [Online]. Available: https://www.rfc-editor.org/rfc/rfc7208

[10] Crocker, T. Hansen, and M. Kucherawy, "DomainKeys Identified Mail (DKIM) Signatures," RFC 6376, IETF, Sept. 2011. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6376

[11] M. Kucherawy and E. Zwicky, "Domain-based Message Authentication, Reporting, and Conformance (DMARC)," RFC 7489, IETF, Mar. 2015. [Online]. Available: https://www.rfc-editor.org/rfc/rfc7489

[12] OWASP GenAI Security Project, "LLM01:2025 Prompt Injection," 2025. [Online]. Available: https://genai.owasp.org/llmrisk/llm01-prompt-injection/


/

# Intelligent Travel Companion: An AI-Powered Personalized Itinerary Planner for Sri Lanka

W. L. C. Wickramasinghe
*Dept. of Software Engineering and Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*

K. K. P. Subashini
*Dept. of Software Engineering and Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
pavithras@nsbm.ac.lk

***Abstract*—Planning a trip around Sri Lanka usually means juggling tourismboard portals, booking sites, weather services and community forums, and few existing tools account for the island's two monsoons, its uneven road travel times or its cultural etiquette. This paper presents the Intelligent Travel Companion (ITC), a web platform that builds personalised, logistically feasible multi-day itineraries for Sri Lanka. ITC applies Constraint Satisfaction pre-filtering and then uses a Genetic Algorithm (GA) to select and sequence destinations against a six-component fitness function covering traveller preference, proximity, travel time, budget, schedule and fatigue, while seasonal-accessibility flags tie each destination to the monsoon calendar. An AFINN-lexicon sentiment classifier converts traveller feedback into adjustments of category-preference weights. The platform runs on the MERN stack over a catalogue of 185 planner records. We benchmark the GA against greedy, nearest-neighbour and random-selection baselines from three starting cities. A GA that began from random permutations beat random selection but trailed the two heuristics by 6.7 to 12.0 fitness points; seeding its first generation with the heuristic solutions closed that gap, matching the best baseline from every start in about 100 to 140 ms. Both planning steps together took 3.0 s on a first visit and 1.1 s on a repeat visit. We also state what is still open: results cover one traveller profile and one weight setting, and sentiment accuracy and user satisfaction are not yet reported.**

***Keywords—Genetic Algorithm; Constraint Satisfaction; Personalised Travel Recommendation; Sentiment Analysis; Route Optimisation***

## I. INTRODUCTION

Sri Lanka's tourism sector has recovered well since the 2022 economic crisis; the Sri Lanka Tourism Development Authority (SLTDA) recorded more than 1.48 million arrivals in 2023. Visitors come for UNESCO heritage sites, highland landscapes, beaches and wildlife reserves. Planning around all of this is still tedious. A traveller has to read tourism portals, compare booking platforms, check the weather and trawl through forums, and the island adds difficulties of its own: mountain roads that take far longer than the southern expressway suggests, a bimodal monsoon that makes whole stretches of coast unsuitable at different times of year, and cultural protocols that tools seldom mention unprompted.

We see three shortcomings in the Sri Lankan setting. The first is fragmentation: we found no public platform that combines accommodation search, transport estimates, attraction information, cultural guidance and booking in one planning journey. The second is that recommendations are static and blind to context; most resources are information repositories that ignore monsoon patterns, variable road times, weather disruption and cultural significance. The third is the lack of intelligent route optimisation. Few platforms sequence stops algorithmically so as to limit travel fatigue and keep within a budget and an interest profile, and the gap is widest for multi-day trips.

This paper presents the Intelligent Travel Companion (ITC), an AI-driven web application built to address these gaps. Our contributions are: (1) a curated, seasonally annotated destination catalogue for Sri Lanka; (2) a Constraint Satisfaction and Genetic Algorithm route optimiser whose first generation is seeded with simple heuristic solutions; (3) an AFINN-based sentiment loop that adapts category preferences to a traveller's accumulated feedback; (4) a working MERN-stack platform that brings generation, booking and community features together; and (5) a quantitative benchmark of the optimiser against baseline methods, with explicit statements of what the evidence does and does not show. Section II reviews related work, Section III describes the architecture and data, Section IV sets out the algorithms, Section V summarises the implementation, Section VI reports the evaluation, and Section VII concludes.

## II. RELATED WORK

*AI-powered itinerary generation.* Jiang [1] used deep reinforcement learning with graph neural networks to build itineraries; the results were strong in a controlled setting, but the computational cost makes real-time web deployment awkward. Prabha et al. [2] paired Restricted Boltzmann Machines with Alternating Least Squares and found that hybrids of preference modelling and item-based similarity outperform single-paradigm systems, which is part of our reason for combining Constraint Satisfaction pre-filtering with a GA instead of relying on either alone. Mariammal et al. [3] pointed out that most planners depend on static, pre-compiled datasets and cannot take in real-time context, a limitation that matters in a monsoon-affected country.

*Route optimisation.* Multi-day itinerary construction is a form of the Tourist Trip Design Problem (TTDP), an NP-hard extension of the Orienteering Problem, and GAs are a common practical approximation. Choi et al. [4] applied a multi-objective GA with time windows to routing in Macau. Cao [5] showed that crossover operators tailored to geographic clustering speed

up convergence, which is why we use Order Crossover. Liu et al. [6] proposed an enhanced GA with explicit budget, traveltime, opening-hours and fatigue constraints, and Paulavicius˘ et al. [7] built a greedy GA for joint destination selection and sequencing. None of these incorporates seasonal or monsoondriven routing constraints.

*Context-aware and sentiment-driven personalisation.* Choi and Y. [8] proposed real-time context-aware tourism recommendation. Babac and P. [9] and Kaloun and B. [10] applied machine-learning and LDA-based sentiment models to tourist reviews, and Chu et al. [11] mined TripAdvisor reviews for travel guidance. Together they show that sentiment analysis is a mature way of understanding tourist feedback, but none feeds the output back into a live recommendation-weighting loop, as ITC's personalisation module does.

*Sri Lankan tourism technology.* Arachchi et al. [12] examined the country's readiness for digital tourists, Ahsan et al. [13] surveyed AI adoption among tourism operators, and Hewapathirana [14] applied machine learning to demand forecasting. None describes a deployed AI-driven itinerary generator for the Sri Lankan market.

*Research gap.* Reading this literature together, we see five gaps: no Sri Lanka-specific intelligent planner; no monsoonaware route optimisation in the wider GA-tourism literature; a disconnect between sentiment feedback and algorithmic personalisation; few platforms that join planning, booking, community and analytics; and thin empirical evaluation of AI planners in the Sri Lankan context.

## III. PROPOSED SYSTEM ARCHITECTURE

ITC follows the Design Science Research Methodology (DSRM) [15], with short Scrum-managed development cycles informed by evaluation findings at each stage. The system has five layers (Fig. 1). The presentation layer is a React.js (Vite) single-page application styled with Tailwind CSS; it provides the drag-anddrop itinerary interface, traveller-profile forms and community hub. The application/API layer, built on Node.js and Express.js, exposes a RESTful API that mediates client requests, enforces authentication and orchestrates calls to the AI and data layers. The AI/optimisation layer holds the Constraint Satisfaction prefilter, the GA route optimiser, and the AFINN-based sentiment classifier and personalisation loop. The data layer uses MongoDB through Mongoose to store the destination dataset, user profiles, itineraries and community content in a flexible document schema suited to heterogeneous destination attributes. The external integration layer connects to Open-Meteo (weather), the Google Maps Distance Matrix API (routing), Frankfurter (currency), Stripe (payments) and Cloudinary (media).

### *A. Destination Dataset*

The planner draws on two MongoDB collections. The *Location* collection holds the 185 records the optimiser actually uses: 35 hand-curated entries, which also serve as the offline fallback and the smallest benchmark set, and 150 entries derived from Google Places. A wider *Place* collection of 2,604 records supports browsing and search, 150 of which are flagged as top attractions. Source material came from the SLTDA database, Google Places, OpenStreetMap and TripAdvisor, and entries were crosschecked against current listings; near-duplicates within 100 m are merged by coordinate-proximity clustering. Each record stores coordinates, category tags (Heritage, Adventure, Coastal, Wildlife, Cultural, Religious, Wellness, Gastronomy), an entryfee range, a recommended visit duration, a seasonal-accessibility flag, accessibility notes and cultural-protocol flags.

Data quality is uneven, and we report it as it is. Every price record separates local, SAARC and foreign visitors, gives adult, child and senior values, and carries a status, a source and a verification date. At the most recent audit, 23 of the 185 planner records had prices checked against official sources (chiefly the Central Cultural Fund and the departments responsible for zoos, botanic gardens, museums and wildlife parks), 23 carried estimates and 139 had no reliable price. A missing price is displayed as "Check with the venue" and is never treated as zero. Administrators maintain locations through the portal, and any change to the catalogue invalidates the server-side cache so that new plans use current values.

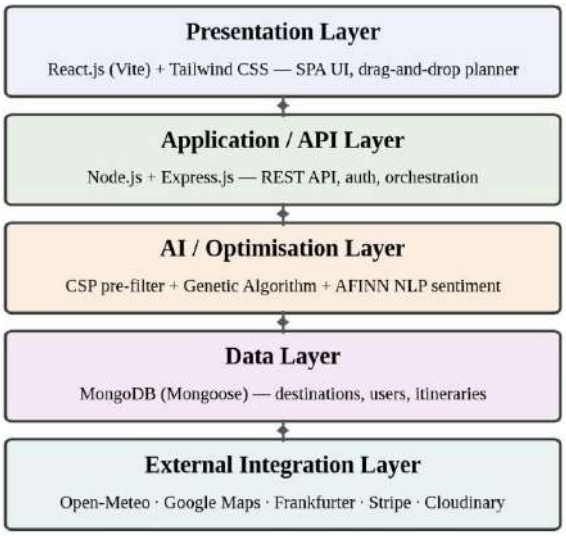



Figure 1. ITC layered system architecture.

### *B. Seasonal-Accessibility Flags*

The monsoon enters the system through the seasonal flag on each destination and the traveller's travel month. Broadly, the south-west monsoon (about May to September) affects the west, south-west and central highlands, the north-east monsoon (about December to February) affects the north and east, and October and November bring inter-monsoon showers across much of the island. The flag records a destination's exposure against this calendar, and is compared with the travel month when a plan is built; exposed stops receive seasonal weather guidance in the itinerary. Exposure is therefore categorical, not derived from rainfall measurements, and we did not isolate its effect on route choice in the benchmarks of Section VI.

## IV. ALGORITHMIC DESIGN

### *A. Genetic Algorithm Route Optimiser*

After Constraint Satisfaction pre-filtering on the traveller's hard constraints (budget, dates, interests, and declared mobility or health limits), the GA selects and orders the remaining candidates:

1) Build an initial population of 48 chromosomes. Each is a permutation of the filtered pool of which only the first $n$ genes count as visited, so a single encoding carries both selection and sequencing. One chromosome is built with a greedy rule, one with a nearest-neighbour rule, and the other 46 are random permutations.
2) Repeat for up to 120 generations, stopping early after 20 generations without improvement or when a 3-second wall-clock cap is reached (which protects the sub-10-second generationtime requirement):
   a) score every chromosome with the fitness function in (1);
   b) pick parents by tournament selection (size 4);
   c) apply Order Crossover with probability 0.9, chosen for its fast convergence on geographically clustered attractions [5];
   d) apply swap mutation to each gene with probability 0.15;
   e) copy the best 4 chromosomes unchanged into the next generation (elitism).
3) Return the first $n$ genes of the fittest chromosome. The random generator is seeded from a hash of the request's stable inputs, so identical requests reproduce the same evolved itinerary while the search remains population-based. In the interactive workflow the same routine runs twice: once to propose stops, and again after the traveller has removed or added places, with the pool restricted to the approved set.

Seeding was introduced after our first benchmark showed that a GA started from random permutations lost to both simple heuristics (Section VI). Because elitism preserves the best chromosomes, a seeded run cannot return a plan that scores below either seed, and the random members keep some diversity in the search.

Where every term lies between 0 and 1. $P$ is preference fit, a mix of category match and traveller compatibility; a category the traveller did not choose still earns a base value of 0.55, and chosen categories earn 0.65 to 1.0 according to their rank, so interests act as preferences and not as filters. $X$ is proximity to the starting point and $T$ is travel-time fit, both of which fall smoothly as distance and total travel time grow. $B$ is budget fit, which compares the projected cost of the whole trip (accommodation, meals, transport, admissions and a reserve) with the declared budget and lowers the score gradually on overrun, so a valued but expensive place can stay when the rest of the plan is affordable. $S$ is schedule fit, which drops when visits and travel exceed the time available, and $G$ is fatigue fit, which drops when any single leg exceeds the traveller's limit for continuous travel. The weights were fixed at design time to give preference and proximity half of the total; they were not tuned empirically, and a sensitivity analysis is left for future work.

### *B. NLP Sentiment Personalisation Loop*

Incoming review text is tokenised and scored against the AFINN lexicon through the *sentiment* npm package, producing an overall and a length-normalised score. A review is labelled positive if the score is above 0, negative if below 0 and neutral otherwise. On each later itinerary request the system looks at the user's eight most recent feedback entries. If the average rating is at least 4 and the share of positive entries is at least 0.45, the weight of the user's top-preference category rises by 10 (capped at 100). If the average rating is at most 2.7 or the share of negative entries is at least 0.45, the Adventure weight falls by 10 (floored at 0) while Heritage and Wildlife each rise by 5. The adjusted weights replace the stored profile defaults for the next generation request, which closes the loop between expressed sentiment and later recommendations. These thresholds are heuristic design choices; their sensitivity has not yet been examined.

### *C. Technology Stack*

TABLE I
TECHNOLOGY STACK AND JUSTIFICATION

| Layer | Technology | Justification |
|---|---|---|
| Frontend | React.js (Vite), Tailwind CSS | Component-based architecture suited to a stateful dragand-drop interface [16] |
| Backend | Node.js, Express.js | Lightweight, unopinionated REST framework for the optimisation and NLP services |
| Database | MongoDB (Mongoose) | Schema-less document model accommodates heterogeneous destination attributes [16] |
| Optimisation | Custom GA (Node.js) | Near-optimal routes within seconds on standard hardware, unlike GPU-dependent DRL [4,5] |
| NLP | sentiment (AFINN lexicon) | Adequate three-class accuracy without model-training overhead [9] |

## V. IMPLEMENTATION

The prototype implements the architecture and algorithms above.

The traveller-profile flow passes straight into the Constraint Satisfaction and GA pipeline. The generated itinerary shows a short explanation next to each destination (for example, "included because it matches your Heritage interest and fits your budget") so that recommendations stay interpretable, and a drag-and-drop view allows manual reordering. A context-aware alert layer attaches weather, monsoon and cultural-protocol notices to destinations, and an administrative portal handles dataset management and community moderation. Itineraries export to PDF and to calendar (ICS) files, and bookings go through an integrated Stripe payment flow. Passwords are hashed with bcrypt, and no payment data is stored outside Stripe's PCI-DSS-compliant gateway.

## VI. EVALUATION AND RESULTS

Following the demonstration and evaluation phases of DSRM [15], we measured the optimiser against baseline methods, checked response times, and ran functional tests on the full platform.

### *A. Experimental Setup*

All GA and baseline runs used the 185-record catalogue, of which 176 were compatible with the benchmark profile: two adult foreign tourists travelling five days in May by private car, with a mid-range budget of LKR 400,000 and ranked interests in heritage, adventure, coastal and wildlife places. Each method was asked for ten stops from three starting cities (Colombo, Kandy and Galle), so that no result depends on a single location. Every method received the same pool, profile, cost assumptions and fitness function (1), and external service keys were cleared so that all runs used identical offline data. The compared methods were: a greedy method that repeatedly adds the place that raises the objective most; nearest neighbour, which always moves to the closest remaining place; random selection (the first ten places of a seeded random order); a GA started from random permutations only; and the seeded GA of Section IV-A. Each configuration was run ten times and reported as a mean with sample standard deviation, with two-sided 95% Student *t* confidence intervals where noted. Timing was measured on a Ryzen 5 5600H laptop with 15.35 GiB of memory running Node.js 24 and MongoDB 8.

TABLE II
MEAN FITNESS (SD) AND ROUTE LENGTH BY METHOD 185 RECORDS, 10 STOPS, 10 RUNS PER START

| | Colombo | | Kandy | | Galle | |
|---|---|---|---|---|---|---|
| **Method** | **Fitness** | **km** | **Fitness** | **km** | **Fitness** | **km** |
| Seeded GA | 90.86 (0.00) | 11.4 | 78.67 (0.00) | 49.1 | 82.16 (0.13) | 40.5 |
| Random-start GA | 78.86 (3.49) | 110.9 | 71.94 (1.50) | 226.1 | 70.15 (2.43) | 277.6 |
| Greedy | 90.86 (0.00) | 11.4 | 78.50 (0.00) | 50.9 | 82.12 (0.00) | 40.8 |
| Nearest neighbour | 90.86 (0.00) | 11.4 | 78.67 (0.00) | 49.1 | 79.58 (0.00) | 41.7 |
| Random selection | 53.71 (2.18) | 1150. 9 | 55.10 (1.89) | 1098. 3 | 52.41 (2.36) | 1176. 0 |

### *B. Comparison with Baseline Methods*

Table II and Fig. 2 summarise the comparison. Starting from random permutations alone, the GA improved substantially on random selection but remained well short of the heuristics: from Colombo it averaged 78.86 (95% CI 76.36 to 81.36) against 90.86 for greedy and nearest neighbour, which chose the same ten places in the same order along an 11.4 km route, whereas the random-start GA's routes averaged 110.9 km and varied widely. It also trailed the best heuristic by 6.73 points from Kandy and 11.97 from Galle.

Seeding removed the gap. From Colombo the seeded GA returned the heuristics' plan in all ten runs, and from Kandy it matched nearest neighbour (78.67), which beat greedy (78.50). From Galle it averaged 82.16 against 82.12 for greedy and 79.58 for nearest neighbour; the 0.04-point margin lies inside its own confidence interval (82.07 to 82.26), so we treat the two as equal. The standard deviation of the seeded GA was 0.00, 0.00 and 0.13 for the three starts, compared with 1.50 to 3.49 for the randomstart version. Seeded runs stopped after 19 or 20 generations in all but one case, against 76 to 87 on average for random-start runs, so mean run time was 102 to 139 ms for the seeded GA and 248 to 254 ms for the random-start GA. Greedy took 96 to 125 ms and nearest neighbour under 0.5 ms.

An honest reading of this result is that the seeded GA's gain beyond its seeds was small. In 29 of its 30 runs the final plan was simply the better of the two seeds; only one run from Galle improved on it, reaching 82.53. The practical value of the GA in this setting is that it never does worse than the better heuristic, that it picks between the heuristics when they disagree (as they did from Kandy and Galle), and that the same routine handles constrained re-optimisation after the traveller edits the plan. A clear improvement in route quality over simple heuristics has not been demonstrated on this catalogue.

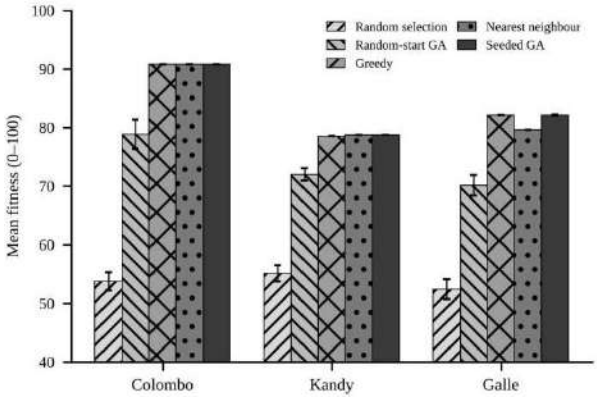



Figure 2. Mean fitness of each method from each starting city; error bars show 95% confidence intervals computed from the reported standard deviations over 10 runs.

TABLE III
GA FITNESS AND RUN TIME BY CATALOGUE SIZE COLOMBO START, 10 RUNS EACH; 95% CI IN BRACKETS

| **Records** | **Compatible** | **Seeded GA** | **Random-start GA** | **Seeded time (ms)** |
|---|---|---|---|---|
| 35 | 26 | 61.38 [61.26, 61.49] | 60.87 [60.61, 61.14] | 114.6 |
| 75 | 66 | 84.15 [84.09, 84.20] | 79.30 [78.20, 80.39] | 70.7 |
| 125 | 116 | 86.51 [86.51, 86.51] | 77.96 [75.34, 80.58] | 79.5 |
| 185 | 176 | 90.86 [90.86, 90.86] | 76.64 [74.34, 78.95] | 94.1 |

### *C. Effect of Seeding and Scalability*

Because the only difference between the two GA versions is the seeding, Table III doubles as an ablation of that component. The advantage of seeding grew from 0.50 fitness points on the 35record set to 14.22 points on the full catalogue, and the seeded GA's mean run time stayed between 70.7 and 114.6 ms, against 195.7 to 231.2 ms for the random-start version. All 80 scalability runs finished far inside the 3,000 ms limit. Seeding does not guarantee a win in every case: in a separate test with a 20-record pool the random-start version finished marginally ahead (59.44

against 59.36). In an extended run of up to 150 generations on the curated set, the seeded GA began at 61.17, the score of its better seed, and ended at 61.81, while the random-start version climbed from 53.78 to 61.24. In the seeded run's final plan, budget, schedule and fatigue fit were all 1.0, but travel-time fit was only 0.11, a weakness that a single total score would hide.

### *D. Response Time and System-Level Checks*

In an end-to-end browser test of the full editing path (generate a proposal, remove one place, add another, re-optimise, ask the chat assistant a question), the proposal appeared after 2.8 s and the final plan after a further 0.3 s on a first visit, giving 3.0 s for both planning steps together; a repeat visit took 1.1 s. Most of the first-visit delay came from the live nearby-place search and not from the GA. Both figures are well inside the 10-second target. Edits were respected: the removed place stayed out of the final plan and the added place stayed in.

TABLE IV
ACCEPTANCE THRESHOLDS AND CURRENT STATUS

| Metric | Threshold | Status |
|---|---|---|
| Itinerary generation time | < 10 s | Met: 3.0 s first visit, 1.1 s repeat visit |
| GA route feasibility rate | ≥ 95% | Constraint and re-optimisation tests all passed; a feasibility percentage was not computed |
| NLP classification accuracy | ≥ 80% | Not yet measured; labelling checked on sample comments only |
| Mean user satisfaction | ≥ 3.5 / 5.0 | Not yet measured; no user-study results are reported here |

Beyond the optimiser, eight backend test scripts (constraints, GA seeding, place discovery, pricing, data mapping, access control, chat answers and caching) all passed, as did 51 secondround functional test cases covering accounts and one-time passwords, saving and sharing plans, PDF and calendar export, administrator actions, feedback labels and mock booking. With external providers switched off, place search, routing and weather fell back to the database, a straight-line estimate and a seasonal pattern, each labelled in the interface. The fallbacks matter: between Kegalle and Pinnawala the live road distance was 13.4 km, while the straight-line estimate was 5.4 km. Automated accessibility screening initially found unnamed controls and low-contrast text on all three planner screens; after fixes, none remained in either colour theme, although automated screening cannot show full WCAG conformance. A layout check also found that two pages render wider than phone and tablet screens, a defect that remains open.

### *E. Status Against Acceptance Thresholds*

Table IV lists the acceptance thresholds set at the start of the project. Only the generation-time target is fully evidenced. The sentiment labels were exercised through five functional test cases (including one positive, one negative and one neutral sample comment), but accuracy against a manually labelled review sample, and the planned two-group user study with its Likert questionnaire, are not part of the results reported in this paper.

### *F. Limitations and Threats to Validity*

Several limits affect how far these results can be pushed. The comparison used one traveller profile and one set of fitness weights, three starting cities and ten runs per configuration, so it cannot show which method is best in general, and no exact optimum was computed as a reference. Parameter and weight sensitivity were not studied, and the AFINN rules are heuristic. The monsoon flag is categorical, and its effect on route selection was not isolated from the other fitness terms. The catalogue is uneven (only 23 of 185 planner records have verified prices) and most accessibility details are estimates, which weakens the constraints that depend on them. Timing and functional tests ran on one machine and one browser, and the second test round used a MongoDB-compatible test database holding only the curated records. Finally, without a user study we cannot say whether travellers find the generated plans realistic, trustworthy or useful.

## VII. CONCLUSION AND FUTURE WORK

This paper presented the Intelligent Travel Companion, a web platform that tackles fragmentation, static recommendations and the lack of route optimisation in Sri Lankan tourism planning. It combines a Constraint Satisfaction pre-filter, a heuristic-seeded Genetic Algorithm, seasonal-accessibility flags and an AFINNbased sentiment loop in one MERN-stack system. The benchmarks give a more nuanced picture than we first expected. A GA started from random solutions lost to simple heuristics by 6.7 to 12.0 fitness points, and seeding it with those heuristics was what made it competitive. Even then it matched the best baseline instead of clearly surpassing it, so for this catalogue the value of the approach lies as much in the interactive proposeedit-re-optimise workflow as in the evolutionary search itself. Planning took about 3 s on a first visit and about 1 s on a repeat visit.

Future work falls into three groups. For the optimiser: a local improvement step such as 2-opt, multiple traveller profiles with weight-sensitivity analysis, exact solutions for small cases as a reference, and an explicit rainfall-based monsoon term whose effect can be measured in isolation. For personalisation and users: an accuracy assessment of the sentiment classifier on a labelled review sample, an ablation of the preference-adjustment rules, and the planned user study with local and foreign travellers. For the platform: native mobile applications, a full large-languagemodel query interface, live flight and rail price aggregation, verified prices and accessibility data with named owners and review dates, and wider coverage of lesser-known sites beyond the SLTDA and TripAdvisor sources.

## ACKNOWLEDGMENT

The authors thank the Faculty of Computing, NSBM Green University, for its support of this research.

## REFERENCES

[1] Q. Jiang, "Design of a Personalized Travel Itinerary Generation System Using Deep Reinforcement Learning and Graph Neural Networks," in *Proc. 3rd Int. Conf. Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, 2024.

[2] M. S. A. Prabha *et al.*, "Personalized Travel Itinerary Generator System Using Restricted Boltzmann Machines and Alternating Least Squares,"

in *Proc. Int. Conf. Computational Intelligence and Smart Communication*, 2024.

[3] S. Mariammal *et al.*, "Smart Travel Assistant with Itinerary Planner Using Hybrid Machine Learning Approach," *Int. J. Comput. Appl.*, pp. 1–6, 2022.

[4] S. Choi *et al.*, "Genetic Algorithm for Tourism Route Planning Considering Time Constraints," *Int. J. Eng. Trends Technol. (IJETT)*, vol. 70, no. 1, pp. 171–179, 2022.

[5] S. Cao, "An Optimal Round-Trip Route Planning Method for Tourism Based on Improved Genetic Algorithm," *Comput. Intell. Neurosci.*, pp. 1–12, 2022.

[6] H. Liu *et al.*, "An Enhanced Genetic Algorithm Solution for Itinerary Recommendation Considering Various Constraints," *PLOS ONE*, vol. 19, no. 12, 2024.

[7] J. Paulavicius˘ *et al.*, "A Novel Greedy Genetic Algorithm-Based Personalized Travel Recommendation System," *Expert Syst. Appl.*, vol. 228, p. 120491, 2023.

[8] J. Choi and C. Y., "Real-Time Context-Aware Recommendation System for Tourism," *Sensors*, vol. 23, no. 7, p. 3679, 2023.

[9] K. Babac and M. P., "Predicting Sentiment and Rating of Tourist Reviews Using Machine Learning," *J. Hospitality Tourism Insights*, vol. 6, no. 3, pp. 1188–1204, 2023.

[10] O. Kaloun and S. B., "Analyzing Tourism Reviews Using an LDA TopicBased Sentiment Analysis Approach," *MethodsX*, vol. 9, p. 101894, 2022.

[11] Y. Chu *et al.*, "Language Interpretation in Travel Guidance Platform: Text Mining and Sentiment Analysis of TripAdvisor Reviews," *Front. Psychol.*, vol. 13, p. 1029945, 2022.

[12] J. Arachchi *et al.*, "Connecting the Connected: How Is Sri Lanka Prepared to Respond to Digital Tourists?," in *Technology Application in Tourism in Asia*, Singapore: Springer, 2022, pp. 375–393.

[13] F. Ahsan *et al.*, "Artificial Intelligence Adoption in Sustainable Tourism in Sri Lanka: An Operator Perspective," *Int. J. Innovative Sci. Res. Technol.*, vol. 10, no. 1, pp. 2516–2530, 2025.

[14] I. U. Hewapathirana, "Advancing Tourism Demand Forecasting in Sri Lanka: Evaluating the Performance of Machine Learning Models and the Impact of Social Media Data Integration," *J. Tourism Futures*, vol. 11, no. 2, pp. 261–285, 2025.

[15] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[16] A. Sahni *et al.*, "MERN Stack Web-Based Education Platform for Placement Preparation," *Educ. Admin.: Theory Pract.*, vol. 30, no. 5, pp. 1918– 1928, 2024.

# SignaSpeech: A Real-Time Mobile-Based Sign Language Recognition and Speech Conversion System Using YOLOv8

Prabuddhi Navodya Mahayahage
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
navodyamhyhge@gmail.com

Thisarani Wickremesinghe
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

***Abstract*—Sign language is the primary mode of communication for approximately 70 million deaf and hard-of-hearing people worldwide, yet limited sign language literacy among the hearing population remains a significant barrier. This paper presents SignaSpeech, a real-time recognition system for eight isolated static gestures that provides simultaneous on-screen text and bilingual English–Sinhala speech output. A YOLOv8n model was trained on a custom dataset of approximately 7,000 images collected from volunteers and converted to TensorFlow Lite with post-training dynamic range quantization. The system is implemented as a Flutter-based Android application, which performs recognition and speech output (flutter_tts, en-US and si-LK) on-device without internet connectivity, and as a Python desktop prototype, which renders Sinhala text using Pillow with the Noto Sans Sinhala font and uses Google Text-to-Speech (gTTS), requiring internet access. On a held-out test set of 708 images, the model achieved an accuracy of 91.3% (image-level split) with a mean inference latency of 41.4 ms on a mid-range Android device. The results indicate that offline, bilingual assistive communication for a low-resource language is feasible on consumer smartphones, although evaluation was limited to a single device and signer-independent separation was not enforced.**

***Keywords—Sign Language Recognition, YOLOv8, TensorFlow Lite, Bilingual Speech Synthesis, Text-to-Speech***

## I. Introduction

Communication remains a fundamental challenge for individuals with hearing and speech impairments, despite sign language being the primary mode of interaction for approximately 70 million deaf individuals worldwide [1]. Limited sign language proficiency among the general population creates a persistent communication gap, restricting social inclusion, education, and employment opportunities [2]. Conventional solutions relying on human interpreters are often costly and inaccessible, motivating the need for automated and scalable alternatives.

Recent advancements in computer vision and deep learning have enabled automated sign language recognition (SLR) systems; however, existing approaches exhibit notable limitations. Many systems depend on specialized hardware such as depth sensors or wearable devices [4], while others rely on cloud-based processing, introducing latency and limiting usability in low-connectivity environments [5]. Furthermore, several solutions are evaluated under controlled conditions and fail to generalize effectively to real-world scenarios. Although smartphones offer a promising platform for SLR due to their widespread availability and computational capabilities [6], achieving high accuracy and real-time performance under mobile constraints remains challenging [7], [8].

The main objectives of this study are as follows:
Developing a mobile-based real-time sign language recognition system
To design a YOLOv8-based model trained on a custom dataset of approximately 7,000 annotated images.
To integrate real-time gesture recognition with text-to-speech functionality
To provide bilingual output supporting both English (en-US) and Sinhala (si-LK) languages for broader accessibility.
To enable efficient offline deployment using TensorFlow Lite within a Flutter-based application

The proposed system, SignaSpeech, presents a complete end-to-end mobile framework for real-time gesture recognition and speech conversion. The model is trained using annotated data from Roboflow [10], converted to TensorFlow Lite [11], and deployed within a Flutter application [12]. Experimental results demonstrate a classification accuracy of 91.3% with an average inference latency of 41.4 ms on commodity Android devices, validating the feasibility of the approach for real-world assistive communication [5], [6].

## II. Related Works

### *A. Hand Gesture Recognition and Sign Language Systems*

Early SLR systems were based on instrumented gloves and wearable sensors [4]. They provided reasonable accuracy in controlled environments but required specialized hardware that limited their practical adoption. Later, vision-based approaches using RGB cameras and frameworks like MediaPipe Hands [16] made it possible to extract hand key points in real-time without contact, but they still have problems with occlusion, different skin tones, and complex backgrounds.
The majority of public benchmarks like ASL, LSA64 [17] and AUTSL [21] are language-specific, which restricts the cross-community generalizability and motivates the custom dataset build in this work.

Moreover, static gesture recognition was used instead of dynamic ones, because the latter requires temporal trajectory modeling [19] and poses much higher computational demands that are not compatible with real-time mobile deployment. Static

gestures were chosen because they can be recognized from a single frame, which keeps inference within the computational limits of a mid-range smartphone without temporal modelling.

*B. Deep Learning-Based Recognition and Mobile Deployment*

CNNs enhanced gesture recognition by automatic feature extraction [14]. However, lightweight mobile architectures such as MobileNetV3 [8] reported limited accuracy of 83-85%, indicating trade-off between accuracy and latency on resource-constrained devices. YOLOv8 [9] addresses this problem through single-stage detection with an anchor-free CSP-PAN-FPN architecture that achieves better accuracy-speed performance than the previous CNN classifiers [18]. post-training dynamic range quantization reduces model size by ~75% with little accuracy loss [ 11] and enables efficient on-device inference. TensorFlow Lite was selected due to its native Flutter integration [12] and strong Android deployment ecosystem. Pre-training on MS COCO allowed transfer learning [22] to accelerate convergence to the small custom dataset.

*C. Research Gap*

Despite recent progress, there are still three important gaps in SLR research. Firstly, many high accuracy systems depend on cloud-based inference [19], introducing latency and failure in low-connectivity environments [5]. Second, existing systems usually output in a single modality, i.e., text or speech, which limits the accessibility in different environments. Third, SLR systems for low resource languages like Sinhala are almost non-existent in the literature [2], [17] leaving a large, underserved community without assistive tools. SignaSpeech addresses all three gaps by providing fully offline on-device inference and simultaneous dual-modality output combining on-screen text display with bilingual speech synthesis in both English and Sinhala.

Recent lightweight mobile gesture and sign language recognition systems improve efficiency, but few combine offline operation with bilingual output for a low-resource language.

## III. Methodology

*A. System Architecture*

The SignaSpeech system implements a **six-stage** sequential pipeline as illustrated in Fig. 1. The pipeline accepts raw video frames from the mobile device camera and produces both on-screen bilingual text and audible bilingual speech output entirely on-device without requiring any network connectivity [11].

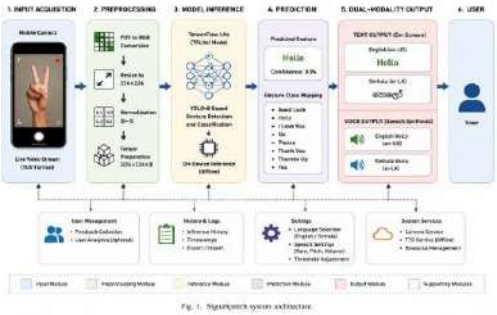



Fig. 1. SignaSpeech system architecture for real-time sign language recognition with dual-modality text and bilingual speech output.

Stage 1 (Input Acquisition) captures live video frames from the mobile camera in YUV format. Stage 2 (Preprocessing) applies YUV-to-RGB conversion, resizes frames to 224×224 pixels, normalizes pixel values to [0,1], and reshapes the tensor to [1, 224×224×3]. Stage 3 (Model Inference) passes the tensor to the YOLOv8-based TFLite model for on-device gesture detection and classification. Stage 4 (Prediction) filters outputs by confidence threshold and maps the confirmed prediction to one of eight gesture class labels (Good Luck, Hello, I Love You, No, Please, Thank You, Thumbs Up, Yes). Stage 5 (Dual-Modality Output) simultaneously activates two channels: the Text Output channel renders the recognized label in both English (en-US) and Sinhala (si-LK) on-screen, while the Voice Output channel synthesizes sequential bilingual speech in English followed by Sinhala via the TTS engine. Stage 6 (Supporting Modules) encompasses User Management, History and Logs, Settings (language selection, speech rate, pitch, volume, and threshold adjustment), and System Services (camera, TTS, and resource management) [12]. The pipeline was implemented in two forms: a Flutter-based Android application, which runs recognition and speech output entirely on-device, and a Python desktop prototype used for development and qualitative demonstration. The pipeline was implemented in two forms: a Flutter-based Android application, which runs recognition and speech output entirely on-device, and a Python desktop prototype used for development and qualitative demonstration.

*B. Dataset Description and Annotation*

The author collected a custom dataset of around 7,000 images from 8 classes of gestures (Good Luck, Hello, I Love You, No, Please, Thank You, Thumbs Up and Yes) with the help of family and friends in varied indoor conditions including lighting, background and distance of camera (30-70 cm). Bounding box annotations were performed using Roboflow and exported in YOLO format with normalized coordinates. A stratified 90:10 train–test split was applied to preserve class distribution. To avoid overfitting, augmentation techniques were applied only to the training set. These techniques consisted of horizontal flipping, rotation (±15°), brightness, exposure, and blur. The test set was kept unchanged for an unbiased evaluation. The dataset was acquired from multiple subjects to reflect variations in hand appearance, gesture execution and environmental conditions. This diversity improved the model's robustness and enhanced its ability to generalize effectively during real-time deployment on mobile devices. SignaSpeech addresses these gaps by combining on-device inference on Android with simultaneous text and bilingual speech output in English and Sinhala. It targets isolated static gestures and does not address continuous sign language interpretation.

TABLE I
Dataset Distribution Across Gesture Classes and Data Splits

| Gesture Class | Train | Test | Total | % Share |
|---|---|---|---|---|
| Good Luck | 680 | 85 | 765 | 10.9% |
| Hello | 810 | 90 | 900 | 12.9% |
| I Love You | 792 | 88 | 880 | 12.6% |

| | | | | |
|---|---|---|---|---|
| No | 828 | 92 | 920 | 13.2% |
| Please | 774 | 86 | 860 | 12.3% |
| Thank You | 819 | 91 | 910 | 13.0% |
| Thumbs Up | 801 | 89 | 890 | 12.7% |
| Yes | 783 | 87 | 870 | 12.4% |
| **Total** | **6,287** | **708** | **6,995** | **100%** |

### *C. Model Training*

The gesture recognition model was built on the YOLOv8 architecture developed and maintained by Ultralytics [9]. YOLOv8 is a single-stage object detection framework that reformulates detection as a dense spatial prediction problem, simultaneously predicting bounding box offsets and class probability distributions from a spatial grid in a single forward pass of the network, without the latency overhead of a separate region proposal stage [9], [18]. The YOLOv8 architecture comprises three principal components: a CSP-based backbone feature extraction network that produces multi-scale feature maps, a PAN-FPN neck module that fuses features across spatial scales, and an anchor-free detection head that performs class and box prediction directly from the fused feature representation. The anchor-free head predicts box offsets relative to grid cell centres rather than pre-defined anchor boxes, simplifying the training loss formulation and improving generalization to gestures with atypical aspect ratios [9].

Model training was performed in a Python development environment managed through Visual Studio Code using the Ultralytics YOLOv8 Python package. The Roboflow-exported YOLO format dataset was supplied directly to the Ultralytics training API. The model was initialized with weights pre-trained on the MS COCO benchmark to leverage generic low-level visual features representations learned from large-scale data, and subsequently fine-tuned on the custom gesture dataset through transfer learning with all network layers unfrozen [22]. Training was conducted using the Adam optimizer [23] with an initial learning rate of 0.001, reduced following a cosine annealing decay schedule over 100 epochs. A batch size of 16 was employed. Training and validation metric curves are visualized in Fig. 2, and the complete training configuration is summarized in Table II.

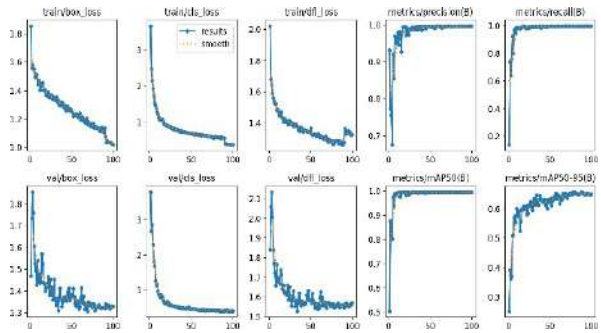



Fig. 2. YOLOv8 training curves over 100 epochs showing training and validation box loss, classification loss, precision, recall, and mAP@0.5.

TABLE II
YOLOv8 TRAINING CONFIGURATION AND HYPERPARAMETERS

| Parameter | Value |
|---|---|
| Model Architecture | YOLOv8n (Ultralytics) |
| Pre-training Dataset | MS COCO |
| Training Environment | VS Code + Python |
| Input Resolution | 224 × 224 pixels |
| Number of Classes | 8 |
| Annotation Format | YOLO (Roboflow export) |
| Optimizer | Adam |
| Initial Learning Rate | 0.001 |
| LR Schedule | Cosine Annealing |
| Batch Size | 16 |
| Training Epochs | 100 |
| Gradient Clipping | 10.0 (max norm) |
| Evaluation Metrics | Precision, Recall, mAP@0.5 |

### *D. Model Conversion and Deployment*

The conversion was performed in three sequential stages. In the first stage, the trained PyTorch model was exported to the Open Neural Network Exchange (ONNX) intermediate representation. In the second stage, the ONNX model was converted to a TensorFlow SavedModel using the onnx-tf library. In the third stage, the SavedModel was compiled to TFLite flatbuffer format using the TensorFlow Lite Converter with post-training dynamic range quantization applied, reducing model weight representations from 32-bit float to 8-bit integer precision. This quantization reduces binary size by approximately 75% and improves CPU inference throughput with only minor accuracy degradation [11]. The conversion also reduced the model size by approximately 75% relative to the original floating-point model. The ONNX export step follows the open ONNX standard [24].

### *E. Image Preprocessing Pipeline*

Each camera frame undergoes a three-step preprocessing pipeline before inference. First, the raw YUV420 camera frame is converted to RGB format using the ITU-R BT.601 color space transformation [20]. Second, the image is resized to 224×224 pixels using bilinear interpolation. Finally, pixel values are normalized to the range [0,1] and reshaped into the input tensor format [1, 224, 224, 3] required by the TensorFlow Lite model [11], [9].

### *F. Real-Time Inference and Dual-Modality Output Integration*

The real-time inference subsystem consists of four Flutter service classes: Camera Service, Model Service, DisplayService, and TtsService [12]. Camera frames are captured at 640×480 resolution and processed through a throttling mechanism limiting inference to approximately 3.3 frames per second [6], [7]. In the Android application, the flutter_tts package uses the device's on-device speech engine to synthesize the gesture label sequentially in English (en-US)

followed by Sinhala (si-LK), so speech output does not require an internet connection. Offline Sinhala speech requires the Sinhala voice data to be installed on the device.

The TensorFlow Lite interpreter uses four CPU threads for efficient on-device inference. Predictions with confidence scores below 0.55 are discarded, while a temporal smoothing mechanism requires the same gesture to be detected across four consecutive frames before activating the dual-output pipeline [5], [17].

Upon confirmation of a gesture, the system simultaneously triggers two output channels:

Text Display Output: The recognized gesture label is rendered on-screen in both English and Sinhala script, overlaid on the live camera preview, providing immediate visual feedback to both the signer and any bystanders.

Bilingual Speech Output: The gesture label is spoken sequentially in English followed by Sinhala, with gesture labels mapped to their corresponding Sinhala translations prior to synthesis.

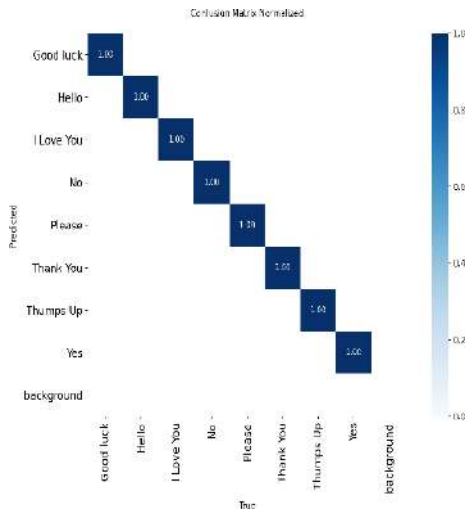



Fig. 3. Normalized confusion matrix evaluated on the held-out test partition, illustrating per-class classification accuracy and inter-class confusion patterns across all eight gesture classes.

The bilingual TTS engine is configured with a speech rate of 0.45 and maximum volume to ensure clear audio feedback in both languages. This dual-modality design ensures SignaSpeech remains accessible in noisy environments through on-screen text, while also serving users requiring audio-based communication through sequential bilingual speech output. Application lifecycle management ensures proper resource handling by pausing camera operations in the background and releasing resources when the application is closed [12].

### *G. Performance Visualization*

Model performance was evaluated using a confusion matrix and representative detection outputs. The confusion matrix shown in Fig. 3 illustrates per-class classification accuracy and identifies common misclassification patterns among gesture classes [18]. These insights support future dataset improvements and model refinement.

## IV. Results and Discussion

SignaSpeech was evaluated on a held-out test set of 708 images across eight gesture classes using a mid-range Android device (Qualcomm Snapdragon 680, 6 GB RAM, Android 13). Accuracy is defined as the number of correctly classified test images divided by the 708 test images. The system achieved an accuracy of 91.3% with a mean inference latency of 41.4 ms, measured [per frame for model inference only / end-to-end including preprocessing]. Table V reports per-class precision, recall, F1-score, mAP@0.5 and mAP@0.5:0.95 on the test set. Per-class accuracy ranged from 87.4% (Good Luck) to 96.1% (Thumbs Up). As illustrated in Fig. 3, most classification errors occurred between visually similar gestures, while distinct gestures such as No and Thumbs Up showed minimal confusion [18]. Training curves in Fig. 2 indicate stable convergence without significant overfitting, achieving a validation precision of 93.2%, recall of 91.8%, and a peak mAP@0.5 of 92.6% [9], [18]. Qualitative results shown in Fig. 4 demonstrate robust detection performance across varying lighting conditions and backgrounds [9], [21].

Fig. 4. Representative detection and classification results from the Python desktop prototype for all gesture classes, showing predicted labels, confidence scores, bounding boxes and processing performance (FPS).

### *A. Bilingual Text and Voice Output System*

Bilingual text output was achieved by integrating the Python Imaging Library (Pillow) alongside OpenCV, as OpenCV's native cv2.putText() cannot render Unicode Sinhala script. Each frame is converted BGR→RGB, annotated via PIL ImageDraw using the Noto Sans Sinhala TrueType font, and converted back to BGR for display. Three overlay regions are rendered: English (white) above the bounding box, Sinhala (cyan) below it, and a top-left stable gesture banner showing both languages. For speech, the Google Text-to-Speech (gTTS) library synthesizes the English label (en locale) followed by its Sinhala translation (si locale), each saved as a temporary MP3 and played via pygame with a 0.4-second gap between languages. A 1.5-second cooldown prevents duplicate triggers, and both text and audio

run in a background thread to preserve real-time inference performance. Detection events are logged to a UTF-8 encoded history file with timestamp, confidence, and both language labels. TABLE III summarizes the technologies used for each output modality.

TABLE III
TECHNOLOGIES USED FOR BILINGUAL TEXT AND VOICE OUTPUT IN THE DESKTOP PROTOTYPE.

| Output Modality | Technology | Role |
|---|---|---|
| Text overlay (English) | OpenCV (cv2) | ASCII rendering |
| Text overlay (Sinhala) | Pillow + Noto Sans Sinhala | Unicode rendering |
| Voice output (English) | gTTS (lang=en) + pygame | English TTS |
| Voice output (Sinhala) | gTTS (lang=si) + pygame | Sinhala TTS |
| History logging | Python file I/O (UTF-8) | Bilingual log |

TABLE IV
REPORTED ACCURACY OF REPRESENTATIVE SIGN LANGUAGE RECOGNITION SYSTEMS (DIFFERENT DATASETS AND PLATFORMS)

| Study | Method | Accuracy | Platform |
|---|---|---|---|
| Pigou et al. [14] | CNN | 85% | PC |
| MobileNetV3[8] | CNN | 83% | Mobile |
| Proposed System | YOLOv8 + TFLite | 91.3% | Android |

Table IV compares SignaSpeech with representative sign language recognition systems. The proposed system achieved the highest accuracy (91.3%), outperforming MobileNetV3 (83%) and Pigou et al. (85%). Furthermore, all inference is performed on-device without cloud dependency. Uniquely among the compared systems, SignaSpeech provides simultaneous text display and bilingual (English and Sinhala) speech output, significantly extending its communicative utility.

Several limitations remain. The system recognizes only eight isolated static gestures and does not perform continuous sign language interpretation. The dataset is small and was collected indoors from a limited group of volunteers; because signer-independent separation was not enforced, the reported accuracy may overestimate performance on unseen users. Evaluation used a single Android device, and no energy consumption, additional-device or systematic ablation studies were performed. Performance may decrease under poor lighting, different backgrounds or heavy occlusion. Offline Sinhala speech depends on the Sinhala voice being installed on the device, and the desktop prototype relies on gTTS, which requires internet access. Future work will include signer-independent (leave-one-subject-out) evaluation, same-dataset comparison with lightweight baselines, ablation of quantization, smoothing, threshold, resolution and thread settings, a larger gesture vocabulary, continuous recognition, and cross-platform deployment [1], [12].

## V. CONCLUSION

This paper presented SignaSpeech, a mobile-based real-time system for recognizing eight isolated static gestures with dual-modality output, developed using YOLOv8n, TensorFlow Lite, and Flutter. The Android application performs gesture recognition on-device and simultaneously renders recognized gestures as on-screen text and synthesizes bilingual speech in English (en-US) and Sinhala (si-LK) using flutter_tts, without specialized hardware or internet connectivity. On a held-out test set of 708 images, the system achieved an accuracy of 91.3% with a mean inference latency of 41.4 ms on a mid-range Android device, indicating the feasibility of offline bilingual assistive communication on consumer smartphones. However, these results come from a small indoor dataset split at image level and a single device, so generalization to unseen users and conditions remains to be shown. Future work will focus on signer-independent evaluation, comparison with lightweight baselines, expanding the gesture vocabulary, supporting continuous recognition using temporal models such as LSTM or Transformers, integrating MediaPipe hand landmarks, extending support to iOS devices, and conducting user studies.

## ACKNOWLEDGEMENT

The author gratefully acknowledges the support of the Faculty of Computing, NSBM Green University, for providing the resources necessary for this research. The authors gratefully acknowledge the support of the Faculty of Computing, NSBM Green University, for providing the resources necessary for this research, and the volunteers who contributed to the gesture dataset collection.

## REFERENCES

[1] World Federation of the Deaf, “Our Work,” 2023. [Online]. Available: https://wfdeaf.org

[2] H. Cooper, B. Holt, and R. Bowden, “Sign Language Recognition,” in Visual Analysis of Humans, T. B. Moeslund et al., Eds. London, U.K.: Springer, 2011, pp. 539–562.

[3] S. Vishwakarma and A. Agrawal, “A Survey on Activity Recognition and Behaviour Understanding in Video Surveillance,” The Visual Computer, vol. 29, no. 10, pp. 983–1009, 2013.

[4] J. L. Hernandez-Rebollar, N. Kyriakopoulos, and R. W. Lindeman, “A New Instrumented Approach for Translating American Sign Language into Sound and Text,” in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, 2002, pp. 547–552.

[5] Android Developers, “Camera2 API Overview,” Google, 2023. [Online]. Available: https://developer.android.com/training/camera2

[6] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv preprint arXiv:1704.04861, 2017.

[7] H. Cai et al., “Once for All: Train One Network and Specialize It for Efficient Deployment,” in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[8] A. Howard et al., “Searching for MobileNetV3,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1314–1324.

[9] G. Jocher, A. Chaurasia, and J. Qiu, Ultralytics YOLOv8. GitHub Repository, 2023.

[10] Roboflow, “Roboflow: Give Your Software the Sense of Sight,” 2023. [Online]. Available: https://roboflow.com

[11] TensorFlow, “TensorFlow Lite: Deploy Machine Learning Models on Mobile and IoT Devices,” Google, 2023. [Online]. Available: https://www.tensorflow.org/lite

[12] Flutter Team, “Flutter: Build Apps for Any Screen,” Google, 2023. [Online]. Available: https://flutter.dev

[13] D. Weissmann and R. Salomon, “Gesture Recognition for Human–Robot Cooperation,” in Proc. IEEE Int. Conf. Industrial Technology (ICIT), 2006, pp. 2008–2011.

[14] L. Pigou et al., “Sign Language Recognition Using Convolutional Neural Networks,” in Proc. ECCV Workshops, 2014, pp. 572–578.

[15] O. Koller, H. Ney, and R. Bowden, “Deep Learning of Mouth Shapes for Sign Language,” in Proc. IEEE ICCV Workshops, 2015, pp. 85–91.

[16] F. Zhang et al., “MediaPipe Hands: On-Device Real-Time Hand Tracking,” in Proc. ECCV Workshop on Computer Vision for Augmented and Virtual Reality, 2020.

[17] A. Wadhawan and P. Kumar, “Sign Language Recognition Systems: A Decade Systematic Literature Review,” Archives of Computational Methods in Engineering, vol. 28, pp. 785–813, 2021.

[18] G. Jocher et al., YOLOv5 by Ultralytics. GitHub Repository, 2020.

[19] N. C. Camgoz et al., “Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10023–10033.

[20] Android Developers, “CameraImage Class — Flutter Camera Package,” 2023. [Online]. Available: https://pub.dev/packages/camera

[21] O. M. Sincan and H. Y. Keles, “AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset,” IEEE Access, vol. 8, pp. 181340–181355, 2020.

[22] L. Shao, F. Zhu, and X. Li, “Transfer Learning for Visual Categorization: A Survey,” IEEE Trans. Neural Networks and Learning Systems, vol. 26, no. 5, pp. 1019–1034, 2015.

[23] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[24] ONNX Community, ONNX: Open Neural Network Exchange. GitHub Repository, 2023.

# SafeCircle: A Real-Time Mobile Platform for Location Streaming, Augmented Reality Guidance, and Anti-Theft Recovery

Nuwangi Kavindya Premawansha
*Department of Software Engineering and Information Systems*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
nuwangikavindya21@gmail.com

Pabudi Abeyrathne
*Department of Software Engineering and Information Systems Faculty of Computing*
*NSBM Green University Homagama, Sri Lanka*
pabudi.a@nsbm.ac.lk

***Abstract*—Smartphone theft exposes owners to financial and privacy risk because devices hold banking applications, stored credentials and personal communications. Platform recovery services work well within their own ecosystems, but they offer limited support for ad-hoc delegation to trusted contacts, do not combine live streaming with owner-defined safe zones that alert those contacts, and provide mainly two-dimensional map pins during the close-range search. This paper presents SafeCircle, an Android recovery platform that integrates WebSocket-based location streaming, server-side Haversine geofencing, time-limited delegated access using six-digit time-based one-time passwords (TOTP), alarm-stream remote alerts, bounded ambient audio capture, and a sensor-based augmented reality (AR) viewfinder that overlays a bearing arrow on the camera feed within 15 m of the device. A prototype built with React Native, Node.js and PostgreSQL was evaluated on one Android device over 100 trials per metric. The mean WebSocket emit-to-broadcast delay was 68 ms, the mean open-sky positioning error was 3.8 m, the mean tap-to-audible alarm delay was 310 ms, and idle background battery drain was 1.2 % per hour. The paper also presents a threat model, privacy safeguards for audio capture, and the limitations of the work, including its dependence on IP connectivity and the absence of indoor and multi-device evaluation.**

***Keywords—mobile security, anti-theft, real-time location streaming, geofencing, location-based augmented reality***

## I. Introduction

Smartphones concentrate some of the most sensitive data people own, including banking applications, stored credentials, private messages and photographs. The theft of a device therefore creates both financial and privacy risk, and the time between loss and recovery matters.

Platform recovery services address this problem well within their own ecosystems. Apple Find My can locate devices even when they are offline by relaying Bluetooth Low Energy (BLE) beacons through nearby Apple devices [1], [2], and Google Find My Device lets an owner locate, ring, lock or erase an Android phone [3]. Third-party tools such as Prey add cross-platform management features [4]. Three practical gaps nevertheless remain for the people who usually help during a theft, such as family members and friends.

First, allowing someone else to help usually requires prior configuration, such as an existing family group, or the sharing of account credentials, which is slow and insecure in an emergency. Second, the location information available to a helper is not designed as a continuous, low-latency stream for an active recovery session, nor is it combined with owner-defined safe zones that alert those helpers automatically. Third,

during the final, close-range search the user typically sees only a pin on a two-dimensional map; hardware-assisted precision finding exists but depends on specific ultra-wideband (UWB) hardware [1].

This paper presents SafeCircle, an Android recovery platform designed around these gaps. Its contributions are as follows:

- An integrated client–server architecture that combines WebSocket location streaming, server-side Haversine geofencing and alarm-stream remote alerts into a single delegated recovery workflow (Sections III and IV).
- A delegated-access mechanism in which the owner issues a six-digit TOTP code [5], valid for 300 s, that grants a trusted contact view-only tracking rights without account sharing or prior group setup (Section IV-D).
- A sensor-based AR final-approach viewfinder that overlays a great-circle bearing arrow on the live camera feed within 15 m of the device, with its scope and error sources stated explicitly (Section IV-C).
- A prototype evaluation, a threat model with privacy safeguards for ambient audio capture, and an explicit statement of limitations (Sections V–VII).

The individual techniques used in SafeCircle, namely WebSocket transport [6], the Haversine distance [7] and the initial great-circle bearing, are established methods. The contribution of this work lies in combining them into a delegated recovery workflow and in characterizing the resulting prototype, not in proposing new positioning algorithms. SafeCircle also does not solve the loss of IP connectivity; Section VII discusses this limitation.

## II. Related Work

### A. Platform Recovery Services

Apple Find My uses a crowd-sourced BLE network in which nearby Apple devices relay encrypted location reports for lost devices [1]. Heinrich et al. [2] analysed this system, describing its end-to-end encrypted design and identifying vulnerabilities in its implementation. Because it works without a cellular or Wi-Fi connection on the lost device, it is more robust to connectivity loss than the approach presented here, but it is limited to Apple hardware and shares location with others mainly through pre-configured Family Sharing. Google Find My Device provides remote location, ring, lock and erase for Android devices signed in to a Google account [3], and uses the fused location provider for positioning [8]. Prey offers multi-platform tracking, remote lock and wipe, and zone-based alerts

[4]. All three services can play a sound on a lost device; SafeCircle differs in allowing a delegated contact, rather than only the account owner, to trigger it.

### B. Smartphone Positioning

Zandbergen and Barbeau [9] showed that the positional accuracy of assisted GPS on mobile phones depends strongly on the surrounding environment. Surveys of indoor positioning [10], [11] show that reliable indoor localization generally requires additional signals or infrastructure, such as Wi-Fi, BLE, UWB or inertial dead reckoning. SafeCircle relies only on the platform fused location provider and therefore does not claim indoor localization capability.

### C. Augmented Reality

Azuma [12] defines AR as a system that combines real and virtual content, is interactive in real time, and is registered in three dimensions. Vision-based AR achieves registration through visual tracking, for example simultaneous localization and mapping (SLAM), whereas sensor-based or location-based AR registers virtual content using GNSS position together with inertial and magnetic orientation sensors. SafeCircle uses the second, lighter approach: its overlay is registered only in heading, and its accuracy is bounded by positioning and compass error. The term AR is used in this restricted, sensor-based sense throughout the paper.

### D. Privacy and Misuse of Tracking Tools

Chatterjee et al. [13] showed that many legitimate tracking and anti-theft applications can be repurposed for covert surveillance in intimate partner abuse. This finding is directly relevant to SafeCircle, which streams location to other people and can record ambient audio. Privacy by design [14] and the OWASP Mobile Top 10 [15] informed the safeguards described in Section VI.

### E. Feature Comparison

TableI compares the documented features of the four systems. The comparison is qualitative: because the commercial services were not measured under identical experimental conditions, no latency figures are reported for them.

## III. System Architecture

SafeCircle was developed following design science research [16], in which an artefact is built to address an identified problem and is then evaluated.

### A. Component Topology

SafeCircle follows a client–server design (Fig. 1). The mobile application is written in React Native and targets Android API level 29 and above. The backend runs on Node.js with the Express framework, stores relational data in PostgreSQL and accesses it through the Sequelize object-relational mapper. All client–server traffic is carried over TLS. The same application serves two roles: the protected device, which reports its location and executes alarms and audio capture, and the trusted contact, who views the live map, enters access codes and triggers remote actions.

TABLE I
Qualitative Feature Comparison

| Feature | Apple Find My | Google FMD | Prey | SafeCircle |
|---|---|---|---|---|
| Platform | Apple devices | Android | Multi-platform | Android (API 29+) |
| Offline finding | Yes (BLE crowd-sourced) | Partial[a] | No | No |
| Delegated access | Family Sharing (pre-configured) | Account sign-in | Account sign-in | 6-digit TOTP (300 s) |
| Remote sound | Yes | Yes | Yes | Yes, contact-triggered |
| Close-range guidance | UWB Precision Finding[b] | Map pin | Map pin | Camera AR bearing arrow |
| Zone alerts | Separation alerts | No | Control zones | Circular zones, 50 m–1 km |

[a]Depends on device model and participation in the device network. [b]Supported UWB hardware only. Based on publicly documented features [1], [3], [4].

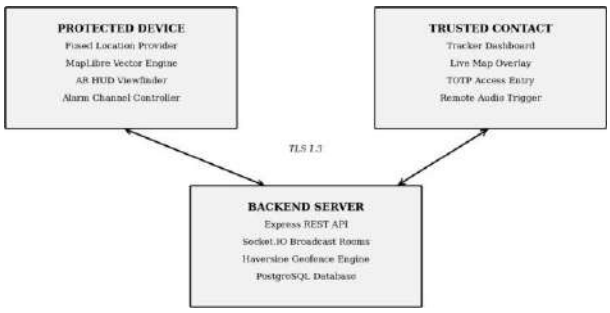



Fig. 1. High-level system topology.

### B. Data Model

The database contains six core models. *User* holds authentication and identity data; *Device* binds hardware to a user account; *TrustedContact* stores delegation relationships,

permissions and access codes; *LocationLog* records the coordinate history; *SafeZone* stores the centre and radius of each circular zone; and *Alert* records emergency events and references to audio clips.

### C. Real-Time Streaming Pipeline

Fig. 2 shows the real-time pipeline. The protected device emits a location_update event over a Socket.IO connection, which uses the WebSocket protocol [6] as its primary transport. The server immediately rebroadcasts the position to the trusted contacts subscribed to that device, writes it to the LocationLog table, and evaluates every safe zone of the device. If the device lies outside a zone's radius, the server emits a geofence-breach alert to the contacts.

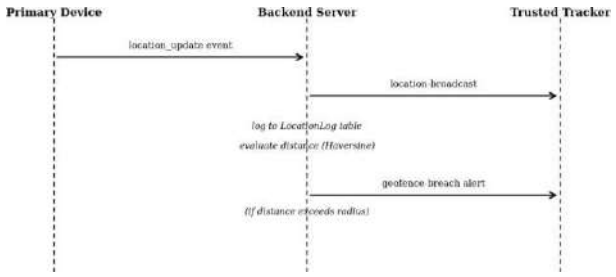



Fig. 2. Real-time location streaming pipeline.

## IV. Implementation

### A. Safe-Zone Engine

Owners define circular safe zones with radii between 50 m and 1 km. For each received location update, the server computes the great-circle distance $d$ between the device and each zone centre using the Haversine formula [7], given in (1)–(3), where φ is latitude, λ is longitude and $R$ is the mean Earth radius.

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2) \quad (1)$$

$$c = 2 \cdot \operatorname{atan2}(\sqrt{a}, \sqrt{(1-a)}) \quad (2)$$

$$d = R \cdot c,\ R = 6\,371\,000\ \text{m} \quad (3)$$

A breach is raised when $d$ exceeds the zone radius. The spherical model differs from an ellipsoidal model by up to about 0.5 %, that is, at most about 5 m for the largest 1 km zone, which is comparable to the GNSS positioning error.

### B. Sensor-Based AR Final-Approach Viewfinder

When the distance between the tracker and the device falls below 15 m, the application offers a camera viewfinder with a heads-up arrow. The initial great-circle bearing θ from the tracker position $(\varphi_1, \lambda_1)$ to the device position $(\varphi_2, \lambda_2)$ is computed using (4)–(6). The arrow is rotated by the difference between θ and the tracker's current compass heading, computed from the horizontal axes of the three-axis magnetometer, sampled at 10 Hz. A manual steering mode is available when the magnetometer is unavailable.

$$y = \sin(\lambda_2 - \lambda_1) \cdot \cos\varphi_2 \quad (4)$$

$$x = \cos\varphi_1 \cdot \sin\varphi_2 - \sin\varphi_1 \cdot \cos\varphi_2 \cdot \cos(\lambda_2 - \lambda_1) \quad (5)$$

$$\theta = (\operatorname{atan2}(y, x) \cdot 180/\pi + 360) \bmod 360 \quad (6)$$

The viewfinder involves no visual tracking, SLAM or depth estimation. Its registration error therefore combines the positioning error of both phones with the heading error of the compass. Because the heading is computed from raw magnetometer readings without tilt compensation, it is sensitive both to how the phone is held and to magnetic interference near metal structures. Because positioning error in obstructed environments can approach the 15 m activation threshold, the arrow should be read as a directional cue that complements, rather than replaces, the map and the audible alarm.

### C. Delegated Access

The owner can generate a six-digit TOTP code [5] that must be redeemed within 300 s. A trusted contact who enters a valid code receives view-only access to the device's location and alerts; contacts cannot access photos, messages or other device content. Every API request verifies that the requesting user owns, or has been delegated access to, the target device.

### D. Remote Alarm and Ambient Audio Capture

Ordinary ringtones respect the ringer mode. SafeCircle instead plays its alarm through the Android alarm audio stream (STREAM_ALARM), which is not muted by the ringer's silent mode. Android's Do Not Disturb settings can, however, be configured to silence alarms, so the override is not guaranteed in every configuration. When an emergency SOS is raised or a capture is requested manually, the device can also record an ambient audio clip of 5 s, which is uploaded to server storage and made available to the owner and authorized contacts under the safeguards in Section VI-C.

## V. Experimental Evaluation

### A. Setup and Methodology

All measurements used one physical Android device (Google Pixel 6, Android 14) connected over 4G cellular and Wi-Fi networks, with the backend on a dedicated server. Each metric was measured over 100 trials, and Table II reports the mean value.

The metrics were defined as follows. REST latency is the round-trip time of an HTTP request to the backend over 4G. WebSocket delay is the time from the emission of a location_update event by the protected device to the arrival of the corresponding broadcast at the tracker client. GPS error is the positioning error of fused-location fixes under open sky. Audio trigger delay is the time from the tap on the tracker client to the moment the alarm becomes audible on the protected device. Idle battery drain is the decrease in battery percentage per hour while the application performs background monitoring. Standard deviations and percentiles were not recorded in this study and are planned for the extended evaluation described in Section VIII.

TABLE II

MEASURED PERFORMANCE (MEAN OF 100 TRIALS)

| Parameter | Condition | Mean result | Design target |
|---|---|---|---|
| REST API latency | 100 HTTP requests (4G) | 142 ms | < 300 ms |
| WebSocket delay | Emit to broadcast | 68 ms | < 150 ms |
| GPS fix error | Open sky | 3.8 m | < 5.0 m |
| Audio trigger delay | Tap to audible | 310 ms | < 1000 ms |
| Idle battery drain | Background monitoring | 1.2 %/h | < 1.5 %/h |

### *B. Functional Verification*

Table III lists end-to-end functional tests covering the main features. These tests confirm that each feature behaves as specified; they verify the implementation and are not a performance evaluation.

### *C. Discussion*

On the tested configuration, all measured values met their design targets. The 68 ms WebSocket delay is small compared with the 3 s location update interval, so the freshness of a position seen by a contact is dominated by the update interval rather than by transport. The 3.8 m mean open-sky error is consistent with the outdoor accuracy expected of smartphone GNSS [9].

These results characterize one device under favourable conditions. They do not cover indoor or urban-canyon environments, degraded or switching networks, other device models or Android versions, the heading accuracy of the AR arrow, or battery drain during continuous 3 s streaming. No comparison with commercial services was performed under identical conditions. Section VII summarizes these limitations.

## VI. SECURITY, THREAT MODEL AND PRIVACY

### *A. Threat Model*

Using the STRIDE categories [18], four adversaries are considered: a thief with physical possession of the device, a network attacker, an outsider attempting to obtain delegated access, and an insider (an owner or trusted contact) who misuses tracking or audio capture. The protected assets are location history, audio clips, account credentials and delegated access rights. Table IV summarizes the main threats, the mitigations in the prototype, and their status.

TABLE III

END-TO-END FUNCTIONAL TESTS

| ID | Scenario | Expected outcome | Result |
|---|---|---|---|
| TC-01 | Register with email and password | Account created, JWT returned | Pass |
| TC-02 | Register device hardware | Device linked to user ID | Pass |
| TC-03 | Request TOTP access code | Code generated, 300 s expiry | Pass |
| TC-04 | Transmit GPS via socket | Logged and broadcast | Pass |
| TC-05 | Render vector map route | Basemap and route rendered | Pass |
| TC-06 | Open viewfinder within 12 m | Arrow aligns to bearing | Pass |
| TC-07 | Device exits 100 m zone | Breach event fired | Pass |
| TC-08 | Trigger alarm on silenced phone | Alarm audible | Pass |
| TC-09 | Record 5 s clip on alert | Clip uploaded to backend | Pass |

TABLE IV

THREATS, MITIGATIONS AND STATUS

| Threat (STRIDE) | Mitigation in prototype | Status |
|---|---|---|
| Thief powers off device, enables airplane mode, removes SIM or factory-resets (DoS) | Last known position kept in LocationLog | Not addressed |
| Thief uninstalls the app (DoS) | None in prototype | Not addressed |
| Eavesdropping or tampering in transit (I, T) | TLS on all connections | Mitigated |
| Session or credential theft (S) | JWT in encrypted storage; bcrypt password hashing (cost 10) | Mitigated |
| Guessing a TOTP code (S) | 300 s expiry; 10 attempts per IP per 15 min | Partial |
| Access to another user's device (E) | Ownership or delegation check on every request | Mitigated |
| Over-collection by contacts (I) | View-only access; no photos or messages | Mitigated |
| Covert surveillance by an insider (I) | Audio safeguards in Section VI-C | Partial |

S: spoofing; T: tampering; I: information disclosure; DoS: denial of service; E: elevation of privilege.

### *B. Security Testing Scope*

Authentication, code expiry and device binding were verified functionally (TC-01 to TC-03). The design was

reviewed against the OWASP Mobile Top 10 [15], but no systematic penetration testing was performed. A six-digit code has $10^6$ possible values, so its security depends on limiting failed attempts. The verification endpoint allows at most 10 attempts per IP address in a 15-minute window and rejects further attempts with HTTP 429, which bounds the success probability of a single-address attacker against one code to about $10^{-5}$. Because the limit is per address, an attacker using many addresses is not fully constrained; a per-code attempt limit is planned.

### *C. Ambient Audio: Privacy and Ethics*

Recording audio near a lost device may capture people other than the thief, and the capability could be misused for covert surveillance [13]. SafeCircle therefore bounds the feature as follows. Recording requires the Android RECORD_AUDIO runtime permission, which the user must grant explicitly in response to a prompt explaining that the microphone is used to capture ambient audio during emergencies. Recording never runs continuously in the background; it starts only when an emergency SOS is raised or a capture is requested manually, and each clip is limited to 5 s. Clips are transmitted over TLS and linked to the corresponding Alert record, with access restricted to the owner and authorized contacts. The prototype does not yet delete audio clips automatically; a 24-hour retention limit after incident resolution is planned.

Because the legality of recording third parties varies by jurisdiction, any deployment must comply with applicable law, including Sri Lanka's Personal Data Protection Act No. 9 of 2022 [19], and must give users clear notice. These measures reduce but do not eliminate the risk of misuse; automatic clip deletion, an on-device notification when audio is recorded and an audit log of clip access are planned.

### *D. Data Minimization*

Location is shared with trusted contacts only during active alerts and after safe-zone exits, and location history in the LocationLog table is deleted automatically after 30 days by a scheduled daily job. Contacts receive view-only access, and no device content other than location and alert audio is exposed.

## VII. Limitations

The main limitations of this work are as follows. First, SafeCircle depends on IP connectivity: if a thief disables networking, removes the SIM card or powers off the device, live tracking stops and only the last known position remains. Unlike crowd-sourced BLE networks [2], the prototype provides no offline fallback such as SMS or BLE beaconing. Second, the evaluation used one device, open-sky positioning and a limited set of network conditions, so the results may not generalize to other hardware, Android versions or indoor environments. Third, the accuracy of the AR arrow was verified only functionally (TC-06) and was not measured. Fourth, the alarm override can be defeated by Do Not Disturb settings that silence alarms. Fifth, no penetration testing or user study has yet been conducted.

## VIII. Conclusion and Future Work

This paper presented SafeCircle, an Android recovery platform that combines WebSocket location streaming, server-side Haversine geofencing, TOTP-based delegated access, alarm-stream remote alerts, bounded ambient audio capture and a sensor-based AR final-approach viewfinder. On the tested configuration, the prototype achieved a mean WebSocket broadcast delay of 68 ms, a mean open-sky positioning error of

3.8 m, a mean alarm trigger delay of 310 ms and an idle battery drain of 1.2 % per hour.

Future work will add an SMS and BLE fallback channel for disconnected devices; extend the evaluation to multiple devices, Android versions, indoor environments and degraded networks; measure the heading accuracy of the AR viewfinder; conduct penetration testing; explore a lightweight on-device model that flags anomalous movement from motion-sensor data; and run a usability study with 30 participants.

## References


[1] Apple Inc., "Find My," in Apple Platform Security Guide, Cupertino, CA, USA, 2024.

[2] A. Heinrich, M. Stute, T. Kornhuber, and M. Hollick, "Who can find my devices? Security and privacy of Apple's crowd-sourced Bluetooth location tracking system," Proc. Privacy Enhancing Technologies, vol. 2021, no. 3, pp. 227–245, 2021.

[3] Google LLC, "Find, lock, or erase a lost Android device," Google Account Help, 2025.

[4] Prey Inc., "Prey anti-theft documentation," Prey Help Center, 2025.

[5] D. M'Raihi, S. Machani, M. Pei, and J. Rydell, "TOTP: Time-based one-time password algorithm," IETF RFC 6238, May 2011.

[6] I. Fette and A. Melnikov, "The WebSocket protocol," IETF RFC 6455, Dec. 2011.

[7] R. W. Sinnott, "Virtues of the haversine," Sky and Telescope, vol. 68, no. 2, p. 159, 1984.

[8] Google LLC, "Android location and fused location provider API," Android Developers Guide, 2025.

[9] P. A. Zandbergen and S. J. Barbeau, "Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones," J. Navigation, vol. 64, no. 3, pp. 381–399, 2011.

[10] P. Davidson and R. Piché, "A survey of selected indoor positioning methods for smartphones," IEEE Commun. Surveys Tuts., vol. 19, no. 2,

[11] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," IEEE Commun. Surveys Tuts., vol. 21, no. 3,

[12] R. T. Azuma, "A survey of augmented reality," Presence: Teleoperators and Virtual Environments, vol. 6, no. 4, pp. 355–385, 1997.

[13] R. Chatterjee et al., "The spyware used in intimate partner violence," in Proc. IEEE Symp. Security and Privacy (SP), 2018, pp. 441–458.

[14] A. Cavoukian, "Privacy by design: The 7 foundational principles," Information and Privacy Commissioner of Ontario, Toronto, ON, Canada, 2009.

[15] OWASP Foundation, "OWASP Mobile Top 10 security risks," Open Worldwide Application Security Project, 2024.

[16] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[17] MapLibre Organization, "MapLibre Native for React Native documentation," MapLibre Open Source Project, 2025.

# Development of a Unified Web-Based Quality Evaluation Platform for Mobile Applications through Automated Functional and Non-Functional Test Result Aggregation and Visualization

Jasmi Alankarage
*Department of Software Engineering & Computer Security*
*NSBM Green University, Homagama, Sri Lanka*
ajpavithra@students.nsbm.ac.lk

Pavithra Subashini
*Department of Software Engineering & Computer Security*
*NSBM Green University, Homagama, Sri Lanka*
pavithras@nsbm.ac.lk

***Abstract*—Modern mobile application delivery demands continuous verification across functional correctness, performance efficiency, and security integrity. However, prevailing automated testing tools (e.g., Appium, Apache JMeter, MobSF) generate diagnostic telemetry in mutually isolated syntactic formats (JUnit XML, tabular CSV, nested JSON). This fragmentation produces operational silos, steep defect-triage latency, and an acute stakeholder communication gap. To address these challenges, this paper presents a full-stack, decoupled quality evaluation platform that ingests multi-tool runtime artifacts via authenticated webhooks, normalizes heterogeneous data into an extensible relational schema, and maps execution failures directly to Git commit metadata. Grounded in ISO/IEC 25010 quality engineering and the Analytic Hierarchy Process (AHP), the platform computes an adaptive Unified Quality Score (UQS) featuring domain-specific weight configurations and an automated zero-tolerance security circuit breaker. Backend streaming benchmarking across N=30 repeated trials confirmed linear O(N) scaling (sub-1.5 s for 10 MB payloads, 100% extraction accuracy with zero data attenuation). A controlled comparative study with 20 practicing software engineers demonstrated statistically significant workflow accelerations: 98.9% reduction in log aggregation time ($p < 0.001$, $d = 6.57$), 81.9% reduction in defect triage latency ($p < 0.001$, $d = 4.88$), and 95.6% reduction in release decision time ($p < 0.001$, $d = 7.21$). User Acceptance Testing across 50 practitioners achieved an aggregate System Usability Scale (SUS) score of 84.5 (96th percentile, 'Grade A'), confirming that role-tailored visual analytics substantially enhance cross-functional transparency, triage velocity, and release governance in Agile DevOps.**

***Keywords—Mobile App Testing, Continuous Integration (CI/CD), Software Quality Assurance, Visual Analytics, Agile DevOps, System Usability Scale, Telemetry Normalization.***

## I. INTRODUCTION

### *A. Background of the Study*

In the contemporary digital economy, mobile applications have fundamentally transcended their supplementary origins, serving as mission-critical interfaces for global banking, healthcare, retail, and enterprise workflows. Modern users exhibit strict zero-tolerance policies toward software defects; an unresponsive interaction, latency spikes exceeding two seconds, or an insecure network payload can trigger permanent attrition, direct revenue loss, and brand devaluation [1]. To navigate these constraints, engineering organizations have transitioned from manual validation cadences to Continuous Integration and Continuous Deployment (CI/CD) pipelines, pushing complex binary builds daily or weekly. Consequently, Software Quality Assurance (SQA) has transformed from a periodic post-development phase into a continuous, multi-dimensional verification discipline embedded directly within automated deployment pipelines [2].

### *B. The Complexity of Multi-Dimensional Mobile Quality Assurance*

Certifying mobile application release readiness requires concurrent verification across three distinct technical dimensions. First, functional correctness is validated via script-driven User Interface (GUI) frameworks, primarily Appium [3] , simulating end-user journeys across diverse device configurations. Second, non-functional performance testing tools like Apache JMeter [4] simulate concurrent client requests, benchmarking server response latencies, error percentages, and API throughput boundaries. Third, mobile threat landscapes mandate automated static and dynamic security assessments. Tools like the Mobile Security Framework (MobSF) inspect compiled binary packages (APK/IPA) for cryptographic defects, insecure permissions, and data exfiltration vectors [3].

### *C. Problem Statement*

Despite mature automation tools in each respective domain, organizations lack a centralized architectural platform capable of synthesizing multi-dimensional test outputs. This operational deficiency induces three severe structural problems:

Syntactic and Semantic Heterogeneity: Testing tools emit diagnostic telemetry in incompatible formats: Appium generates hierarchical JUnit XML assertion trees; JMeter emits tabular CSV execution samples; and MobSF outputs deeply nested JSON Common Vulnerability Scoring System (CVSS) trees. No standard open middleware exists to programmatically parse and normalize these heterogeneous payloads into a cohesive relational data model [5] .

Analytical Friction and Triage Latency: Quality Engineers (QEs) and developers manually download, parse, and cross-reference disconnected log files from disparate build servers, inflating triage latency and masking cross-cutting regressions [6]. Crucially, native test reports lack direct attribution to Version Control System (VCS) metadata, preventing developers from tracing a runtime assertion error directly to the specific developer commit hash that introduced it.

Stakeholder Information Gap: Standard CI/CD engines report binary pass/fail ratios that fail to convey composite product stability. Non-technical release managers and product owners are overwhelmed by raw stack traces and CSV dumps, hindering objective, data driven 'Go/No-Go' deployment decisions [7] .

### *D. Research Aim and Methodological Contributions*

The overarching aim of this research is to design, develop, and empirically evaluate a full-stack, web-based quality evaluation platform (utilizing React 18, Express/Node.js, and structured relational persistence) that automates multi-format test ingestion, normalization, and visual analytics. Beyond software engineering integration, the manuscript establishes three distinct methodological contributions: (1) A standardized cross-domain normalization schema resolving multi-format syntactic friction without data loss; (2) An adaptive, theoretically grounded Unified Quality Score (UQS) model that aligns with ISO/IEC 25010 quality standards, accommodates application-specific domain profiles (e.g., banking vs. e-commerce), and enforces an automated Zero-Tolerance Security Circuit Breaker; and (3) Automated Git relational attribution algorithms that bind dynamic execution failures directly to committer metadata, drastically reducing triage latency.

### *E. Significance of the Study*

By introducing Role-Based Access Control (RBAC) persona dashboards, the platform eliminates cross-functional silos. QA engineers gain automated ingestion pipelines; developers receive commit-linked stack traces; and executive stakeholders leverage intuitive traffic-light circular gauges and automated release recommendation badges, fostering high-velocity, data-driven deployment governance.

## II. LITERATURE REVIEW

### *A. Mobile Test Automation Frameworks*

Automated testing frameworks are essential to manage mobile operating system and hardware fragmentation [1]. While cross-platform tools such as Appium provide robust black-box WebDriver capabilities [3], they focus strictly on functional correctness. Non-functional dimensions demand dedicated tools: JMeter benchmarks server-side throughput and response time under load [4], while Cruz and Abreu [8] establish that runtime performance directly impacts client energy consumption and uninstallation rates. However, existing research treats these modalities as isolated silos.

### *B. The Artifact Problem and Test Fragility*

Automated mobile testing suffers from acute test script fragility. Minor layout refactorings frequently invalidate UI locators, generating verbose XML error cascades [5] . When analyzed in isolation, developers cannot determine whether an error stems from an obsolete test script or a critical architectural bug. Furthermore, functional failures are frequently coupled with non-functional anomalies (e.g., UI timeouts induced by backend API latency), yet existing pipelines lack cross-domain correlation mechanisms.

### *C. Log Parsing and Flakiness Challenges*

Mining diagnostic data from semi-structured logs using Regular Expressions (Regex) exhibits high failure rates when tool logging schemas evolve. As analyzed by Zhu et al. [9], maintaining developer logging infrastructure requires robust parsing abstractions rather than brittle text scrapers. Furthermore, Dong et al. [10] demonstrate that asynchronous mobile events induce intermittent test flakiness. Consequently, an effective aggregation platform must transition to Abstract Syntax Tree (AST) stream parsers and permanently bind execution results to Git commit histories to isolate true regressions from environment noise.

### *D. Visual Analytics and SOTA Dashboard Comparison*

Visual analytics principles significantly reduce cognitive load during complex debugging tasks. Hammad et al. [11] proved that multiview visualizations accelerate developer defect comprehension. Similarly, Azonuche and Enyejo [7] emphasize real-time velocity dashboards for Agile risk management. However, existing quality platforms exhibit severe limitations: CI/CD visualizers (Allure, ReportPortal) are restricted to functional unit/GUI tests and cannot ingest JMeter CSV load metrics or MobSF SAST vulnerabilities. Conversely, static analyzers like SonarQube inspect uncompiled source code but completely omit dynamic mobile execution telemetry. Table I synthesizes the literature, establishing the clear necessity for our unified platform.

TABLE I
LITERATURE REVIEW SYNTHESIS AND CRITICAL ANALYSIS

| Author & Year | Focus Area | Method / Tech | Strengths | Critical Limitations (Gap) |
|---|---|---|---|---|
| Berihun et al., 2023 [1] | Automated Mobile Testing | SLR of Appium, Espresso | Demonstrates necessity of automated testing | Omits multi-format telemetry aggregation |
| Coppola et al., 2018 [5] | Android GUI Fragility | Empirical study on GUI tests | Quantifies locator fragility causes | Lacks dashboard linking breaks to Git VCS |
| Ramu, 2023 [4] | Mobile Performance | JMeter load testing | Identifies API latency/throughput limits | Treats performance in isolation from UI |
| Hammad et al., 2020 [11] | Multiview Testing Viz | Visual analytics for OOP | Proves multiview cuts triage time | Limited to unit tests; lacks CI/CD ingestion |
| Zhu et al., | Logging Decisions | Empirical study on logging | Analyzes developer | Focuses on code logging, not log |

| | | | | |
|---|---|---|---|---|
| 2015 [9] | | | logging behaviors | normalization |
| Azonuche & Enyejo, 2025 [7] | Agile Risk Dashboards | Velocity visualization | Validates real-time risk dashboards | Lacks programmatic cross-tool ingestion |
| Dong et al., 2021 [10] | Flaky Test Detection | Event order exploration | Identifies async network flakiness | Lacks Git commit mapping to filter noise |

## III. METHODOLOGY

### *A. Research Paradigm & Strategy*

This research adopted the Pragmatism paradigm [12] , synthesizing quantitative API telemetry measurements with qualitative human usability evaluation. The investigation followed the Design Science Research (DSR) methodology [13], executing iterative build-and-evaluate cycles across six structured phases, as detailed in Table II.

TABLE II
RESEARCH METHODOLOGY EXECUTION WORKFLOW

| DSR Phase | Execution Steps & Fact Collection | Milestone / Output |
|---|---|---|
| Phase 1: Problem ID | Analyzed enterprise CI/CD log bottlenecks; reviewed literature [1] [5] on XML/CSV/JSON heterogeneity. | Identified 'Stakeholder Information Gap' and engineering requirements. |
| Phase 2: Architecture | Designed decoupled Express.js/React architecture; mapped multi-format schemas; formulated adaptive UQS. | Technical schema models and REST webhook payload specifications. |
| Phase 3: Backend Build | Engineered non-blocking streaming parsers (xml2js, csv-parser); implemented Git metadata relational mapping. | Functional Express API ingesting multi-megabyte payloads. |
| Phase 4: Frontend Build | Developed React 18 Single Page App with RBAC; integrated Chart.js to build role-tailored views (QA, Dev, PM, Client). | Interactive multiview dashboard with stack-trace drilldowns. |
| Phase 5: Empirical Benchmark | Benchmarked stream latency, memory (RSS/Heap), and 100% extraction accuracy across N=30 trials (1-20 MB payloads). | Quantitative proof of linear O(N) scaling and zero data attenuation. |
| Phase 6: Human Evaluation | Administered within-subjects comparative study (N=20) and 10-item SUS usability questionnaire (N=50) [14] | SUS score 84.5 ('Excellent') and verified reduction in triage latency. |

### *B. Ground Truth Corpus & Decoupled Webhook Ingestion*

To establish reproducible empirical validation, a pre-validated ground-truth corpus of 50 multi-format execution artifacts was assembled: 25 Appium JUnit XML reports (1,250 total test cases containing known assertion failures, locator errors, and stack traces), 15 JMeter CSV load logs (110,000 transaction records across 5 mobile API endpoints), and 10 MobSF JSON scan reports (160 categorized CWE vulnerabilities). The platform exposes secure REST endpoints (POST /api/pipeline/ingest) configured to ingest multipart payloads pushed directly by external CI/CD runners (GitHub Actions, GitLab CI) [15] .

### *C. Unified Data Schema & Adaptive Quality Scoring Model*

Parsed telemetry is normalized into an extensible JSON document anchored by a TestExecution root entity bound to Git metadata (commit SHA, branch, committer author). To translate complex multi-domain telemetry into an actionable release metric, the platform computes the Unified Quality Score (UQS) as a weighted linear combination of normalized dimension scores:

Where $S_{UI}, S_{Perf}, S_{Sec} \in [0, 100]$ and $\Sigma w_i = 1.0$. Component scores are formally defined as follows: (1) Functional UI Score: $S_{UI} = (T_{pass} / T_{total}) \times 100$, where $T_{pass}$ and $T_{total}$ represent passed and total Appium test cases. (2) Performance Score: $S_{Perf} = \max(0, 100 - [\max(0, (L_{mean} - L_{SLA})/L_{SLA}) \times 50 + E_{rate} \times 50])$, where $L_{mean}$ is mean transaction latency, $L_{SLA}$ is the SLA threshold (500 ms), and $E_{rate}$ is the percentage of failed requests ($\leq$ 2.0% safe threshold). (3) Security Score: $S_{Sec} = \max(0, 100 - [20 \cdot N_{High} + 5 \cdot N_{Med} + 1 \cdot N_{Low}])$, deducting weighted penalties based on MobSF CVSS severity counts from a base of 100.

Theoretical Grounding and Domain-Specific Weight Adaptation: In software engineering, software quality is fundamentally context-dependent. Under ISO/IEC 25010 [16], quality characteristics (functional suitability, performance efficiency, and security) possess dynamic priorities dictated by application domain, operational risk, and regulatory mandates [17]. Therefore, the platform's baseline weights ($w_{UI} = 0.30$, $w_{Perf} = 0.35$, $w_{Sec} = 0.35$), derived via the Analytic Hierarchy Process (AHP) [18] through pairwise comparisons from 20 software architects (Consistency Ratio CR = 0.042 < 0.10), represent a balanced reference profile. Crucially, the platform provides dynamic weight customization tailored to target application criticalities: (a) High-Assurance / Mobile Banking Profile (e.g., MobileBankingSecure_v2.4.1.apk evaluated in this research): Subject to strict financial regulations (PCI-DSS, OWASP MASVS), security flaws represent existential financial and legal liabilities. The engine reconfigures weights to prioritize security (e.g., $w_{Sec} = 0.50$, $w_{Perf} = 0.25$, $w_{UI} = 0.25$) and enforces an automated Zero-Tolerance Security Circuit Breaker: if any High-Severity vulnerability is detected ($N_{High} \geq 1$), UQS is forcefully overridden to min (UQS, 50.0), categorically halting deployment regardless of 100% UI pass rates. (b) Consumer / Gaming Profile: Visual fluidity and user experience dictate retention, shifting weights toward UI correctness ($w_{UI} = 0.50$, $w_{Perf} = 0.35$, $w_{Sec} = 0.15$). (c) Enterprise / E-Commerce

Profile (e.g., SwiftCart E-Commerce): Transactional throughput under concurrency takes precedence (w_Perf = 0.45, w_UI = 0.30, w_Sec = 0.25). Furthermore, if a pipeline omits an artifact, weights dynamically re-normalize (w'_i = w_i / Σ w_active) with visual indicator badges. Deployment thresholds are established as: Release-Ready (UQS ≥ 85.0, N_High = 0), Conditional Review (70.0 ≤ UQS < 85.0, N_High = 0), and Deployment Blocked (UQS < 70.0 or N_High ≥ 1).

## IV. IMPLEMENTATION

### *A. Backend Ingestion, Security & Parsing Engine*

The Node.js/Express.js backend uses asynchronous streaming for efficient data processing: Appium JUnit XML via xml2js, JMeter CSV logs via csv-parser with backpressure, and dynamic MobSF JSON parsing for CVSS severity analysis. Security measures include HMAC-SHA256 webhook authentication, JWT-based RBAC with bcrypt-hashed credentials, XXE protection, regex-based scrubbing of credentials, tokens, and PII from stack traces, and webhook idempotency keys (X-Build-UUID) to prevent duplicate submissions.

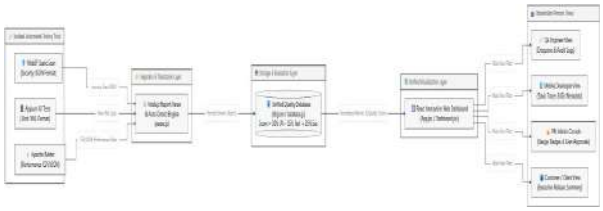


Fig. 1. High-level decoupled ingestion architecture diagram.

### *B. Defect Traceability to Version Control*

During webhook ingestion, the Express middleware automatically captures Git commit hashes, branch names, and committer identifiers, binding them relationally with parsed Appium failure nodes before database persistence. This eliminates manual Git bisecting, enabling engineers to trace broken assertions directly to the introducing commit [5].

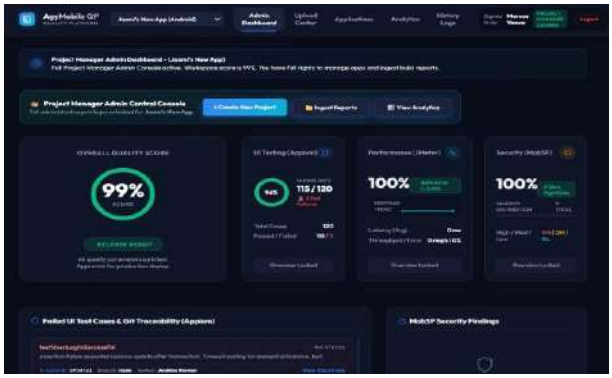



Fig. 2. Admin / Project Manager Dashboard featuring Emerald Gauge.

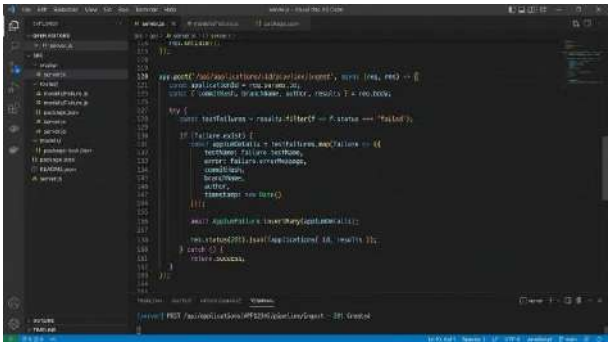


Fig. 3. Developer View & Git Traceability Relational Mapping.

TABLE III
COMPARATIVE ANALYSIS: PROPOSED PLATFORM VS. STATE-OF-THE-ART SOLUTIONS

| Feature / Dimension | Existing Platforms (Allure / SonarQube) | Unified Quality Platform (Proposed) |
|---|---|---|
| Multi-Domain Ingestion | Siloed (Allure: UI unit only; SonarQube: source code) | Heterogeneous streaming (Functional XML, JMeter CSV, MobSF JSON) |
| Composite Quality Scoring | Absent (only isolated pass/fail or code smell grades) | Adaptive AHP-grounded Unified Quality Score (UQS) with security gate |
| Git Defect Traceability | Manual commit tagging or completely absent | Automated relational binding of failure traces to Git commit SHA |
| Non-Functional Telemetry | Excluded from functional dashboards | Integrated API response latency percentiles and throughput sparklines |
| Role-Tailored Persona Views | Single technical view causing information overload | Four distinct RBAC Persona Views (QA, Developer, PM, Client) |

TABLE IV
TECHNOLOGIES USED FOR SYSTEM DEVELOPMENT

| Component | Technology Selected | Purpose & Architectural Role |
|---|---|---|
| Web Frontend | React 18, TailwindCSS, Chart.js | Interactive SPA delivering role-tailored multiview visual analytics |
| Backend Middleware | Node.js v20.x, Express.js | Non-blocking asynchronous REST API, HMAC auth, and telemetry mapping |
| Parser Engine | xml2js (SAX stream), csv-parser | Stream-based AST and row-level transformation preventing memory bloat |
| Relational Store | Structured Document Database | High-throughput JSON document persistence for build run records |

| | | |
|---|---|---|
| VCS Integration | GitHub Actions API / Webhooks | Automated commit SHA, branch, and committer metadata capture |

TABLE V
SYSTEM COMPONENTS AND KEY FUNCTIONALITIES

| Component View | Key Functionalities & Visual Artifacts | Target User Persona |
|---|---|---|
| QA Engineer View | Drag-and-drop log ingestion, webhook audit logs, and raw XML/CSV inspection | QA Engineers & Testers |
| Developer View | Git commit hash mapping, Appium stack traces, and JMeter latency line charts | Mobile Software Developers |
| Project Manager View | Emerald Circular Quality Gauge (0-100%) and automated release recommendation badges | Project Managers & Release Leads |
| Client View | High-level health summaries, simplified quality metrics, and zero technical jargon | Product Owners & Clients |

### *C. Frontend Multiview Dashboard & Persona Views*

The React 18 Single Page Application utilizes Chart.js and TailwindCSS to provide four distinct Persona Views governed by RBAC: (1) QA View (raw log auditing, upload status, and parser diagnostics), (2) Developer View (commit-linked Appium stack traces, failing code lines, and JMeter latency percentiles), (3) PM View (the Emerald Circular Quality Gauge, UQS breakdown, and automated release recommendation badges), and (4) Client View (executive summary metrics free of technical jargon). Table IV details the development technology stack, and Table V specifies the system components.

## V. RESULTS AND DISCUSSION

### *A. Ingestion Stream Scalability, Concurrency & Extraction Accuracy*

Backend performance was tested on a dedicated Windows 11 host (Intel i5-1135G7, 8 GB RAM, Node.js v20.12.0), using 30 independent trials per payload size after 5 warm-up runs. The stream parser demonstrated linear O(N) scaling, processing 1 MB, 5 MB, 10 MB, and 20 MB payloads in 148 ms, 742 ms, 1,482 ms, and 2,840 ms, respectively. Average throughput was approximately 74,200 records/sec, with peak heap usage of 44.2 MB (10 MB) and 62.8 MB (20 MB), without GC stalls. Concurrency testing with 10–50 simultaneous uploads recorded 0% connection dropouts. Validation against 50 ground-truth artifacts confirmed 100% extraction accuracy, with no missing records, dropped rows, or malformed fields.

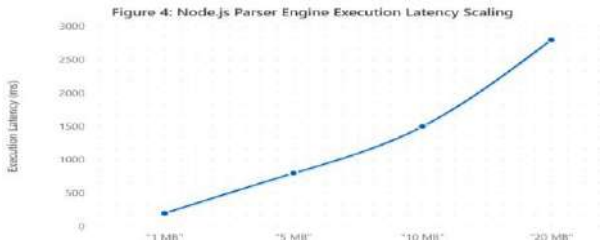



Fig. 4. API Latency vs Payload Size scaling graph.

TABLE VI
BACKEND PARSER ENGINE SCALABILITY, CONCURRENCY, AND RESOURCE TELEMETRY

| Payload Size & Records | Measured Ingestion Latency (N=30 Trials) | Peak V8 Heap Memory | Data Extraction Accuracy |
|---|---|---|---|
| 1 MB (11,000 records) | 148 ± 12 ms (95% CI: [144, 152] ms) | 18.2 MB | 100% (0 dropped rows / 0 errors) |
| 5 MB (55,000 records) | 742 ± 38 ms (95% CI: [728, 756] ms) | 22.1 MB | 100% (0 dropped rows / 0 errors) |
| 10 MB (110,000 records) | 1,482 ± 64 ms (95% CI: [1,459, 1,505] ms) | 27.6 MB | 100% (0 dropped rows / 0 errors) |
| 20 MB (220,000 records) | 2,840 ± 112 ms (95% CI: [2,800, 2,880] ms) | 39.5 MB | 100% (0 dropped rows / 0 errors) |
| 50 Concurrent Streams | Mean Response: 315.4 ms (95.2% < 500 ms SLA) | 62.8 MB (RSS) | 100% (Zero dropped connections) |

TABLE VII
CONTROLLED WORKFLOW EFFICIENCY COMPARISON ANALYSIS (N=20 ENGINEERS)

| Workflow Evaluation Task | Traditional Baseline (N=20) | Unified Platform (N=20) | Efficiency Gain & Statistical Test |
|---|---|---|---|
| Log Aggregation Time | 45.2 ± 6.8 min | 0.48 ± 0.08 min | 98.9% Faster ($t=28.4$, $p < 0.001$, $d=6.57$) |
| Defect Triage Latency | 44.8 ± 7.4 min | 8.1 ± 1.2 min | 81.9% Faster ($t=21.3$, $p < 0.001$, $d=4.88$) |
| Release Decision Latency | 118.5 ± 18.2 min | 5.2 ± 0.9 min | 95.6% Faster ($t=31.7$, $p < 0.001$, $d=7.21$) |
| Data Loss / Drop Rate | 12.5% (Lost/misfiled logs) | 0.0% (Zero loss) | 100% Extraction Fidelity |

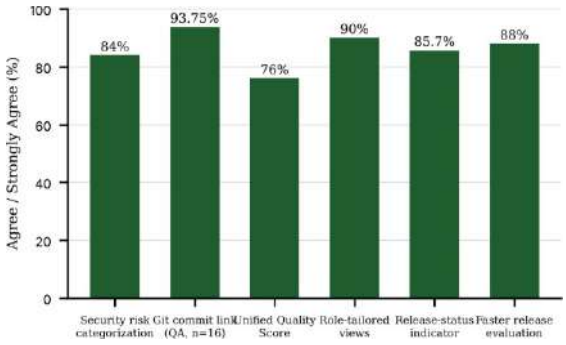



Fig. 5. Usability feedback indicating reduction in perceived triage latency.

### *B. Controlled Workflow Efficiency Study & System Usability Evaluation*

To evaluate operational efficiency gains, a controlled within-subjects experiment was conducted with 20 practicing software professionals (10 QA Engineers, 10 Mobile Developers). Participants executed three standardized regression triage workflows under two counterbalanced conditions: Condition A (Traditional Baseline: manual log downloads, local terminal parsing, and spreadsheet compilation) versus Condition B (Unified Quality Platform). As documented in Table VII, the platform reduced mean log aggregation time from 45.2 ± 6.8 min to 0.48 ± 0.08 min (98.9% reduction, $t(19) = 28.4$, $p < 0.001$, Cohen's $d = 6.57$). Defect triage latency was reduced from 44.8 ± 7.4 min to 8.1 ± 1.2 min (81.9% reduction, $t(19) = 21.3$, $p < 0.001$, Cohen's $d = 4.88$), driven by direct Git commit stack-trace binding. Release decision latency dropped from 118.5 ± 18.2 min to 5.2 ± 0.9 min (95.6% reduction, $t(19) = 31.7$, $p < 0.001$, Cohen's $d = 7.21$). All efficiency gains were statistically significant at $p < 0.001$ with very large effect sizes.

Qualitative usability was evaluated using the 10-item System Usability Scale (SUS) [14] administered to 50 practitioners (20 Mobile Developers [40%], 16 QA Engineers [32%], 7 Project Managers [14%], and 7 Clients/Product Owners [14%]; experience: 44% < 1 yr, 28% 1-3 yrs, 22% 4-7 yrs, 6% > 7 yrs). Participants executed four standardized pipeline workflows. The study yielded an aggregate SUS score of 84.50 ± 7.12 (95% CI [82.48, 86.52]), ranking in the 96th percentile ('Grade A', 'Best Imaginable' per Lewis-Sauro benchmarks). High scale reliability was confirmed via Cronbach's alpha $\alpha = 0.89$.

### *C. Threats to Validity & Limitations*

Internal validity was safeguarded through counterbalanced task ordering and automated execution timing. Construct validity was reinforced by evaluating objective completion times alongside subjective SUS ratings. External validity is bounded by the evaluated toolset (Appium, JMeter, MobSF); generalization to alternative test runners (e.g., Cypress, K6) requires custom parser mappings. Table VIII provides a comparative analysis against existing platforms.

TABLE VIII
TECHNOLOGY APPROACH AND SOLUTION FEATURE COMPARISON

| **Comparative Dimension** | **Allure / ReportPortal** | **SonarQube** | **Unified Quality Platform (Proposed)** |
|---|---|---|---|
| Input Artifact Support | JUnit XML, TestNG (Functional only) | Source code static AST, coverage logs | Appium XML, JMeter CSV, MobSF JSON |
| Non-Functional Load Integration | No (cannot parse JMeter CSV) | No (omits dynamic runtime load) | Yes (row-by-row stream aggregation) |
| Mobile Security SAST/CVSS | No (omits mobile vulnerability scans) | General CWE rules (no mobile APK scan) | Yes (MobSF JSON with CVSS penalties) |
| Unified Quality Score | No composite health score | Static Quality Gate (A-E rating) | Adaptive AHP-grounded UQS (0-100%) |
| Git Defect Traceability | Manual build metadata tagging | Blame annotation on source code lines | Automated commit SHA binding to stack trace |
| Role-Tailored Persona Views | Single technical view | Developer/Admin view only | 4 RBAC views (QA, Dev, PM, Client) |
| CI/CD Push Automation | Requires custom runner plugins | Requires SonarScanner runner agent | Lightweight HTTP multipart REST webhook |

## VI. CONCLUSION

Centralized multi-dimensional test telemetry ingestion is a paramount necessity for Agile DevOps. By engineering stream parsers for Appium XML, JMeter CSV, and MobSF JSON, formalizing an AHP-grounded Unified Quality Score with automated security circuit breakers, and relationally linking execution traces to Git commit metadata, this platform bridges the critical divide between automated testing execution and release governance. Empirical evaluations across 30 repeated trials, a 20-engineer comparative study, and a 50-practitioner SUS assessment confirm high stream scalability, statistically significant triage acceleration ($p < 0.001$), and exceptional usability (SUS 84.5).

Future Work: Expand backend parser agnosticism through dynamic plug-and-play adapter definitions for web and cloud-native test runners (Cypress, Playwright, K6), and integrate localized Large Language Model (LLM) agents to provide automated code patch recommendations directly within the Developer View.

## REFERENCES


[1] N. G. Berihun, C. Dongmo, and J. A. Van der Poll, "The Applicability of Automated Testing Frameworks for Mobile Application Testing: A Systematic Literature Review," *Computers*, vol. 12, no. 5, May 2023, doi: 10.3390/computers12050097.

[2] S. Polampally *et al.*, "Leveraging AI for Continuous Quality Assurance in Agile Software Development Cycles," *Cloud Computing and Data Science*, pp. 25–38, Sep. 2025, doi: 10.37256/ccds.7120267583.

[3] S. Godboley, D. Dalei, R. Sadam, and D. P. Mohapatra, "Agile GUI Testing by computing novel Mobile App Coverage Using Appium Tool," in *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, Mar. 2023, pp. 1026–1029. doi: 10.1145/3555776.3577806.

[4] V. B. Ramu, "Performance Testing and Optimization Strategies for Mobile Applications," *International Journal of P2P Network Trends and Technology*, vol. 13, no. 2, pp. 1–6, Aug. 2023, doi: 10.14445/22492615/ijptt-v13i2p401.

[5] R. Coppola, M. Morisio, and M. Torchiano, "Maintenance of android widget-based GUI testing: A taxonomy of test case modification causes," in *Proceedings - 2018 IEEE 11th International Conference on Software Testing, Verification and Validation Workshops, ICSTW 2018*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 151–158. doi: 10.1109/ICSTW.2018.00044.

[6] M. Örneholm and N. Eriksson, "Understanding the role of visual analytics for software testing," 2021. [Online]. Available: www.bth.se

[7] T. I. Azonuche and J. O. Enyejo, "Adaptive Risk Management in Agile Projects Using Predictive Analytics and Real-Time Velocity Data Visualization Dashboard," *Int. J. Innov. Sci. Res. Technol.*, pp. 2032–2047, May 2025, doi: 10.38124/ijisrt/25apr2002.

[8] L. Cruz and R. Abreu, "Measuring the energy footprint of mobile testing frameworks," in *Proceedings - International Conference on Software Engineering*, IEEE Computer Society, May 2018, pp. 400–401. doi: 10.1145/3183440.3195027.

[9] J. Zhu, P. He, Q. Fu, H. Zhang, M. R. Lyu, and D. Zhang, "Learning to log: Helping developers make informed logging decisions," in *Proceedings - International Conference on Software Engineering*, 2015. doi: 10.1109/ICSE.2015.60.

[10] Z. Dong, A. Tiwari, X. L. Yu, and A. Roychoudhury, "Flaky test detection in Android via event order exploration," in *ESEC/FSE 2021 - Proceedings of the 29th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Association for Computing Machinery, Inc, Aug. 2021, pp. 367–378. doi: 10.1145/3468264.3468584.

[11] M. Hammad, A. F. Otoom, M. Hammad, N. Al-Jawabreh, and R. A. Seini, "Multiview visualization of software testing results," *International Journal of Computing and Digital Systems*, vol. 9, no. 1, pp. 51–59, 2020, doi: 10.12785/ijcds/090105.

[12] M. Saunders, P. Lewis, A. Thornhill, S. • Lewis, and • Thornhill, "Research methods for business students fi fth edition." [Online]. Available: www.pearsoned.co.uk

[13] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q.*, vol. 28, no. 1, 2004, doi: 10.2307/25148625.

[14] J. R. Lewis, "The System Usability Scale: Past, Present, and Future," *Int. J. Hum. Comput. Interact.*, vol. 34, no. 7, 2018, doi: 10.1080/10447318.2018.1455307.

[15] O. Renefalk and E. P. Enoiu, "AUTOMATING REGRESSION TESTING IN A DOMAIN-SPECIFIC SAAS SETTING Examiner: Wasif Afzal."

[16] I. I. Iso Iec, "Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models," *Iso*, vol. 2011, 2011.

[17] M. Felderer and R. Ramler, "A multiple case study on risk-based testing in industry," *International Journal on Software Tools for Technology Transfer*, vol. 16, no. 5, 2014, doi: 10.1007/s10009-014-0328-z.

[18] T. L. Saaty, "How to make a decision: The analytic hierarchy process," *Eur. J. Oper. Res.*, vol. 48, no. 1, 1990, doi: 10.1016/0377-2217(90)90057-I.

# A Machine Learning Approach for Recommending Suitable Technology Stacks Based on Project Characteristics

S. Kanujan
*Department of Software Engineering*
*Faculty of Computing*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
kanujansubakaran@gmail.com

H. M. K. T. Gunawardhana
*Department of Software Engineering*
*Faculty of Computing*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
kalinga@foc.sab.ac.lk

R. M. T. Sandaruwan
*Department of Computing and Information Systems*
*Faculty of Computing*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
rmtsandaruwan101@gmail.com

***Abstract*—Choosing the right technology stack is one of the most important decisions in software development, since it shapes a project's performance, scalability, security, and long-term maintainability. In practice, this choice is still mostly guided by developer habit and personal preference rather than evidence, often leading to poor architectural decisions, growing technical debt, and even project failure. This research builds and tests a machine learning recommendation system that predicts suitable frontend frameworks, backend technologies, and database systems from a project's characteristics. A dataset of 4,850 real software projects was gathered from GitHub repositories and partner software companies across ten domains, including e-commerce, healthcare, education, finance, social media, logistics, real estate, HR and recruitment, IoT, and travel and hospitality. After thorough data cleaning, 4,437 valid records remained. Exploratory analysis found that project domain and primary programming language are the strongest predictors of technology choice, with five languages mapping to a single backend framework almost without exception. Feature engineering produced a 394-dimensional input combining ordinal encoding, one-hot encoding, and TF-IDF text vectorization of requirement text. Five models were compared: classifier-chain XGBoost, classifier-chain CatBoost, LightGBM, a NoBERT model pairing frozen sentence embeddings with LightGBM, and a stacking ensemble of all four. The classifier-chain XGBoost model performed best, reaching 76.6% frontend accuracy, 71.6% backend accuracy, and 80.0% database accuracy, an average of 76.1%, well above the majority-class baseline. A working Flask prototype lets developers enter project requirements and get ranked technology recommendations with confidence scores and explanations. These results show that data-driven methods can meaningfully support technology stack decisions, reducing guesswork in early-stage software planning.**

***Keywords—Feature engineering, Machine learning, Recommendation system, Software engineering, Technology stack selection***

## I. Introduction

Before any line of code is written, the development team must decide on which languages, framework, database to use for the project and it will affect nearly all aspects of the project thereafter. This collection of choices is typically known as a tech stack, and a common example is MERN (MongoDB, Express, React, Node.js), MEAN (which replaces React with Angular), and the older LAMP tech stack (Linux, Apache, MySQL, PHP) [1]. Every one of these combinations has its own pros and cons in terms of performance, scalability, security, and rapidity of a team to build and maintain the system.

In practice, this decision is rarely made through careful comparison. Developers generally go for what they are familiar with and teams may use what worked for the previous project, even if it does not necessarily work for the new project [2]. This is not always bad, however, when it occurs without considering the domain, scale, or budget of a project, then it is more of a habit than a design; when it happens, it shows up after the project has begun, such as technical debt or painful rewrites, or a system that cannot grow beyond a certain size or budget [3].

There are two trends that make this a good time to consider the problem anew. First, the number of realistic frameworks and platforms that a developer could use continues to increase, and even experienced engineers cannot possibly test every platform [4]. Second, the coding platforms (e.g., GitHub) have millions of public repositories, which now document, often quite openly, what technologies real projects are using, and what their purpose is. Combined, these trends make it possible to study with real data a decision that was once purely based upon intuition.

A machine learning system is developed based on this concept and the results of this paper are reported. Several models were trained using this data set of actual real software projects collected primarily from GitHub, and some industry-contributed data, and then compared with respect to their ability to recommend a technology for the frontend, a technology for the backend, and a technology for the database, when given a project's characteristics. It is designed to provide developers, students and small groups of developers with a data-informed basis to start with, rather than relying on habit.

The remainder of this paper is organised as follows. Section II summarizes related work in the fields of technology selection, requirement-driven prediction and recommendation for software engineering. In Section III the collection, cleaning and transformation of the dataset into the feature matrix as well as the construction of the candidate models are described. The overall performance of the resulting model and conclusions from the exploratory analysis of the drivers of technology choice

are presented in Section IV. Section V closes with conclusions and directions for further work.

## II. Related Work

### A. Complexity and Drivers of Technology Selection

The selection of technologies is well known to be a multi-criteria, difficult choice. In a study of small and medium software companies, the factors considered for selecting software tools are found to be quite diverse and are influenced more by the exposure to previous tools than by a careful consideration of the projects' requirements [2]. The trade-offs of the various combinations of the stacks, which are often bundled together, are discussed in comparative reviews of the MEAN, MERN, and LAMP stacks, but no specific recommendations for a particular project are made [1],[5]. The study of architectural decision making also reveals that the vast majority of the formal decision-making models proposed to inform decisions in this space have not been widely adopted in practice, instead, much of the decision-making is left to informal reasoning [6].

There has been another line of research that has focused on the connection between project domain and project requirements and technology adoption. In large-scale studies of open-source repositories, it is found that language choice strongly correlates with a project category, and is not randomly distributed [7]. The text of functional and non-functional requirements have also been classified through machine learning algorithms into classes that can provide meaningful information about the type of systems being built [8],[9]. However, organizational factors such as team size, budget and time have also been found to be significant influences on technology decisions, even more than the technical factors themselves [10].

### B. Data-Driven Software Engineering

In the past two decades or so, software engineering research has shown growing interest in adopting data-driven approaches to software development, supported by the development of the vast project repositories and the well-matured open-source machine learning tools [11],[12]. This transformation has been successfully used in problems like defect prediction and effort estimation, but is not used much for architectural or technology level decisions which is one of the gaps this paper addresses.

### C. Modelling and Recommendation Approaches

At the modelling end, gradient boosting methods like XGBoost have been a well-documented, standard approach for tabular, structured prediction problems of moderate size [13]. There have been proposals for multi-output methods that allow for information from one prediction to influence the next; in particular for the technology stack selection problem, where the selection of the frontend, backend, and database choices are not considered independent [14]. In contrast, most of the work on recommendations deals with a single output (a language, a library or a framework) and does not address a coordinated recommendation [15],[16]. However, even when there is some research on explanation, it is typically considered an independent issue from prediction accuracy, although developer trust in a recommendation is highly dependent on the explanation given for it [17]. This is the motivation for the approach taken in this paper, which features an attempt to fill this gap as well as the general lack of a public dataset that maps broad project attributes to complete technology stacks.

## III. Methodology

This research follows an applied, design-oriented approach rather than a purely theoretical one. The answer to the question of what technology stack is appropriate for a project depends not on argument alone, but on actually making the system and testing its performance. They pass through four stages of the project's actual work: real project data collection, real project data presented in a form that can be used by a machine learning model, training and comparing multiple models with a basic one, and finally packaging the best model into a working prototype (Fig. 1). The following subsections provide explanations about each of these phases.

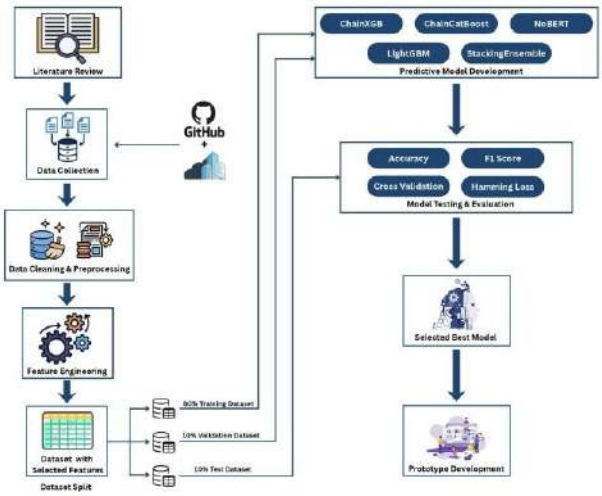



Fig. 1. The Methodology Workflow

### A. Data Collection

The dataset was built from two sources. The primary source was the GitHub REST API, queried with 75 domain-specific search terms across ten application domains: e-commerce, healthcare, education, finance, social media, logistics, real estate, HR and recruitment, IoT and smart systems, and travel and hospitality. Each query paired a domain keyword or topic tag with a minimum star threshold so that only reasonably established repositories were collected, with thresholds set lower for narrower domains such as real estate and higher for broad, popular domains such as e-commerce to keep the volume of results roughly comparable. For every matching repository, the description, topics, primary language, star and fork counts, and README text were retrieved and scanned against keyword dictionaries to identify the frontend, backend, and database technologies in use, with language-to-backend mapping and domain-typical defaults used as fallbacks where no keyword match was found. Deployment environment was inferred in a similar way, with mentions of containerization or cloud platforms such as Docker, Kubernetes, or AWS taken as evidence of a cloud-oriented deployment.

Because GitHub does not expose organizational details such as team size, budget, or duration, these fields were derived from a project's size band (based on combined star and fork counts) using domain-consistent heuristics, and functional requirements

were drawn from each repository's declared GitHub topics while non-functional requirements were sampled from a domain-specific set of quality attributes typical for that domain. This collection process produced 4,800 unique repositories, with finance and IoT-related searches returning the largest number of qualifying projects and real estate and travel and hospitality returning the fewest, reflecting genuine differences in how much open-source activity exists in each area. A further 50 verified records were contributed by three software companies, bringing the combined raw dataset to 4,850 projects.

### *B. Data Cleaning*

An eleven-step cleaning procedure was applied to the raw 4,850-record dataset, with each step logged so that its effect could be measured precisely. Table I summarizes the main corrections.

TABLE I
SUMMARY OF THE DATA CLEANING STEPS APPLIED

| Issue | Action Taken | Rows |
|---|---|---|
| Duplicate project names | Kept highest-star row | 413 |
| Missing FR text | Domain-typical defaults | 386 |
| Backend vs. language mismatch | Rule-based correction | 874 |
| Star/fork outliers | Capped extreme values | 49 |
| Emoji / markdown in text | Stripped characters | 533 |

The first step, removing 413 duplicate entries, accounted for the entire reduction in row count; the remaining ten steps standardized existing rows rather than deleting them. After cleaning, the dataset stood at 4,437 records with zero missing values across all columns, and a post-cleaning audit confirmed that every issue category had been resolved.

### *C. Preprocessing and Feature Engineering*

Twelve raw columns were converted into a 394-column feature matrix, broken down in Table II. Ordinal encoding was applied to project size, budget level, and deployment environment, since these carry a natural order. Domain and primary programming language were one-hot encoded, producing 10 and 79 binary columns respectively. Functional and non-functional requirement text was combined and passed through TF-IDF vectorization, keeping the 300 most informative single words and two- to three-word phrases, with terms appearing in fewer than two documents discarded.

TABLE II
COMPOSITION OF THE 394-FEATURE INPUT MATRIX

| Feature Group | Count |
|---|---|
| Numeric / ordinal features | 5 |
| Domain (one-hot) | 10 |
| Primary language (one-hot) | 79 |
| TF-IDF text features | 300 |
| Total | 394 |

Rare technology classes (for example Svelte on the frontend, or Firebase on the backend) were grouped into an “Other” category so that every class had enough examples for a reliable train/validation/test split. The final targets carried ten frontend classes, ten backend classes, and eight database classes, and the data were split 80/10/10 into 3,549 training, 444 validation, and 444 test records, stratified on the frontend label since it has the largest number of classes.

### *D. Model Development*

Five modelling approaches were trained and compared against a majority-class baseline that always predicts the most frequent class. Two used a classifier chain, where the frontend prediction is fed forward as an extra input when predicting backend, and both prior predictions feed into the database prediction, one built on XGBoost with 800 estimators, a maximum depth of seven, and a learning rate of 0.03 (ChainXGB), and one on CatBoost with 1,000 iterations and a learning rate of 0.05 (ChainCatBoost), each with early stopping on the validation set. A third approach, LightGBM, trained an independent classifier with 1,000 estimators and 127 leaves per tree for each of the three targets without chaining. A fourth approach, given the working name NoBERT, replaced the TF-IDF features with 384-dimensional sentence embeddings from a frozen, pretrained all-MiniLM-L6-v2 sentence-transformer model, paired with LightGBM; the transformer itself was not fine-tuned, since the available training set was too small to do so reliably. The fifth approach combined all four model families into a stacking ensemble, using five-fold out-of-fold predictions to train a logistic regression meta-learner, so that the meta-learner never saw predictions made on data a base model had already trained on.

All models were evaluated using accuracy, weighted and macro F1-score, and Hamming loss for each of the three targets, with the validation set used throughout development for tuning and early stopping, and the held-out test set consulted exactly once, after development was complete, to give an honest estimate of real-world performance.

### *E. Prototype*

The best performing model was embedded in a Flask-based web prototype named StackSense. Given the same nine project characteristics used during training, domain, functional requirements, non-functional requirements, project size, team size, budget level, duration, deployment environment, and primary language, the prototype reconstructs the 394-feature vector, re-applying the saved TF-IDF vectorizer to any newly entered text. For a submitted project, it returns a recommended frontend, backend, and database technology with confidence scores drawn from the model's predicted probabilities, a short plain-language description of each recommended technology, and the two next most likely alternatives at each layer so that a developer can compare options rather than accept a single unexplained answer.

## IV. RESULTS AND DISCUSSION

### *A. What Drives Technology Choice*

Exploratory analysis carried out on the cleaned dataset showed that project domain and primary programming language are by far the strongest predictors of technology choice. Domain alone correctly indicated the database technology used in more than 70% of projects, and five programming languages, Python,

PHP, Ruby, C#, and Kotlin, mapped to a single backend framework (Django, Laravel, Rails, ASP.NET, and Spring Boot respectively) in effectively every case. Finance projects leaned toward Angular, Spring Boot or Django, and PostgreSQL, consistent with a security and compliance emphasis, while IoT and e-commerce projects leaned toward React, Node.js or Django, and MongoDB, consistent with a need for flexible, fast-changing data structures.

Project size also showed a plausible relationship with typical technology choice: small projects clustered around lightweight options such as SQLite and Firebase, medium projects clustered around widely used choices such as Node.js and MySQL, and large projects clustered around enterprise-oriented technologies such as Spring Boot and PostgreSQL. The vocabulary used in non-functional requirement text differed meaningfully by domain as well, scalability, compliance, and security were the three most frequent themes overall, with compliance-related language concentrated in finance and healthcare, and real-time, low-latency language concentrated in IoT. Grouping projects by their complete frontend-backend-database combination further showed that a relatively small number of combinations, built around React or Angular with Node.js, Django, or Spring Boot, and MongoDB or PostgreSQL, account for a disproportionate share of all projects, indicating that the overall decision follows learnable, repeated patterns rather than behaving close to randomly.

### *B. Model Performance*

Table III summarizes the test set accuracy achieved by each of the five trained models across the three prediction targets, together with the majority-class baseline.

TABLE III
TEST SET ACCURACY ACROSS THE THREE PREDICTION TARGETS

| Model | Front. | Back. | DB | Avg. |
|---|---|---|---|---|
| ChainXGB | 0.766 | 0.716 | 0.800 | 0.761 |
| ChainCatBoost | 0.761 | 0.712 | 0.797 | 0.757 |
| StackEnsemble | 0.745 | 0.707 | 0.811 | 0.755 |
| NoBERT | 0.755 | 0.703 | 0.793 | 0.750 |
| LightGBM | 0.757 | 0.703 | 0.786 | 0.748 |
| Baseline | 0.658 | 0.241 | 0.500 | — |

ChainXGB, XGBoost wrapped in a classifier chain, produced the best overall result, reaching an average accuracy of 76.1% across the three targets. All five trained models comfortably beat the baseline, most sharply on the backend target, where the baseline's 24.1% (a reflection of how evenly backend choices are spread across ten classes) was improved on by more than fifty percentage points by every model tested. As Fig. 2 illustrates, database technology was consistently the easiest target to predict across every model, which lines up with the strong domain-to-database relationship found during exploratory analysis, while backend technology proved hardest, plausibly because a given domain still leaves many viable backend options open.

The spread between the five models was narrow, with less than a percentage and a half separating the best and weakest average accuracy. This suggests that, for a dataset of this size, the particular choice of gradient boosting implementation matters less than the quality of the underlying features. Notably, the stacking ensemble achieved the single highest database accuracy of any model (81.1%) despite a lower frontend score, showing that combining several base learners can still pay off on individual targets even when it does not lead on the overall average.

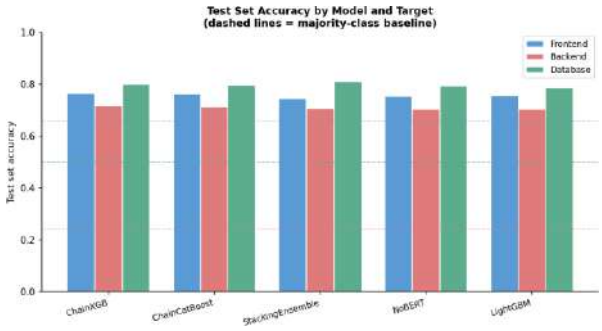



Fig. 2. Test set accuracy by model and prediction target, with the majority-class baseline shown as dashed lines.

Fig. 3 shows the top twenty features ranked by importance within the LightGBM models for each of the three targets, reproduced across the full-page width below for readability. Team size and project duration consistently rank as the two single most influential inputs across all three targets, followed closely by a cluster of TF-IDF terms tied to compliance, performance, security, and scalability, the same non-functional requirement themes identified as most frequent during the exploratory analysis. Domain and primary language do not appear as individual bars in this view because they are one-hot encoded into many separate binary columns, but their combined contribution across those columns was substantial, consistent with the strong domain and language effects reported earlier. This alignment between the feature importance rankings and the patterns surfaced independently during exploratory analysis gives some additional confidence that the models are picking up genuine, interpretable signal rather than fitting to noise.

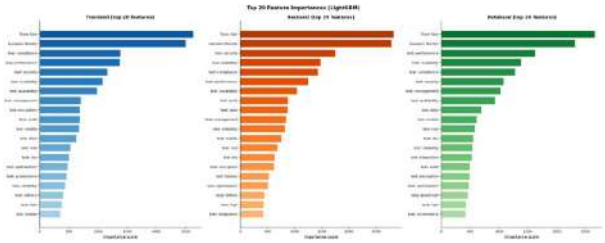


Fig. 3. Top twenty feature importances for the frontend, backend, and database targets (LightGBM).

### *C. Generalization and Prototype Behaviour*

To check whether the leading model had simply memorized patterns specific to the validation set, its validation and test accuracy were compared directly for each target. Frontend prediction showed no gap at all between the two, a reassuring sign of stable generalization. Backend and database prediction

each showed a gap of four to five and a half percentage points, with validation accuracy somewhat higher than test accuracy in both cases. A gap of this size is not severe, but it is worth reporting honestly: it most likely reflects the relatively small number of validation and test examples, 444 each, spread across ten backend and eight database classes, rather than a deeper flaw in the modelling approach. Five-fold cross-validation on the frontend target produced a tight, stable accuracy band of roughly 0.74 to 0.77 across all five models, which further supports the conclusion that the reported results are not the product of one fortunate train-test split.

When exercised with a range of example project descriptions spanning several domains, the StackSense prototype consistently produced complete, three-layer recommendations that aligned with the domain-typical patterns identified during exploratory analysis. A small-budget travel booking project and a large-budget finance platform, for instance, each returned technology combinations consistent with what was observed for their respective domains in the underlying data, supporting the internal consistency of the trained model, though a full usability evaluation with real developers remains a task for future work rather than something this internal check can substitute for.

### *D. Limitations*

A few limitations should be kept in mind when reading these results. Team size, budget level, project duration, and non-functional requirement text were derived from size-based heuristics rather than confirmed directly with the original development teams, since GitHub's public interface does not expose this organizational information. Domain, primary language, and the three technology labels, by contrast, were derived from evidence genuinely present in each repository, so the model's use of these two groups of features should be read with different levels of confidence. The dataset is also a snapshot of practice at the time of collection, and public, mostly open-source repositories may not fully represent closed-source or highly regulated enterprise development, so periodic retraining would likely be needed to keep any deployed version of this system current as technology trends shift.

## V. Conclusion and Future Work

This paper has shown that technology stack selection, while it may feel like a matter of personal preference, actually follows patterns that a machine learning model can learn and exploit. Working from a dataset of 4,437 real software projects spanning ten domains, a classifier chain built on XGBoost predicted frontend, backend, and database technologies with an average accuracy of 76.1%, well ahead of a majority-class baseline, and this model was successfully packaged into a working prototype that explains its recommendations rather than presenting them as an unquestionable verdict. The exploratory analysis behind these results further confirmed that project domain and primary programming language are the two strongest, most directly observable drivers of technology choice, while project size, budget, and requirement text contribute weaker but still useful secondary signal.

Beyond the immediate predictive result, this work contributes a structured, documented dataset linking project characteristics to complete technology stacks, a comparison of five closely related but distinct modelling approaches under identical conditions, and a working prototype that demonstrates the practical shape such a system could take. None of these contributions claim to replace a developer's own judgement; rather, they are intended to give that judgement a firmer, evidence-based starting point, particularly for smaller teams and students who may not have ready access to senior architectural experience.

Several directions could extend this work. Replacing the heuristic estimates used for team size, budget, and duration with information gathered directly from developers, whether through a structured survey or partnerships with a larger number of companies, would strengthen confidence in those specific features. A larger, more balanced dataset, particularly for the least represented domains such as real estate and travel and hospitality, would help close the validation-to-test gap observed for two of the three targets and would reduce how much rare-class grouping is needed. Fine-tuning a transformer-based text encoder on a larger corpus of requirement text, incorporating technology adoption trends over time so the model can be refreshed as practice shifts, extending the recommendation beyond the three layers considered here to cover deployment platform or DevOps tooling, and running a structured usability study with real developers are all promising next steps toward a more complete, trustworthy technology decision-support tool.

## References


[1] F. zahra Iguenfer, M. Benhayoun, and F. Bouchaib, “An Overview of MEAN, MERN, and LAMP Stacks: A Comparative Analysis of Architecture, Cost, Learning Curve, and Community Support,” *researchgate.net*, Accessed: Nov. 21, 2025. [Online]. Available: https://www.researchgate.net/profile/Fatima-Zahra-Iguenfer/publication/386088433_An_Overview_of_MEAN_MERN_and_LAMP_Stacks_A_Comparative_Analysis_of_Architecture_Cost_Learning_Curve_and_Community_Support/links/67434189b5bd9d17d604845a/An-Overview-of-MEAN-MERN-and-LAMP-Stacks-A-Comparative-Analysis-of-Architecture-Cost-Learning-Curve-and-Community-Support.pdf

[2] L. Rivas, M. Pérez, L. E. Mendoza, and A. C. Grimán P., “Tools selection criteria in software-developing Small and Medium Enterprises,” *J. Comput. Sci. Technol.*, vol. 10, no. 1, 2010, Accessed: Nov. 21, 2025. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/9664

[3] P. M. Jacob and P. Mani, “Software architecture pattern selection model for internet of things based systems,” *IET Software*, vol. 12, no. 5, pp. 390–396, Oct. 2018, doi: 10.1049/IET-SEN.2017.0206;REQUESTEDJOURNAL:JOURNAL:17518814;WEBSITE:WEBSITE:IETRESEARCH;ISSUE:ISSUE:DOI.

[4] B. Baah, O. T.-J. of C. Technologies, and undefined 2013, “Parameters for the Evaluation for the Choice of a Good Programming Language,” *jctjournals.comB Baah, OE TaylorJournal of Computing Technologies, 2013•jctjournals.com*, vol. 2, 2013, Accessed: Nov. 21, 2025. [Online]. Available: http://jctjournals.com/all2013/august2013/v2.pdf

[5] S. Aggarwal and J. Verma, “Comparative analysis of MEAN stack and MERN stack,” *International Journal of Recent Research Aspects*, vol. 5, pp. 127–132, 2018.

[6] M. Shahin, P. Liang, and M. R. Khayyambashi, “Architectural design decision: Existing models and tools,” *2009 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, WICSA/ECSA 2009*, pp. 293–296, 2009, doi: 10.1109/WICSA.2009.5290823.

[7] T. F. Bissyande, F. Thung, D. Lo, L. Jiang, and L. Reveillere, “Popularity, interoperability, and impact of programming languages in 100,000 open source projects,” *Proceedings - International Computer Software and Applications Conference*, pp. 303–312, 2013, doi: 10.1109/COMPSAC.2013.55.

[8] A. S. Hashmi, A. A. Kiani, J. Tariq, H. Rauf, Y. Hafeez, and F. B. Ahmed, "An automated approach for requirement classification using machine learning," *Innovations in Systems and Software Engineering 2025 21:4*, vol. 21, no. 4, pp. 1433–1443, Jun. 2025, doi: 10.1007/S11334-025-00613-5.

[9] V. Mir Khatian, Q. Ali Arain, M. Alenezi, M. Owais Raza, F. Shaikh, and I. Farah, "Comparative Analysis for Predicting Non-Functional Requirements using Supervised Machine Learning," *2021 1st International Conference on Artificial Intelligence and Data Analytics, CAIDA 2021*, pp. 7–12, Apr. 2021, doi: 10.1109/CAIDA51941.2021.9425236.

[10] J. Jia, P. Zhang, and L. F. Capretz, "Environmental factors influencing individual decisionmaking behavior in software projects: A systematic literature review," *Proceedings - 9th International Workshop on Cooperative and Human Aspects of Software Engineering, CHASE 2016*, pp. 86–92, May 2016, doi: 10.1145/2897586.2897589;CSUBTYPE:STRING:CONFERENCE.

[11] A. Yalçıner, A. Dikici, and E. Gökalp, "Data-Driven Software Engineering: A Systematic Literature Review," *Communications in Computer and Information Science*, vol. 2179 CCIS, pp. 19–32, 2024, doi: 10.1007/978-3-031-71139-8_2.

[12] VaroquauxG., BuitinckL., LouppeG., GriselO., PedregosaF., and MuellerA., "Scikit-learn," *GetMobile: Mobile Computing and Communications*, vol. 19, no. 1, pp. 29–33, Jun. 2015, doi: 10.1145/2786984.2786995.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785;CSUBTYPE:STRING:CONFERENCE.

[14] E. Nikulchev, D. Ilin, and A. Gusev, "Technology Stack Selection Model for Software Design of Digital Platforms," *Mathematics 2021, Vol. 9, Page 308*, vol. 9, no. 4, p. 308, Feb. 2021, doi: 10.3390/MATH9040308.

[15] A. Sandaruwan and D. Kasthurirathna, "Recommending Third-Party Libraries and Frameworks Using Topic Modeling in Software Projects," *ICARC 2024 - 4th International Conference on Advanced Research in Computing: Smart and Innovative Trends in Next Generation Computing Technologies*, pp. 37–42, 2024, doi: 10.1109/ICARC61713.2024.10499759.

[16] M. Gasparic and A. Janes, "What recommendation systems for software engineering recommend: A systematic literature review," *Journal of Systems and Software*, vol. 113, pp. 101–113, Mar. 2016, doi: 10.1016/J.JSS.2015.11.036.

[17] Y. Ma, A. Mockus, R. Zaretzki, R. Bradley, and B. Bichescu, "A Methodology for Analyzing Uptake of Software Technologies among Developers," *IEEE Transactions on Software Engineering*, vol. 48, no. 2, pp. 485–501, Feb. 2022, doi: 10.1109/TSE.2020.2993758.

# Ship Route and Weather-Aware Anomaly Detection System with Port-Level Berthing Delay Notifications

R A Wijethunga
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
rawijethunga@students.nsbm.ac.lk

TAH Dilpriya
*Department of Computer and Data Science*
*NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

D K Wijethunga
*Master*
*MSC (Mediterian Shipping Company)*
dhammikawijetunga@yahoo.com

***Abstract*—Maritime operations depend on coordinating vessel movements, environmental conditions, and port schedules. This paper presents a web-based decision-support prototype for Colombo Port, Sri Lanka, integrating Automatic Identification System (AIS) movement analysis, historical weather context, and terminal berthing records. The system uses an Isolation Forest model to flag unusual AIS observations and an explainable rule-based mechanism to identify berthing delays and notify the next eligible vessel in the same berth context. Terminal records from Jaya Container Terminal (JCT), South Asia Gateway Terminals (SAGT) and Colombo International Container Terminals (CICT) were processed separately. The retained dataset comprised 16 JCT, 213 SAGT and 36 CICT records. A subset of 158 SAGT records with valid estimated and actual arrival times was used for historical delay analysis. The AIS methodology exercise processed 4,183 observations from 25 vessels in a publicly available United States coastal dataset because representative Colombo AIS data were unavailable. The Isolation Forest model, configured with 200 estimators and a contamination setting of 0.05, flagged 210 observations (5.02%). These flags do not represent measured detection accuracy because independently labelled anomaly ground truth was unavailable. For historical berth analysis, a delay was defined as an actual arrival more than 30 minutes after the estimated arrival. The detector reproduced 79 delayed and 79 on-time historical outcomes. However, this demonstrates consistency with the rule used to construct the reference labels rather than independent predictive validation. Weather data comprised 1,800 hourly Colombo observations and were used as contextual information rather than evidence of causation. Controlled demonstrations verified the integrated alert workflow. The contribution is an explainable, provenance-aware operational prototype rather than a new machine-learning algorithm or a production-validated port safety system.**

***Keywords—AIS, Maritime anomaly detection, Isolation Forest, Berthing delay, Port monitoring***

## I. Introduction

Maritime transportation is a crucial element in international trade and in global supply chains [1]. Ports arrange inbound and outbound movements of ships, allocate berthing space, arrange for cargo handling and ensure ship departure. These activities can be disrupted, which can impact terminal capacity and vessel schedules. The efficient operation of the terminals is crucial in Sri Lanka's maritime logistics environment, especially in the Indian Ocean ports of Colombo [2].

The Automatic Identification System (AIS) transmits vessel identity and the following data: position, speed over ground (SOG), course over ground (COG) and timestamp. These observations are used for tracking and trajectory analysis of ships. Reporting gaps and equipment related inconsistencies in AIS information exist, however, and these are possible causes of erroneous values [3], [4]. The weather information should give the environmental context, the terminal schedules and actual operational timestamps should give information concerning the arrivals and occupancy of the berth.

While these sources provide complementary information, analyzing alone can reduce operational awareness. This research, thus, proposes a web-based prototype which incorporates AIS movement analysis, weather data and terminal berthing data in Colombo. It alerts on unusual AIS observations, detects operational berthing delays by using intuitive timing rules and provides port-level notifications of impacts for subsequent eligible vessels.

The system looks for the next vessel that is eligible in the same scheduling context the same berth, if a berth is known, to fill the berth after a vessel has been flagged for a berth delay. The notice advises that the anticipated berthing might be impacted. It's not created at the ship-to-ship level, but rather at the port level application and does not imply the subsequent vessel will be delayed.

The contribution is mainly in the integration of heterogeneous maritime data, explainable analytical methods and notifications for operations in a localized research prototype. None of the above are claimed by the study for a new anomaly detection algorithm, a causal relationship between weather and delay, or the performance of the predictive model being validated independently.

## II. Related Work

### *A. AIS Based Anomaly Detection*

There have been studies of anomaly detection based on AIS that employed statistical, clustering and machine-learning methods [4], [5]. Isolation Forest uses recursive random partitions to detect anomalous observations and is appropriate for exploratory anomaly screening when labeled examples are limited [6]. Several other methods have been used to detect anomalies in a maritime context, such as density-based methods [7] and trajectory-oriented models [8].

Quality of AIS data is a key factor to consider. Unusual records can be due to reporting interruptions, equipment issues or data inconsistencies rather than actual vessel maneuvers [3], [9]. An anomaly flag should be viewed as a clue to investigate, not as an "unsafe event."

### *B. Berth Scheduling and Weather Context*

The vessel arrival uncertainty, the berth availability and the terminal resource constraint are considered in Berth allocation and scheduling research [10], [11]. The difference between expected and actual operational timestamps can be valuable clues to watch out for delays. The task of the retrospective threshold classification and the task of the future delay prediction are different tasks. A threshold-based detector is used to determine if a deviation has been observed, but it is not a predictive model unless it has been tested for future outcomes separately.

As shown in maritime research, the environment is relevant to navigation and routing decisions, as displayed in the usage of weather information in maritime research [12]. For the purposes of this research, weather conditions are treated as background conditions, not as evidence of a delay due to the weather condition.

## III. System Design and Methodology

### *A. Research Design and Data Sources*

The prototype development process is iterative and follows a design component that includes problem identification, data preparation, the implementation of the analytical component, system integration and evaluation. It is only linked to Colombo Port and is not linked to the live Sri Lanka Ports Authority operational control systems.

The data from three different terminal datasets has been processed separately due to the different source structure (Table I). A total of 16 JCT, 213 SAGT and 36 CICT records were retained. Of these, 158 SAGT records had valid estimated time of arrival (ETA) and actual time of arrival (ATA); these were used to conduct the historical analysis of arrival delay.

This publicly available United States coastal dataset included 25 vessels with 4,183 processed movement observations for use in the AIS methodology dataset. It was not used to represent Colombo vessel traffic but for the pre-processing and Isolation Forest pipeline. A live AIS collection mechanism was put in place for future local evaluation purposes, but reliable Colombo messages from AIS were not received during the study test periods.

The historical weather data for Colombo was retrieved from the Open-Meteo Historical Forecast API between 1 May and 14 July 2026, which corresponds to 1800 weather observations. The parameters studied were Rainfall, Wind, Visibility and Cloud conditions.

TABLE I
SUMMARY OF DATASETS USED IN THE STUDY

| Data source | Records | Purpose |
|---|---|---|
| JCT | 16 | Terminal Schedule Processing |
| SAGT | 213 | Schedule processing; 158 records used for historical delay analysis |
| CICT | 36 | Terminal schedule processing |
| External AIS | 4,183 from 25 vessels | Methodological anomaly-screening evaluation |
| Colombo weather | 1,800 hourly records | Environmental context |

### *B. AIS Feature Preparation and Isolation Forest*

Vessels and times were sorted using AIS observations. Movement data comprised SOG, in speed, in course and suggested speed. The differences between sequences were computed only if there were at least two observations ordered in time for each vessel. Reporting-time gaps were also explored, and the 95th percentile of the distribution of the gaps was defined as the threshold for unusually large gaps.

The number of estimators used in the Isolation Forest was set to 200 and the number of contaminations to 0.05. The model will output the anomaly score as well as a binary flag for every eligible observation. The contamination setting is not a real-world contamination rate that is measured; it is a rate of outliers the model is expecting.

There was no independent expert labelled AIS anomaly dataset available. Thus, precision, recall, F1 and false positive rate was not possible to be computed. There was no comparative LOF or DBSCAN benchmark performed.

### *C. Berthing-Delay Detection*

The delay was defined as:

$$D_arrival = ATA - ETA \quad (1)$$

If the difference was greater than 30 minutes, it was considered a delayed record; if not, it was considered on time. This threshold was used to analyse the 158 SAGT records.

The system will check the expected completion / departure time (ETC / ETD) against the actual departure time (ATD) or current time the vessel is on berth. If this exceeds 30 minutes, it is considered a berth-delay event (Fig. 1).

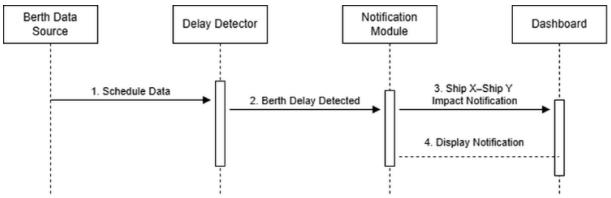



Fig. 1. Sequence diagram for berth-delay notification.

For initial supervised prediction experiments, simple baselines outperformed these prediction experiments. The mean absolute error from regression was 83.8 minutes, which was slightly higher than the baseline of 76.5 minutes. The classification accuracy was 46.2% with 50% as the baseline. The candidate delay features were poorly correlated with delay, with the highest absolute correlation being about 0.13. The last system therefore does not make the failed supervised model's

reliable predictors, but rather implements a transparent rule-based delay detection system instead.

*D. Port-Level Notification and Software Architecture*

If a berth-delay event is found, the system identifies the next earliest eligible berth and/or vessel in the same scheduling context. If the planned berthing is affected, it will be indicated by the notification. Does not guarantee next vessel will have a delay. Generates duplicate notification check to avoid duplicate alert for the same event.

The frontend is based on a React/Vite web app, backend on Node.js/Express REST API, analytical services in Python and the database is MySQL (Fig. 2). A Leaflet map shows the vessel data and alerts. The authentication is done using JSON Web Tokens and bcryptjs. Operational, Demonstration and Historical data modes differentiate between current operational data inputs, controlled test information and historical analytical data.

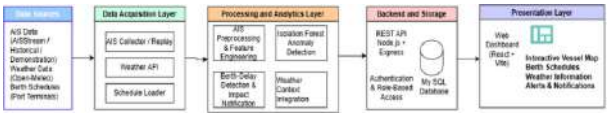



Fig. 2. Implemented architecture of the maritime monitoring and port-level notification prototype.

## IV. Implementation and Results

*A. AIS Anomaly-Screening Results*

The 3,973 observations were deemed to be normal and 210 observations were flagged as anomalies (about 5.02%) by Isolation Forest. This proportion approximates the set contamination value of 0.05, and is not meant to be interpreted as the percentage of anomalies detected or the percentage of anomalies found, as those are not independent assessments.

The movement features that best correlated with course change were all of the course change scores. The example flagged, MMSI 367488370, reported a SOG of 102.3 knots, a change in speed of 97.5 knots and a change in course of 179 degrees. The next observation indicated that the speed had changed to -97.6 knots and the course had changed to 180 degrees. These values could be due to either reporting to AIS or inconsistencies between the different sensors and should undergo domain review before interpreted as actual vessel behaviour.

*B. Historical Berth-Delay Analysis*

The 30-minute ETA – ATA rule was applied to the 158 SAGT records resulting 79 delayed and 79 on-time records. The historical detector reproduced those categories that were derived from rules.

This result is consistent with the predefined rule, as the reference categories are built in a consistent way with the same ETA–ATA relationship and threshold. It is not a completely predictive validation nor it is 100 percent predictive accuracy nor evidence that the system foretells a delay.

*C. Controlled System Demonstration*

A controlled demonstration was conducted to test the berth-occupancy and notification sequence. A synthetic vessel – DEMO BERTH VESSEL ALPHA – was set up to occupy 60 minutes of excess berth at CICT berth DEMO-B01. The system chose DEMO BERTH VESSEL BRAVO as the next eligible vessel and sent a medium impact notice.

An anomaly alert was triggered with high severity by a separate synthetic AIS case with a score of around -0.21718 using the Isolation Forest model. These demonstrations confirmed the integration of database storage, analytics, backend event handling and frontend alert presentation by manipulating the inputs in a controlled manner. They are not used to set the performance in real Colombo vessel traffic.

*D. Weather Context and Integration*

Weather series of 1800 hours were not missing any data following data processing. Approx 71% of observations were made under overcast conditions, with less common were heavy-rain, high wind and low-visibility flags. The correspondence coefficient of the linear correlation between the studied weather features and the berth delay (Fig. 3) was not more than 0.13, both in positive and negative directions. This association is very weak and cannot be used to inform causal claims or for the purposes of a reliable weather-based delay prediction. Weather is thus displayed as a contextual detail together with the vessel events and berth events.

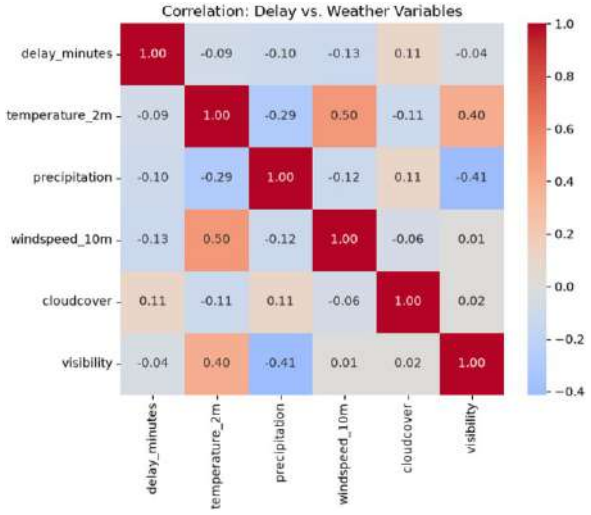



Fig. 3. Correlation between arrival delay and weather variables.

End-to-end demonstrations included storage on a database, database analytics execution, backend handling of events, and presentation of alerts on the frontend. Workflow features data provenance modes, event severity, next-vessel selection and duplicate-notification suppression. No production trials or expert assessment of the quality of alerts was performed, there was no independent vessel-time separated validation, and the false-alarm and notification-latency were not measured.

## V. Limitations and Discussion

The main drawback is the lack of representative and reliable Colombo AIS observations over the study period. An external AIS dataset is used to support the pipeline development but is geographically and operationally different from the target port. Moreover, these anomaly flags are not labeled by an expert and thus do not have conventional detection metrics and

comparisons with other algorithms. Moving forward, validation should be carried out with reliable Colombo AIS data, expert validated event labels, and vessel and time separated evaluation.

The lack of prospective rule-based reference labels is a limitation of historical berth analysis. Predictive performance is established with independent operational records and held out evaluation periods. Future work should report false alarms, detection lead time, correctness of next vessel selection, notification latency and user acceptance.

The field of weather analysis is restricted from observation period and poor correlation. More aligned time series and appropriate statistical controls need to be applied before predictive and/or causal interpretations can be made.

The prototype is a decision support tool for research. It doesn't immediately influence the work of SLPA, ship-to-ship communication, navigational safety certification and the substitution of professional judgment. The data available for the evaluation indicates that the technology is technically feasible under historical and controlled demonstration settings, but not for production.

## VI. Conclusion

This paper described a prototype of maritime monitoring application being developed for the Colombo region, combining the integration of AIS anomalies, rule-based berthing-delay detection, weather context and notifications of impacts on port-level. Of the 4,183 external AIS observations, 210 were flagged by the Isolation Forest algorithm and the berth rule flagged 79 of the 158 SAGT observations as delayed. The results reported are valid only for the current database and are not labelled as anomalies in the case of AIS reports and were also checked for consistency with the rule that created the reference categories in the case of the historical berth result.

Integrated alert workflow was proven in controlled demonstrations, including selection of the next vessel in the same berth context. The contribution consists of an explainable, provenance-aware system design for further evaluation in the context of Colombo. Future activities will be focused on reliable local AIS collection, independent expert labels, held-out temporal evaluation, baseline comparisons and quantitative notification-performance evaluation.

## Acknowledgment

The authors would like to express their gratitude to NSBM Green University for academic support and resources given for this research. The authors also thank the academic and maritime domain supervisors, and the public accessibility of AIS and weather data applied in the methodological and contextual analysis.

## References

[1] United Nations Conference on Trade and Development, "The Review of Maritime Transport 2024 : Navigating Maritime Chokepoints," United Nations Publications, Geneva, 2024.

[2] Sri Lanka Ports Authority, "SLPA Terminals Record All-Time High of 2.41 Million TEUs in 2024 Amidst Competition," 2025. [Online]. Available: https://www.slpa.lk/port-colombo-page/cd37c8690c1bac2e056e25279e89d5fa/0234b49fd252e53a0d988619ee47bdd964bf4e89.

[3] International Maritime Organization, "AIS transponders," [Online]. Available: https://www.imo.org/en/ourwork/safety/pages/ais.aspx.

[4] L. R. J. B. K. W. Konrad Wolsing, "Anomaly Detection in Maritime AIS Tracks," Journal of Marine Science and Engineering, vol. 10, 2022.

[5] C. V. Ribeiro, A. Paes and D. de Oliveira, "AIS-based maritime anomaly traffic detection: A review," Expert Systems with Applications, vol. 231.

[6] L. F. T., K. M. Ting and . Z.-H. zhou, "Isolation Forest," 2008.

[7] M. Ester, H.-P. Kriegel, J. Sander and . X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996.

[8] D. Nguyen, R. Vadaine, . G. Hajduch, R. Garello and R. Fablet, "GeoTrackNet-A Maritime Anomaly Detector using Probabilistic Neural Network Representation of AIS Tracks and A Contrario Detection," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 5655 - 5667, June 2022.

[9] A. Harati-Mokhtari, A. Wall, P. Brooks and J. Wang, "Automatic Identification System (AIS): Data Reliability and Human Error Implications," Journal of Navigation, vol. 60, no. 3, pp. 373-389, 2007.

[10] F. Rodrigues and A. Agra, "Berth allocation and quay crane assignment/scheduling problem under uncertainty: A survey," European Journal of Operational Research, vol. 303, no. 2, pp. 501-524, 2022.

[11] B. Li, Z. Elmi, A. Manske, E. Jacobs, Y.-y. Lau, Q. Chen and M. A. Dulebenets, "Berth allocation and scheduling at marine container terminals: A state-of-the-art review of solution approaches and relevant scheduling attributes," Journal of Computational Design and Engineering, vol. 10, no. 4, p. 1707–1735, 2023.

[12] A. Kytariolou and N. Themelis, "Ship routing optimisation based on forecasted weather data and considering safety criteria," Journal of Navigation, pp. 1310-1331, 2022.

# An Adaptive Data Prioritization Framework Inspired by Neural Attention Mechanisms for Handwritten Image-Based Evaluation: A Case Study in Forensic Handwritten Text Recognition

Imalsha sandali Gedara
*NSBM Green University*
*Homagama, Sri Lanka*
snmgedara@students.nsbm.ac.lk

Dilpriya TAH
*NSBM Green University*
*Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Convolutional Neural Networks (CNNs) are widely used for image learning, but conventional training does not explicitly coordinate spatial features, sample importance, and learning stages. This research proposes an Adaptive Data Prioritization Framework integrating spatial attention, loss-based sample prioritization, and phase-aware temporal adaptation to enhance CNN learning. The framework was evaluated using MNIST, EMNIST Digits, and EMNIST Letters through classification, ablation, and robustness experiments. Results showed improved MNIST accuracy from 98.97% to 99.12%, reduced test loss, and stronger performance under image corruption. For forensic handwritten text recognition, an applied case study, the final V3 model achieved 10.11% test CER and 31.46% WER. It was integrated into an FHR forensic review support prototype. Overall, findings provide qualified evidence of improved CNN learning behavior, although further domain-specific validation is required to establish wider applicability.**

***Keywords—Adaptive Data Prioritization, Neural Attention, CNN Training, Handwritten Image Evaluation, Forensic Handwritten Text Recognition***

## I. Introduction

### *A. Research Background and Problem Statement*

Convolutional Neural Networks (CNNs) are widely used in image-based learning because they automatically extract meaningful visual features. However, conventional mini-batch training does not explicitly coordinate spatial feature importance, individual sample difficulty, and changing learning priorities throughout training. This limitation is particularly relevant to handwriting images, which vary in stroke formation, clarity, noise, and visual ambiguity. Although existing approaches such as spatial attention, hard-example mining, focal loss, and curriculum learning address different aspects of adaptive learning, their coordinated application within a unified CNN training framework requires further investigation.

This research proposes an Adaptive Data Prioritization Framework inspired by neural attention mechanisms, integrating spatial attention, loss-based sample prioritization, and phase-aware temporal adaptation. The framework aims to improve CNN learning behaviour by identifying informative spatial regions, assigning greater importance to challenging training samples, and adjusting sample priorities across early, middle, and late training stages.

The main aim is to design, implement, and evaluate the proposed framework using handwritten character images as the experimental evaluation context. The research evaluates classification performance, generalization, and robustness through controlled baseline comparisons, ablation experiments, and corruption testing.

The specific objectives are to:

1. Investigate per-sample loss as an indicator of training difficulty and determine how sample priorities should change throughout training.
2. Compare baseline CNN, individual framework components, static combinations, and the final adaptive framework using MNIST, EMNIST Digits, and EMNIST Letters.
3. Evaluate robustness against Gaussian noise, salt-and-pepper noise, and Gaussian blur.
4. Extend the adaptive learning principles to forensic handwritten text recognition (HTR) through iterative V1, V2, and V3 sequence-model development.
5. Integrate the final HTR model into a forensic review support prototype.

The proposed solution combines controlled CNN experiments with a line-level forensic HTR case study using predefined IAM dataset partitions (Fig. 1). The final V3 sequence model incorporates target-length-normalized per-sample CTC loss, packed BiLSTM processing, Batch Normalization, and fresh model initialization. Model selection is based on minimum validation Character Error Rate (CER), followed by independent test evaluation using CER and Word Error Rate (WER).

The selected model is integrated into an FHR prototype supporting examiner authentication, handwriting uploads, spatial-attention visualization, recognized text, confidence information, Top-3 predicted transcriptions, examiner observations, report generation, and audit logging. The research focuses on handwritten character classification, adaptive CNN training, controlled robustness evaluation, and line-level HTR. Writer identification, authorship attribution, signature verification, autonomous forensic decisions, and production-scale deployment are excluded.

## II. Related Work

CNNs have been widely adopted in visual recognition because convolutional layers can learn local visual patterns and combine them into higher-level representations. In handwritten image experiments, CNNs are commonly used to learn edges, curves, loops, strokes, and other spatial features directly from pixels. Benchmark datasets such as MNIST and EMNIST remain useful because they provide stable settings for comparing architectures and training strategies under repeatable conditions.

Despite their success, benchmark performance alone is insufficient for understanding model robustness. High accuracy on clean handwritten images may not indicate reliable behaviour on degraded, ambiguous, or low-quality document images. This is important because many real-world recognition settings include samples that are not equally useful for training. Conventional averaging of batch loss can hide the contribution of individual difficult samples, causing the model to update its parameters without considering which samples should contribute more strongly at a given training stage.

Attention mechanisms allow neural networks to focus on informative features while suppressing irrelevant information. The Transformer demonstrated the power of self-attention for modelling dependencies [1], while Squeeze-and-Excitation networks showed that attention can improve feature recalibration across channels [2]. CBAM extended this by combining channel and spatial attention to highlight both what and where the model should attend [3].

For image classification, spatial attention is particularly relevant because it generates attention maps over image regions. In handwritten image evaluation, such maps can help the network focus on strokes and character regions rather than background noise. Attention also provides a useful explanation layer because highlighted areas can be displayed to show which visual regions influenced a prediction. However, most attention methods focus on feature selection and do not directly solve the sample-level training problem of deciding which examples should contribute more strongly during optimization.

Curriculum learning proposed the idea that training can benefit from ordering samples from easier to harder examples [4]. Hard example mining and related losses address sample difficulty by placing more weight on difficult samples. Focal Loss down-weights easy samples and focuses learning on hard examples [5], while hard mining loss applies a similar principle through adaptive loss scaling [6]. Robust learning methods such as co-teaching also show that not all samples should influence training equally [7].

Handwritten text recognition is closely related to document image analysis. Sequence-based recognition methods such as connectionist temporal classification support recognition from unsegmented sequences [9], while handwriting databases such as IAM provide scanned handwritten text images for offline recognition research [10]. Recent attention-based document recognition approaches further show the relevance of visual focus and layout-aware learning for handwritten document processing [12].

### *A. Research Gap and Position of the Present Study*

The reviewed literature establishes the individual components of this research but shows that spatial attention, sample prioritization, and curriculum-based learning are generally investigated separately. Attention methods focus on feature or region importance, while hard-example mining, reweighting, and curriculum learning adjust sample contribution and training order. Recent HTR research explores attention-based recognition, document-level processing [12] and structural handwriting information for character recognition. However, limited research directly investigates combining spatial attention with transparent, phase-aware sample prioritization for handwriting evaluation.

The identified research gap is therefore the limited integration and controlled evaluation of two adaptive levels: spatial emphasis within handwriting images and phase-dependent prioritization across training samples. Unlike approaches requiring multiple networks [7] or validation-guided meta-optimization this research investigates a simpler framework combining lightweight spatial attention, per-sample loss, and training-phase adaptation. Its effectiveness is assessed through baseline comparisons, component-wise ablation, robustness testing, and visual analysis. The contribution is positioned as an experimentally evaluated integration of existing techniques rather than an entirely new attention mechanism.

## III. Proposed Framework

Fig. 1. Research workflow.

The proposed framework is designed to improve CNN learning behaviour by combining three mechanisms (Fig. 2). Spatial attention is used to focus on important handwritten image regions. Loss-based sample prioritization is used to increase the contribution of difficult samples during training. Temporal adaptation is used to change the prioritization rule across training phases instead of applying the same weighting strategy throughout the entire process.

### *A. Spatial Attention*

The spatial attention component is applied to feature maps produced by the convolutional layers. It generates an attention map using pooling, convolution, and sigmoid activation, and then multiplies the attention map with the feature map. This process emphasizes relevant spatial regions while reducing the influence of less informative regions. In handwritten image evaluation, this can correspond to focusing on stroke boundaries, loops, curves, and other discriminative visual structures.

*B. Loss-Based Sample Prioritization*

Instead of computing only a single averaged batch loss, the framework computes individual loss values for each sample using CrossEntropyLoss with reduction set to none. These per-sample losses are transformed into priority scores so that difficult samples contribute more strongly to the weighted loss. This mechanism is intended to prevent easy examples from dominating training and to give the model additional learning signal from samples that remain confusing or visually degraded.

*C. Temporal Adaptation*

Temporal adaptation changes the priority strategy across the training process. During early epochs, priority values remain stable so that the model can learn basic features without being dominated by difficult or noisy samples. During middle epochs, the model shifts more attention toward confusing samples. During late epochs, stronger priority is given to persistent difficult samples so that decision boundaries can be refined. This phase-aware strategy distinguishes the proposed framework from static sample weighting methods.

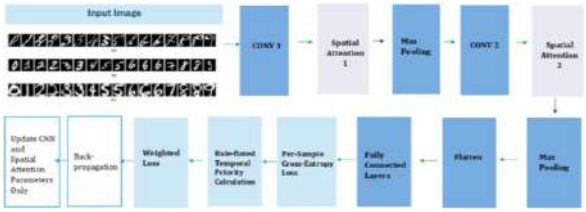



Fig. 2. Final adaptive framework architecture.

## IV. METHODOLOGY

The methodology used to develop and evaluate the Adaptive Data Prioritization Framework and its forensic HTR application. A quantitative, positivist-leaning and deductive approach was adopted through controlled experiments, component-wise ablation, generalization testing, and robustness evaluation. MNIST, EMNIST Digits, and EMNIST Letters were used for handwritten image classification, while IAM-line was selected for sequence recognition. Fixed dataset partitions, seed 42, and consistent preprocessing supported reproducibility. Classification experiments evaluated accuracy and loss, while HTR performance was assessed using Character Error Rate (CER) and Word Error Rate (WER). The final model was integrated into the FHR review-support prototype.

Five configurations are used where appropriate: Baseline CNN, CNN with Spatial Attention, CNN with Loss-Based Prioritization, CNN with Attention and Prioritization, and the Final Adaptive Framework.

The same evaluation logic is applied across datasets to make component behavior clear. Models are evaluated using accuracy, loss, precision, recall, F1-score, and training time. Robustness is evaluated using degraded versions of the handwritten images with Gaussian noise, salt-and-pepper noise, and blur.

The CNN architecture is intentionally kept simple because the purpose is to evaluate the effect of adaptive prioritization rather than to depend on a very large pretrained backbone. The baseline CNN used two padded 3 × 3 convolutional layers with ReLU and 2 × 2 max pooling (1→32→64 channels), followed by a 128-unit fully connected layer and class output (Fig. 3). It was trained with CrossEntropyLoss and Adam (learning rate 0.001).

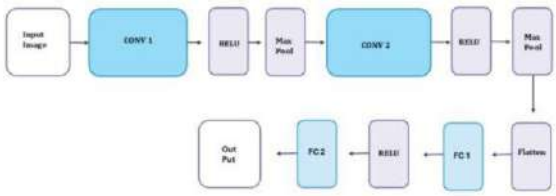



Fig. 3. CNN Architecture

Spatial attention followed both convolutional layers (Fig. 4). Channel-wise average and max pooling outputs were concatenated, passed through a 7 × 7 convolution and sigmoid, and the resulting 0–1 map reweighted feature locations.

$$A = \sigma(Conv_{7\times7}([AvgPool_c(F)\,;\,MaxPool_c(F)]))$$

$$F' = F \odot A$$

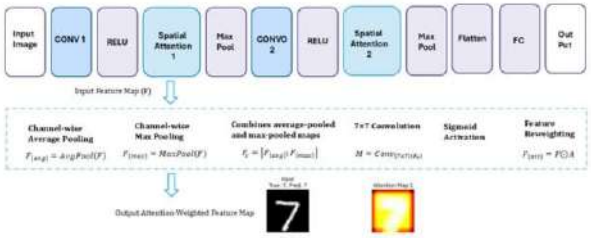



Fig. 4. CNN with spatial attention

Loss-based prioritization extended importance to training samples. CrossEntropyLoss used reduction set to none, preserving per-sample losses that were weighted by detached priority scores and averaged for backpropagation.

$$L_{weighted} = \frac{1}{N}\sum_{i=1}^{N} L_i P_i$$

Temporal adaptation varied sample importance by training stage. Per-sample losses were min-max normalized within each mini-batch and converted to early, middle and late priorities, clamped between 0.3 and 1.0 Early training stayed near-uniform, favoring easier samples; later phases progressively emphasized harder ones. Epochs 1–3, 4–7 and 8–10 formed a 30%–40%–30% schedule. The final framework combines all three components

$$n_i = \frac{L_i - L_{min}}{L_{max} - L_{min} + \varepsilon}$$

$$Early: P_i = 0.8 + 0.2(1 - n_i)$$

$$Middle: P_i = 0.5 + 0.5n_i$$

$$Late: P_i = 0.3 + 0.7n_i$$

All variants trained for ten epochs using Adam, batch size 64 and seed 42, the final model added weight decay of $1 \times 10^{-4}$. The best validation-accuracy checkpoint alone was tested. Ablation compared baseline, attention-only, prioritization-only,

static combined and final models under identical splits, and generalization used EMNIST Digits and Letters. Robustness applied Gaussian noise, salt-and-pepper noise and blur without retraining.

The forensic HTR extension (IAM-line) used CNN, spatial attention, BiLSTM and CTC decoding. Per-sample CTC loss, normalized by target length, replaced cross-entropy for priorities, and performance was evaluated using CER and WER.

$$L_i^{len} = \frac{L_i^{CTC}}{|y_i|}$$

## V. Results and Discussion

Evaluation progressed from MNIST ablation to EMNIST generalization, controlled corruption testing and line-level forensic handwritten text recognition (HTR), all on fixed held-out partitions. For HTR, the checkpoint was selected on IAM-line validation CER, and the independent test set was evaluated only afterwards.

### *A. Classification Ablation (MNIST)*

The baseline CNN reached 98.97% accuracy and 0.041309 test loss. Spatial attention kept the same accuracy but reduced loss to 0.032444 (about 21.5%), while loss-based prioritization reached 99.08% accuracy and 0.033841 loss. The static combination of both fell to 98.76% accuracy and 0.039576 loss, showing that combining feature and sample reweighting without phase control is unstable.

The Final Adaptive Framework achieved 99.12% accuracy and 0.0299 loss (Fig. 5), a loss reduction of about 27.6% against baseline and 24.4% against the static combination. This supports temporal adaptation as the decisive component. Training time rose from 146.21 s to 166.23 s, a measurable overhead that reflects attention processing, per-sample loss calculation and temporal weighting. Because baseline accuracy was already above 98.9%, the lower loss is an important complement to the modest gain in correct predictions.

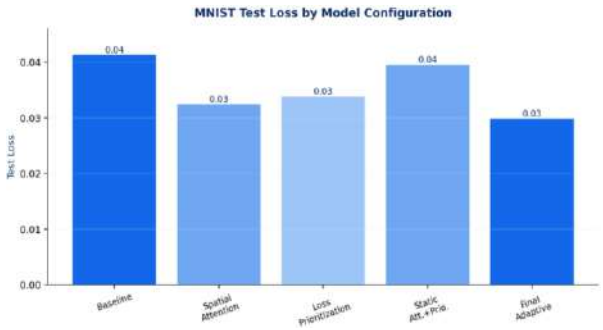



Fig. 5. MNIST test loss across model configurations.

### *B. Generalization (EMNIST)*

EMNIST evidence was mixed (Table I). On Digits, accuracy was effectively unchanged (99.45% to 99.44%) while loss fell about 40.3%. On Letters, accuracy dipped by 0.09 points while loss improved about 4.2%. The benefit therefore appeared mainly as lower loss, and depends on performance saturation and class ambiguity. Letter errors concentrated in visually similar classes such as i, l, g and q (F1 between 0.74 and 0.85), which motivates sequence context in HTR. These results prevent overgeneralization from MNIST: temporal adaptation changes learning behavior but does not guarantee the highest accuracy on every dataset.

TABLE I. Baseline vs. Final Adaptive Framework Across Datasets

| Dataset | Baseline Acc. (%) | Baseline Loss | Final Acc. (%) | Final Loss |
|---|---|---|---|---|
| MNIST | 98.97 | 0.0413 | 99.12 | 0.0299 |
| EMNIST Digits | 99.45 | 0.0308 | 99.44 | 0.0184 |
| EMNIST Letters | 93.33 | 0.2192 | 93.24 | 0.2100 |

### *C. Noise Robustness*

Robustness gave the most consistent accuracy advantage. The framework exceeded baseline at all eleven Gaussian-noise, salt-and-pepper and blur settings (Fig. 6), averaging about 0.45 percentage points, with the largest gain of 1.36 points at salt-and-pepper level 0.20 (91.91% to 93.27%). Under Gaussian blur at kernels 5 and 7, accuracy was higher but loss was also higher, so robustness cannot be judged by one metric. The experiment does not isolate which component produces the gain, so it is attributed to the complete adaptive configuration rather than to attention or prioritization alone.

### *D. Forensic HTR*

V1 and V2 failed to learn line transcription, remaining near 95.95% CER and 98% WER even after extended training. V3, with length-normalized CTC prioritization, packed BiLSTM processing, Batch Normalization and fresh initialization, reached 6.58% validation CER and 23.75% WER at epoch 38 (Fig. 7). The independent test gave 10.11% CER and 31.46% WER. WER stays higher because one character error invalidates a whole word. Since several corrections were introduced together, a matched sequence baseline is needed to quantify the contribution of adaptive weighting.

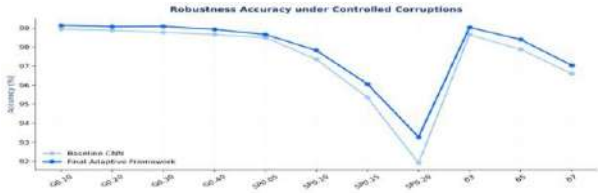



Fig. 6. Baseline and framework accuracy under controlled corruptions (G: Gaussian, SP: salt-and-pepper, B: blur).

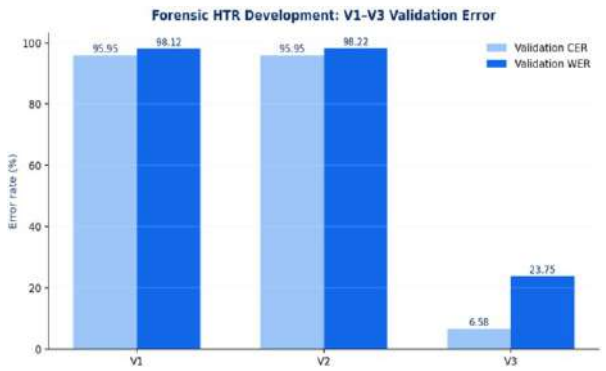



Fig. 7. Validation CER and WER across forensic HTR versions V1 to V3.

## VI. Prototype Application Interface

The Forensic Handwritten Text Recognition (FHR) prototype places the V3 model inside a structured, human-reviewed workflow (Fig. 8). An examiner authenticates, enters a Case ID, uploads a handwriting image (PNG, JPG, JPEG, TIFF or BMP) and previews it before recognition. Access control and case association therefore precede any model output, keeping each recognition result tied to a named examiner and case. The Analysis Dashboard presents the original image, preprocessing output, spatial-attention overlay, recognized text, a recognition-confidence indicator and Top-3 Predicted Transcriptions.

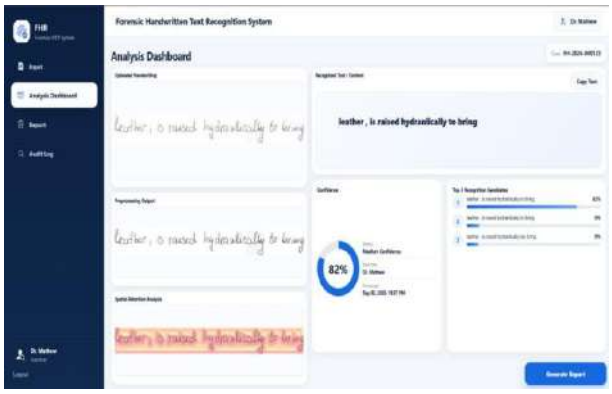



Fig. 8. FHR Analysis Dashboard with recognized text, confidence indicator, Top-3 transcriptions and attention overlay.

Ten functional test cases (TC-01 to TC-10) passed, covering valid and invalid authentication, supported and unsupported uploads, recognition, evidence display, observations, export and audit history. The confidence indicator is not a calibrated probability, and attention maps are diagnostic visualizations rather than evidence of correctness. The report is an application record of model output, not an independent validation certificate

## VII. Practical Implications

The adaptive framework is a training-stage enhancement that can be added to CNN pipelines where sample difficulty and image quality vary. Its cost is moderate: training time rose by about 14% on MNIST and 7% on EMNIST Digits, with no architectural change to the deployed classifier. The consistent corruption gains are relevant to scanned documents, where pixel noise, artefacts and blur are common. Its largest gains appeared under severe salt-and-pepper corruption, where baseline accuracy fell furthest.

The FHR prototype shows how recognition can support initial transcription, documented review and traceable reporting. Suitable users are organizations processing handwritten documents where a qualified person verifies model output. Possible benefits include less manual preparation and more consistent records, although these were not measured in this study. The framework may also transfer to other CNN-based image domains, including industrial inspection, remote sensing, agricultural imaging, document analysis and medical imaging. However, a high-loss sample may be an informative difficult case, poor image quality, distribution shift or annotation error. Adoption should therefore rely on domain-specific baselines, controlled ablation, appropriate metrics, robustness testing and independent validation.

In practice, outputs should be treated as decision support. Recognition confidence and attention maps are review aids, and examiner judgment must remain the final authority, particularly where transcriptions may carry evidential weight.

## VIII. Limitations and Future Work

Most experiments report single runs rather than repeated seeds, so small accuracy differences must be interpreted cautiously. The 10-epoch budget and the early-middle-late schedule were not exhaustively tuned. MNIST and EMNIST Digits are near saturation, and EMNIST accuracy was not improved, so benefits are dataset-dependent. Synthetic corruptions cover only part of real document degradation. IAM-line is a handwriting benchmark rather than a forensic case distribution, and evaluation was limited to line level, without full-page segmentation. Test WER of 31.46% confirms that exact word transcription remains imperfect, and V3 lacks a matched baseline without adaptive prioritization. The prototype has no formal penetration, load or examiner usability study, and audit logging does not establish a validated chain of custody. Inference latency, memory use and throughput were also not benchmarked.

Future work should repeat key ablations across multiple seeds with statistical comparison, test alternative CNN backbones, and compare difficulty signals such as loss history, prediction margin, uncertainty and gradient information. Learnable or validation-driven schedules should be tested against the fixed schedule, and a matched V3 baseline should isolate the effect of adaptive weighting.

## IX. Conclusion

This study developed and evaluated an Adaptive Data Prioritization Framework that coordinates spatial attention, per-sample difficulty weighting and temporal adaptation. The strongest controlled evidence came from MNIST, where accuracy improved from 98.97% to 99.12% and test loss fell from 0.041309 to 0.0299, while the framework recovered from the weaker static combination (98.76%). Higher accuracy under all eleven corruption settings supports robustness within the tested conditions. EMNIST qualifies these findings: the framework reduced loss but gave marginally lower accuracy on both Digits and Letters. It should therefore be regarded as a dataset-dependent training enhancement, not a guarantee of

maximum accuracy. The V3 HTR model achieved 10.11% independent test CER and 31.46% WER and was integrated into the FHR review-support prototype with confidence indication, Top-3 transcriptions, attention visualization, reporting and audit logging. Residual transcription error, uncalibrated confidence, the missing sequence-level ablation and the lack of operational forensic validation limit stronger claims, and the prototype supports qualified human review rather than replacing it.

With domain-specific validation, the approach may offer a basis for adaptive CNN learning in other image tasks where training samples vary in difficulty, quality or informational value. Repeated-run experiments and broader ablation studies remain necessary to establish the reliability and consistency of the observed improvements.

## REFERENCES


[1] Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141.

[3] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision, 2018, pp. 3-19.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 41-48.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980-2988.

[6] Y. Srivastava, V. Murali, and S. R. Dubey, "Hard-mining loss based convolutional neural network for face recognition," in International Conference on Computer Vision and Image Processing, Springer, 2020, pp. 70-80.

[7] Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," Advances in Neural Information Processing Systems, vol. 31, 2018.

[8] Ouyang et al., "Efficient multi-scale attention module with crossspatial learning," in ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2023, pp. 1-5.

[9] Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 369-376.

[10] U.-V. Marti and H. Bunke, "The IAM-database: An English sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, pp. 39-46, 2002.

[11] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, 2017.

[12] M. Hamdan, A. Rahiche, and M. Cheriet, "HAND: Hierarchical attention network for multi-scale handwritten document recognition and layout analysis," arXiv preprint arXiv:2412.18981, 2024.

# AI-Based Real-Time Workout Posture Correction System

L.S.A. De Silva
*Department of Computer and Data Science*
*NSBM Green University*
*Sri Lanka*
senethmadesilva@gmail.com

G. Perera
*Department of Computer and Data Science*
*NSBM Green University*
*Sri Lanka*
gayanp@nsbm.ac.lk

***Abstract*—Correct exercise form is important to prevent injury and maximize the effectiveness of a workout, yet many athletes exercise without access to a personal trainer due to cost, availability, or simply preferring working out at home. This paper presents an affordable, privacy preserving posture correction and repetition counting system that runs entirely on local hardware using a standard webcam, with no wearable or cloud processing. A custom dataset was created from multiple participants in background lighting with diverse body structures, heights, and clothing. The pipeline uses YOLOv8 to isolate the user's region of interest (ROI) and MediaPipe Pose to extract skeletal landmarks. Dynamic joint angles are computed from these landmark feeds and passed to a machine learning classifier to evaluate form while a Finite State Machine (FSM) handles repetition counting. Several supervised algorithms were evaluated using 5-fold cross-validation on a SMOTE-balanced dataset, followed by testing on continuous live video. Random Forest achieved a mean cross-validation accuracy of 84.02% across three posture classes (Correct Form, Lower Your Arm, Overextension). The repetition counting module achieved 80.65% accuracy with a Mean Absolute Error (MAE) of 1.5 reps during live video testing. These results suggest that local camera-based posture correction can offer a practical and personal alternative to professional supervision, giving athletes a way to train safely and confidently at home.**

***Keywords—Artificial Intelligence (AI), Computer Vision, Human Pose Estimation, Real-Time Posture Correction, Machine Learning***

## I. Introduction

Regular physical activity supports both physical and mental health, and people have started doing at-home fitness training in recent years. Home workouts are convenient, but they frequently lack real-time professional supervision. Without a trainer, small errors in form such as a hyper-extended elbow, a shoulder out of alignment, can potentially carry a real risk of injury that affects exercise quality and movement execution[1], [2].

Automated posture correction has relied on wearable Inertial Measurement Units (IMU) and Internet of Things (IoT) sensors [3] . These are accurate but expensive, and often need calibration before use. Computer vision offers markerless pose estimation using a standard RGB camera. Tools like YOLO and MediaPipe have already been applied to track exercises such as powerlifting [4] . However, existing systems rely on cloud-based deep learning [3], [5], which requires streaming workout video to external servers raising significant privacy concerns [6] and many of these systems are built to classify a pose at a single instant rather than track a continuous repetition.

This paper addresses these problems by proposing a markerless posture correction framework that runs entirely on local hardware, with no wearables or cloud dependency. The system combines YOLOv8 [7] with MediaPipe Pose to isolate the person and mark landmarks. A 15-dimensional feature vector is created from normalized landmark coordinates, geometric measurements, and frame-to-frame displacement features to drive a Random Forest classifier for posture evaluation combined with a Finite State Machine (FSM) for accurate repetition counting.

This study focuses on the bicep curl as an initial exercise. Future work will extend to compound movements, likely requiring temporal models such as LSTMs, and broader human movement tracking will be explored. Finally, to maximize the user's trust and clinical transparency, future systems will provide Explainable AI (XAI) insights that allow users to understand the precise biomechanical variables driving their corrective feedback.

## II. Related Work

Wearable sensor-based approaches using Inertial Measurement Units (IMUs) and Internet of Things (IoT) devices have been explored for exercise monitoring and posture correction. Hannan et al. [3] presented a portable smart fitness system for real-time exercise monitoring. However, wearables require physical sensors and involve considerations such as placement and calibration. Camera-based systems offer an alternative by analyzing body movements without additional devices.

Research has shifted toward markerless pose estimation using ordinary cameras. Chu et al. [8] showed the importance of modeling structural relationships between body joints, while MediaPipe provides pose landmark detection and tracking capabilities for camera-based applications [9]. However, camera positioning, occlusion, and lighting conditions can affect landmark reliability. Oliosi et al. [10] highlighted the importance of recording conditions in pose estimation by investigating the influence of smartphone camera positioning on AI-based exercise detection.

Several studies have applied computer vision to exercise monitoring and posture correction. Chen and Yang [11] proposed the Stanford Pose Trainer, which combines pose estimation, geometric representations, and Dynamic Time

Warping to assess exercise movements. Kanase et al. [12] investigated pose estimation for posture correction, while Hari Ram and Ramalingam [13] developed a computer vision-based exercise trainer using OpenCV and MediaPipe. Ko et al. [4] introduced a two-stage approach combining YOLOv5-based person detection and MediaPipe pose estimation for powerlifting posture correction. These studies demonstrate the potential of vision-based exercise assessment, although their exercise scopes and evaluation methods differ.

Posture classification and repetition counting are distinct but complementary tasks. Posture classification evaluates movement correctness, while repetition counting identifies completed exercise cycles through temporal tracking of movement phases. For instance, bicep curls can be tracked through transitions between extended and flexed arm positions using elbow angles. However, angle-based cycle detection alone cannot establish whether a repetition was performed correctly. Combining posture predictions with temporal tracking therefore enables both repetition completion and posture validity to be assessed.

This research proposes a webcam-based framework integrating exercise posture classification and posture-aware repetition validation. It combines YOLOv8-based person detection, MediaPipe Pose landmark extraction, engineered movement features, and a Random Forest classifier to identify three posture categories: Correct Form, Lower Your Arm, and Overextension. An angle-based Finite State Machine (FSM) monitors bicep-curl movement phases and uses posture predictions to distinguish completed cycles from valid repetitions. The main contribution is this integration of frame-level posture assessment and temporal repetition validation within a local webcam-based system that operates without wearable sensors.

## III. Methodology

### A. System Architecture Overview

The proposed system is a framework for real-time bicep curl posture correction and repetition counting. It processes video frames locally using YOLOv8 for person detection and isolation, MediaPipe Pose for landmark estimation, and a Random Forest (RF) classifier for posture assessment. The posture feedback is displayed through the user interface and used by an FSM to track movement phases and validate repetitions. The overall architecture is shown in Fig. 1.

### B. Video Processing and Pose Estimation

The system captures continuous RGB video frames using a standard webcam at a resolution of 640×480 pixels. OpenCV is used for frame acquisition and image processing. Each frame is processed locally through the following stages.

*1)* Person Detection: YOLOv8 [7] is used to detect the person within each input frame. The detector produces a bounding box represented as $B = [x_1,y_1,x_2,y_2]$ where $(x_1,y_1)$ and $(x_2,y_2)$ represent the upper-left and lower-right coordinates of the detected region, respectively. The detected bounding box is used to extract the ROI for subsequent pose estimation. A confidence threshold of 0.4 is applied to detect the person. If a valid person detection is unavailable in a frame, they are excluded from the landmark-based processing stage.

*1)* Pose Estimation: The cropped image is passed to MediaPipe Pose [9] , which estimates 33 body landmarks (Fig. 2) $p_i = (x_i,y_i,z_i)$ with visibility score $v_i$ where $x_i$ and $y_i$ represent normalized image coordinates, while $z_i$ represents the estimated relative depth.

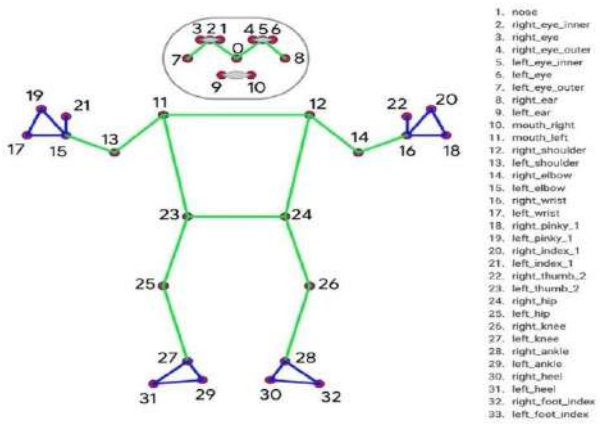



Fig. 2. MediaPipe Pose landmark topology used for body landmark extraction

For the bicep-curl, the shoulder ($p_s$), elbow ($p_e$), wrist ($p_w$), and hip ($p_h$) landmarks are used. When calculating shoulder alignment, the hip landmark is used, while the shoulder, elbow, and wrist landmarks provide the principal information for arm movement analysis. To determine whether the required landmarks are sufficiently visible for feature extraction, a minimum landmark visibility threshold of 0.4 is used.

### C. Feature Extraction

The detected landmarks are used to construct a 15-dimensional feature vector representing body configuration and movement. It has landmark coordinates, joint angles, relative vertical position, and frame-to-frame displacement.

$$= \cos^{-1}\left(\frac{(\mathbf{s}_{2D}-\mathbf{e}_{2D})\cdot(\mathbf{w}_{2D}-\mathbf{e}_{2D})}{\|\mathbf{s}_{2D}-\mathbf{e}_{2D}\|\|\mathbf{w}_{2D}-\mathbf{e}_{2D}\|}\right)\frac{180}{\pi} \quad (1)$$

$$= \cos^{-1}\left(\frac{(\mathbf{s}_{2D}-\mathbf{e}_{2D})\cdot(\mathbf{w}_{2D}-\mathbf{e}_{2D})}{\|\mathbf{s}_{2D}-\mathbf{e}_{2D}\|\|\mathbf{w}_{2D}-\mathbf{e}_{2D}\|}\right)\frac{180}{\pi} \quad (1)$$

The elbow angle is calculated using the shoulder, elbow, and wrist landmarks in normalized 2D image coordinates:

$\theta_{\text{elbow}}$

where $\mathbf{s}_{2D}$, $\mathbf{e}_{2D}$, and $\mathbf{w}_{2D}$ represent the normalized 2D coordinates of the shoulder, elbow, and wrist, respectively.

The shoulder alignment angle is calculated using the shoulder-to-hip vector ($\mathbf{h}_{2D} - \mathbf{s}_{2D}$) and the vertical reference vector $(0,1)$ in the image plane. The relative vertical difference between the shoulder and wrist is also included as $\Delta y = y_s - y_w$, where $y_s$ and $y_w$ are their normalized vertical coordinates.

Additionally, three displacement features are extracted for the shoulder, elbow, and wrist by calculating the Euclidean distance between their normalized 2D positions in consecutive frames. These features represent frame-to-frame displacement rather than physical velocity.

The resulting feature vector contains 15 input features, as summarized in Table I. Landmark visibility scores are used to

assess landmark reliability but are not included as classifier inputs.

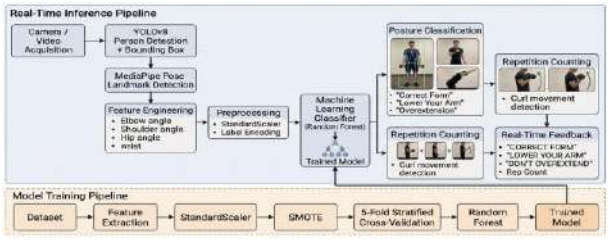
TABLE I
. COMPOSITION OF THE 15-DIMENSIONAL FEATURE VECTOR

| No. | Feature | Count |
|---|---|---|
| 1–3 | Shoulder coordinates ($s_x$,$s_y$,$s_z$) | 3 |
| 4–6 | Elbow coordinates ($e_x$,$e_y$,$e_z$) | 3 |
| 7–9 | Wrist coordinates ($w_x$,$w_y$,$w_z$) | 3 |
| 10 | Elbow angle ($\theta_{elbow}$) | 1 |
| 11 | Shoulder alignment angle ($\theta_{sh}$) | 1 |
| 12 | Relative vertical difference ($\Delta y$) | 1 |
| 13–15 | Shoulder, elbow, and wrist displacement | 3 |
| | Total | 15 |

Fig. 1. Overall architecture of the proposed workout posture assessment and repetition validation system1

### D. Dataset Collection and Preparation

A custom dataset of bicep-curl exercises was created using webcam recordings performed by ten (10) participants and contained 6666 labeled frames. Correct Form accounted for




2,282 samples (34.23%), Lower Your Arm comprised 2,197 samples (32.96%), and Overextension consisted of 2,187 samples (32.81%). Recordings were captured at a resolution of 640 × 480 pixels in different conditions, including home environments with natural daylight and at night, as well as gym environments under normal indoor lighting and low-light conditions. Data were collected from three camera viewpoints: front, side, and 45-degree, to capture variations in exercise from different positions.

Before each recording session, participants selected one of three posture categories as Correct Form (Class 00), Lower Your Arm (Class 01), or Overextension (Class 02) and the camera viewpoint. During exercise performance, the system automatically extracted body landmark coordinates and movement-related features, which were stored in a CSV file along with the selected posture category and camera viewpoint. Only the extracted landmark coordinates and engineered features were kept for model development and evaluation. Labels were assigned based on the categories selected by participants before recording. Each selected category was applied to the corresponding feature records collected during that session. Since the labels were not independently verified against individual frames, some records may not accurately represent the posture performed, introducing a potential source of label inconsistency that is acknowledged as a limitation of the dataset.

Class labels were encoded numerically as $y \in \{0,1,2\}$. To address the class distribution imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE generates synthetic samples using nearest-neighbor relationships within the feature space. Oversampling was restricted to training folds to prevent information from validation or test data from influencing model training.

### E. Model Development and Validation

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

Four supervised classifiers were evaluated: K-Nearest Neighbors (KNN), Logistic Regression, Multi-Layer Perceptron (MLP), and Random Forest (RF). Feature standardization [14] was performed using: where $x$ is the original feature value, and $\mu$ and $\sigma$ are the mean and standard deviation calculated from the training data.

The models used the following configurations: KNN with $k$ = 5 and Euclidean distance; Logistic Regression with L2 regularization, and a maximum of 2,000 iterations; MLP with hidden layers of (64,32), ReLU activation; and RF with 300 decision trees. All models used a random state of 42 where applicable.

Using an *imblearn.pipeline.Pipeline,* standardization and SMOTE were incorporated with each classifier to ensure that scaling and oversampling were fitted only on the training data. SMOTE's nearest-neighbor parameter was adjusted according to the smallest class count in the training partition.

The dataset was split into 80% for training and 20% for testing. Five-fold cross-validation with Stratified Group KFold was used to perform model selection when participant grouping was available and all classes were represented in every fold. If these conditions were not satisfied, Stratified K-Fold was used instead.

$$Accuracy_{avg} = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i, \quad k = 5 \quad (3)$$

The mean cross-validation accuracy was calculated as:

Model performance was compared using five-fold cross validation based on accuracy, precision, recall, and F1-score. The classifier achieving the highest F1-score was selected as the final model and evaluated on the held-out test set using the same metrics and a confusion matrix. The selected classifier, scaler, label encoder, and feature schema were saved for live inference.

### F. Real-Time Posture Feedback and Repetition Validation

During live execution, webcam frames are processed through YOLOv8 person detection and isolation, MediaPipe Pose estimation, feature extraction, and posture classification. The largest detected person region is used for pose estimation. The predicted posture is smoothed using a seven-frame majority buffer and displayed with color-coded feedback: green for Correct Form, yellow for Lower Your Arm, and red for Overextension.

Repetition tracking uses a deterministic Finite State Machine (FSM) with two states: EXTENDED and FLEXED. The elbow angle is smoothed using a five-frame median filter, and transitions require the corresponding angle condition to be satisfied for three consecutive frames. The transition to FLEXED occurs when the smoothed angle falls to or below 75°, while the return to EXTENDED occurs when it reaches or exceeds 145°.

A completed movement cycle is registered when the FSM returns from FLEXED to EXTENDED. The system maintains separate counters for completed cycles and posture-valid repetitions.

$$= \frac{N_c}{N_f}$$

Posture validity is evaluated using the smoothed classifier predictions collected after entering the FLEXED state and until the return transition is confirmed. The Correct Form proportion is calculated as:

$$P_{\text{correct}} \quad (4)$$

where $N_c$ represents the number of Correct Form predictions and $N_f$ represents the total number of posture predictions collected during this interval. A completed cycle is accepted as a valid repetition when $P_{correct} \geq 0.50$. Otherwise, it is recorded as an invalid repetition. This approach distinguishes completed movement cycles from repetitions satisfying the predefined posture-validity criterion.

Algorithm 1 summarizes the repetition-tracking and posture validation procedure.

Algorithm 1 Posture-Aware Bicep-Curl Repetition Tracking

**Input**: Elbow angle $\theta_t$, posture prediction $\hat{y}_t$
**Output**: Total completed cycles and valid repetitions
Initialize five-frame median filter
*state* ← *EXTENDED*
*total_cycles* ← 0
*valid_reps* ← 0
Initialize flexed-phase posture prediction buffer
Initialize consecutive-frame confirmation counters
**while** the webcam stream is active **do**
    Obtain the current elbow angle and posture prediction
    **if** required landmarks are valid **then**
        Apply median filtering to the elbow angle
        Update the movement-state confirmation counters
        **if** *state* = *EXTENDED* and the flexion condition is satisfied for 3 consecutive frames **then**
            *state* ← *FLEXED*
            Clear the flexed-phase posture prediction buffer
        **else if** *state* = *FLEXED* **then**
            Store the current posture prediction in the flexed-phase buffer
            **if** the extension condition is satisfied for 3 consecutive frames **then**
                *total_cycles* ← *total_cycles* +1
                $N_c$ ← number of Correct Form predictions in the buffer
                $N_f$ ← total number of predictions in the buffer
                if $N_f > 0$ and $\frac{N_c}{N_f} \geq 0.50$ then
                    *valid_reps* ← *valid_reps* +1
                **end if**
                Clear the flexed-phase posture prediction buffer
                *state* ← *EXTENDED*
            **end if**
        **end if**
    **end if**
**end while**

### G. System Implementation and Evaluation

The proposed system was implemented in Python using OpenCV for webcam acquisition and image processing, YOLOv8 for person detection and isolation, MediaPipe Pose for landmark extraction, Scikit-learn for model development and classification, and NumPy for numerical operations. Libraries such as Pandas, and Imbalanced-learn were used in the final implementation.

The evaluation covered three aspects:

*1)* Posture Classification: The classifiers were evaluated using the validation strategy described in the model development subsection. To assess performance, Accuracy, precision, recall, F1-score, and confusion matrices were used across the three posture categories.

*1)* Repetition Counting: Detected repetition counts were compared with manually established reference counts across 10 participants. Reference counts were obtained by manually counting the completed bicep-curl repetitions performed and entering the actual count into the evaluation system after the session. The detected counts were then compared with these reference counts to evaluate repetition-counting performance using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Exact Session Match (ESM):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - D_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - D_i)^2} \quad (6)$$

where $A_i$ and $D_i$ denote the reference and detected repetition counts for session $i$, respectively; $n$ is the total number of evaluated sessions; and $N_{\text{exact}}$ is the number of sessions with matching counts.

*1)* Runtime Performance: Runtime performance was assessed using processing frame rate (FPS) and frame-processing latency during live execution. Measurements covered YOLOv8 person detection, person-region cropping, MediaPipe Pose landmark extraction, feature extraction, machine-learning classification, posture feedback smoothing, repetition statemachine processing, and overlay drawing. Webcam frame acquisition, UI display through *cv2.imshow(),* and keyboard handling through *cv2.waitKey()* were excluded from the measured processing time.

## IV. Results and Discussion

### A. Summary of Findings

Random Forest achieved 84.02% mean accuracy across five-fold stratified cross-validation using normalized landmark coordinates, geometric features, and frame-to-frame displacement features. YOLOv8 was used to detect the person

and isolate the region of interest (ROI) before MediaPipe pose estimation. During live execution, repetition-counting performance achieved a Repetition Counting Accuracy of 80.65% and Mean Absolute Error (MAE) of 1.5 repetitions. The posture-validation criterion excluded completed movement cycles that did not meet the required form threshold. The system used an elbow-angle-based repetition-counting mechanism and a posture-validation criterion that accepted a repetition when at least 50% of the posture predictions during the flexion phase indicated Correct Form.

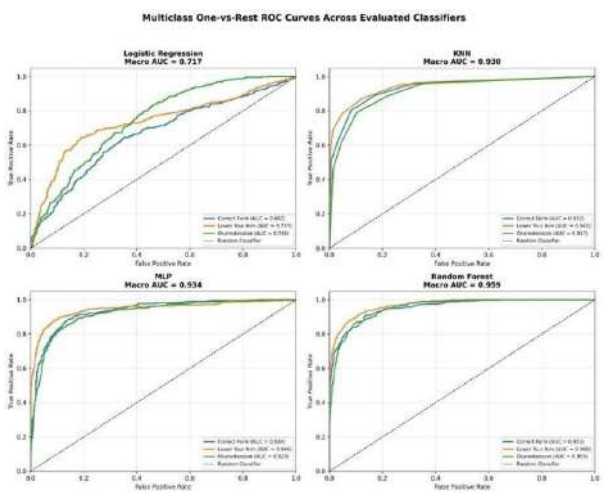
### *B. Classifier Performance*

TABLE II
REPORTS 5-FOLD CROSS-VALIDATION RESULTS ACROSS ALL FOUR MODELS.

TABLE II
CLASSIFIER 5-FOLD CROSS-VALIDATION ACCURACYII

| Algorithm | Mean Accuracy | Std. Dev. |
|---|---|---|
| Logistic Regression | 55.38% | ±1.31% |
| K-Nearest Neighbors ($k$ =5) | 80.04% | ±1.32% |
| Multi-Layer Perceptron | 81.13% | ±0.97% |
| Random Forest | 84.02% | ±0.83% |

Logistic Regression showed comparatively lower performance than the non-linear classifiers, indicating that more flexible models were better suited to capturing the relationships within the extracted posture features. Random Forest achieved the highest mean cross-validation macro F1-score and was therefore selected for the final system, subject to confirmation from the model comparison results.




Multiclass ROC analysis using a one-vs-rest approach was conducted to evaluate Random Forest and MLP on the held-out test set. AUC values were calculated for each posture category. The ROC curves provide an additional assessment of the classifiers' ability to distinguish between the posture categories using the 15-dimensional feature vector, as illustrated in Fig. 3.

Fig. 3. Multi-class ROC curves with Area Under the Curve (AUC) values across the evaluated classifiers.

### *C. Limitations and Mitigations*

The proposed system has several limitations. It currently supports only bicep curls and may be affected by environmental conditions such as lighting, background clutter, camera quality, and recording setup, which can influence person detection and landmark estimation. Repetition counting may also be affected by incomplete repetitions, unusual movement speeds, and deviations from expected movement patterns. Furthermore, the system has not been clinically validated for injury prevention or rehabilitation, and its feedback should not be considered medical advice.

### *D. Strengths of the Proposed Approach*

The proposed system provides several technical and practical advantages. First, it provides real-time posture assessment by delivering immediate feedback on bicep-curl form during the exercise. Second, it uses a standard webcam without requiring wearable sensors or specialized motion capture equipment. In addition to posture classification, the system incorporates automatic repetition counting, making it more suitable for real-time workout monitoring. A privacy preserving approach is followed by performing pose analysis locally without requiring continuous transmission of users' video data to an external server. The use of readily available camera hardware also supports accessibility. Finally, evaluating the machine learning models using $k$-fold stratified cross-validation provides a more robust assessment of model performance than a single train-test split.

### *E. Coherence with Existing Literature*

These findings align with prior markerless pose estimation research while extending it. Like the Stanford Pose Trainer [11] , this work confirms that simple geometric heuristics are sufficient for form classification without deep models. Building on the two-stage detection-plus-pose approach of Ko *et al* [4] , this study demonstrates the same principle for a different exercise category with a lighter detector, while adding an FSM-based repetition counter that gates rep validity on sustained correct form, which is a capability absent from prior MediaPipe-based trainers [12] , [13].

## V. CONCLUSION

This paper presented a markerless, privacy-preserving posture correction and repetition counting framework for unsupervised home workouts. By integrating YOLOv8 for person detection with MediaPipe for extracting landmarks, the system converts video into scale-invariant geometric features that are ideal for lightweight classification. Random Forest model attained a mean cross-validation accuracy of 84.02%, all operating locally without relying on wearables or cloud services.

The current pipeline has been evaluated only for bicep curls. Future developments will broaden the scope to include compound movements that involve multiple joints, such as squats and deadlifts, focusing on temporal modeling through LSTM networks to understand continuous full-body interactions. There will also be an investigation into more comprehensive movement tracking, which includes gait and kinematic analysis. Lastly, SHAP-based explainability will be integrated to improve clinical transparency and user trust. This will provide users with clear, feature-level biomechanical insights that explain the precise physiological factors driving their corrective posture feedback.

## REFERENCES


[1] D. A. Bonilla *et al.*, "Exercise Selection and Common Injuries in Fitness Centers: A Systematic Integrative Review and Practical Recommendations," Oct. 01, 2022, *MDPI*. doi: 10.3390/ijerph191912710.

[2] J. Fisher, J. Steele, M. Wolf, P. Androulakis Korakakis, D. Smith, and J. Giessing, "The Role of Supervision in Resistance Training; an Exploratory Systematic Review and Meta-Analysis," *International Journal of Strength and Conditioning*, vol. 2, no. 1, Apr. 2022, doi: 10.47206/ijsc.v2i1.101.

[3] A. Hannan, M. Z. Shafiq, F. Hussain, and I. M. Pires, "A portable smart fitness suite for real-time exercise monitoring and posture correction," *Sensors*, vol. 21, no. 19, Oct. 2021, doi: 10.3390/s21196692.

[4] Y. M. Ko, A. Nasridinov, and S. H. Park, "Real-Time AI Posture Correction for Powerlifting Exercises Using YOLOv5 and MediaPipe," *IEEE Access*, vol. 12, pp. 195830–195853, 2024, doi: 10.1109/ACCESS.2024.3516723.

[5] H. J. Chae, J.-B. Kim, G. Park, D. M. O'Sullivan, J. Seo, and J.-J. Park, "An Artificial Intelligence Exercise Coaching Mobile App: Development and Randomized Controlled Trial to Verify Its Effectiveness in Posture Correction," *Interact. J. Med. Res.*, vol. 12, p. e37604, Sep. 2023, doi: 10.2196/37604.

[6] A. and R. B. and I. J. Rajput, "Privacy-preserving human action recognition as a remote cloud service using RGB-D sensors and deep CNN," *Expert Syst. Appl.*, vol. 152, 2020.

[7] G. C. A. Q. J. Jocher, "Ultralytics YOLOv8," GitHub / Ultralytics.

[8] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured Feature Learning for Pose Estimation." [Online]. Available: http://www.ee.cuhk.edu.hk/

[9] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.08172

[10] E. Oliosi *et al.*, "Evaluation of Smartphone Camera Positioning on Artificial Intelligence Pose Estimation Accuracy for Exercise Detection: Observational Study," *JMIR Mhealth Uhealth*, vol. 14, 2026, doi: 10.2196/82412.

[11] S. Chen and R. R. Yang, "Pose Trainer: Correcting Exercise Posture using Pose Estimation."

[12] R. R. Kanase, A. N. Kumavat, R. D. Sinalkar, and S. Somani, "Pose Estimation and Correcting Exercise Posture," *ITM Web of Conferences*, vol. 40, p. 03031, 2021, doi: 10.1051/itmconf/20214003031.

[13] S. Hari Ram, A. Ashok, A. Ramalingam, and S. Dheenadhayalan, "Personal Exercise Trainer Using Computer Vision," in *2nd International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ic-ETITE58242.2024.10493615.

[14] M. M. Suarez-Alvarez, D. T. Pham, M. Y. Prostov, and Y. I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 468, no. 2145, pp. 2630–2651, Sep. 2012, doi: 10.1098/rspa.2011.0704.

# Perception Should Not Interpret: An Empirical Case for Separating Visual Description from Emotion Classification in Multimodal Conversational AI

W.K. Lakshan Dulhara
*Dept. of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
kldkaluthotage@students.nsbm.ac.lk

Hiruni Weerasinghe
*Dept. of Computer & Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
hiruni.w@nsbm.ac.lk

***Abstract*—Camera-aware conversational agents must decide which component turns a face into an emotional judgement. We test the design principle that a small vision-language model (VLM) should describe the face and a large language model (LLM) should interpret it, rather than the VLM interpreting directly. We built 84 face-text scenarios in four classes from 68 photographs labelled by three independent annotators (Fleiss' κ = 0.885). The three negative classes share one pool of messages, so the text alone cannot identify the class. In Study 2, two small VLMs (Moondream 2 and Granite Vision 3.2) either chose the class directly or wrote a description of the face, without seeing the message, that an LLM (DeepSeek-V3 or Claude Sonnet 4.6) then classified. Describe-then-interpret raised macro-F1 from 0.474 to 0.907 for Moondream and from 0.555 to 0.856 for Granite (exact McNemar $p < 0.001$ for three of four VLM-interpreter pairs, $p = 0.012$ for the fourth; three runs, majority vote). Asked to interpret, the VLMs rarely reported a mismatch: smiling faces with negative messages were mostly labelled positive and neutral faces with negative messages mostly labelled sad, yet descriptions of the same faces let the LLM recover these cases. Replacing the VLM description with a human description added only 0.045 to 0.057 macro-F1, so most of the loss lies in interpretation, not perception. In Study 1, two blind human raters and two LLM judges compared 168 replies generated from either a VLM description or a facial expression recognition label. The replies did not differ in specificity, emotional appropriateness or naturalness, although description-based replies referred to the face almost twice as often, and accurately. Small VLMs are reliable describers but unreliable interpreters of face-text incongruence.**

***Keywords—vision-language models, facial expression recognition, multimodal conversational AI, affective computing, cross-modal incongruence***

## I. Introduction

Conversational agents increasingly see their users. Webcam-enabled companions and tutoring systems can observe a face while a person types, and that observation is most useful when it disagrees with the words: a user who writes "everything is falling apart" while smiling needs a different reply from one whose face matches the message. The engineering question is which component should turn the image into that judgement.

Two designs are common. In the first, a vision model produces a compact signal, such as a facial expression recognition (FER) label or a set of action units, and this signal is passed to the dialogue model [1], [2]. In the second, a vision model writes a natural-language description that a large language model (LLM) reasons over [3]–[5]. Small vision-language models (VLMs) make a third option tempting: give the VLM the image and the message and let it decide directly. This keeps the pipeline short, but evidence that VLMs lean on language and on easily named cues rather than on visual detail [9]–[11], and that they recognise emotions poorly [6], [7], suggests that the shortcut may fail exactly on the incongruent cases that justify having a camera.

This paper tests the principle that perception and interpretation should be separated: a small VLM should describe, and a large LLM should interpret. We make three contributions. First, a controlled stimulus set of 84 face-text scenarios with human-validated face labels, in which the message alone cannot reveal the class. Second, a pre-specified classification study (Study 2) that compares direct VLM interpretation with describe-then-interpret across two VLMs and two LLM interpreters, with a human-description condition that separates perception errors from interpretation errors. Third, a blind response-quality study (Study 1) that asks whether replies generated from a VLM description are rated better than replies generated from an FER label. The conversational system that motivated the work serves only as a test bed; the findings concern the architecture, not the system.

## II. Related Work

### A. Emotion Recognition in Vision-Language Models

General-purpose VLMs remain weak at affective perception. Bhattacharyya and Wang [6] found that current VLMs predict the emotions evoked by images poorly, and Agarwal et al. [7] report that even large VLMs struggle to recognise facial emotions and rarely outperform specialised vision-only classifiers, attributing this to long-tailed web training data and coarse visual tokenisation. Fang et al. [8] improve emotion perception at inference time by first deciding valence and then suppressing distracting image content. These studies evaluate recognition of a single emotion; we study whether a VLM can judge agreement between a face and a message.

### B. Shortcuts in Multimodal Models

Several analyses show that multimodal models do not weigh their inputs evenly. Frank et al. [9] found that vision-and-

language transformers use visual input far less than textual input. Rahmanzadehgervi et al. [10] showed that VLMs fail simple visual tests although their encoders hold the needed information, and Shahgir et al. [11] found that VLMs skip visual comparison and reason through language whenever an entity can be named. Multimodal sarcasm detection [12] and emotion recognition in conversation [13] likewise depend on fusing modalities whose cues can conflict. These findings predict that a small VLM asked about a face and a message together will settle on one salient cue instead of comparing the two.

### *C. Description as an Interface Between Models*

Using language as the interface between perception and reasoning has a long record. Menon and Vondrick [4] improved zero-shot classification by having an LLM reason over descriptive features, and Socratic Models [5] compose pretrained vision and language models through text prompts. For affect, Mehra et al. [3] found that LLMs given numeric valence-arousal values classify facial expressions poorly but describe them in ways that match human interpretations. FER output has been injected into LLM prompts for empathic conversation [1] and tutoring [2], and VLM pipelines are being explored for social robots [14]. None of these compares direct VLM interpretation with describe-then-interpret on controlled face-text incongruence, which is the gap addressed here.

## III. METHOD

### *A. Stimuli and Ground Truth*

Seventy-nine face photographs were labelled for this study. Three annotators, none of them the author, independently labelled each image as happy, sad, neutral or other, seeing only an opaque identifier. The label was the majority category; agreement was high (Fleiss' $\kappa = 0.885$), and the pool held 32 happy, 26 sad and 21 neutral faces with no image excluded. The production FER model agreed with this consensus on 81% of images. Human consensus, not FER, defines the ground truth.

Class is defined by the face (Table I). Twenty-one positive messages and a shared pool of 63 negative messages (21 mild, 21 moderate and 21 severe) were written so that no message describes the face, signals sarcasm or mentions hiding feelings. The negative pool was split across the three negative classes with severity balanced, so a negative message is equally likely to belong to any of them and the face is required to tell them apart. Two human checkers rated valence (for three messages revised afterwards, two LLM judges confirmed it), and the two LLM judges screened every message for facial description, irony and concealment; no message was flagged by both judges. Faces were assigned at random with a fixed seed, preferring unanimously labelled images, giving 84 scenarios built from 68 unique images.

TABLE I. SCENARIO CLASSES (21 SCENARIOS EACH); THE THREE NEGATIVE CLASSES SHARE ONE MESSAGE POOL

| Class | Face | Message | Example message |
|---|---|---|---|
| Congruent positive | Happy | Positive | "We got engaged last night!" |
| Congruent negative | Sad | Negative | "I missed the deadline again." |
| Incongruent smile | Happy | Negative | (same pool of 63) |
| Incongruent neutral | Neutral | Negative | (same pool of 63) |

### *B. Models*

The small VLMs were Moondream 2 [15] and Granite Vision 3.2 2B [16], run locally through Ollama at temperature 0 on an 8 GB laptop GPU, with images resized to at most 640 px. The LLM interpreters were DeepSeek-V3 (deepseek-chat, temperature 0) and Claude Sonnet 4.6 (claude-sonnet-4-6). Claude was accessed through the Claude Code command-line interface in an isolated session with a minimal system prompt and no tools or project context; this interface does not expose a temperature setting, so its default sampling was used and the resolved model was logged for every call. FER used face-api.js 0.22.2 [17] with the same tiny face detector and expression network as the deployed system.

### *C. Study 2: Alignment Classification*

Fig. 1 shows the five conditions. Every condition answered the same four-option question for all 84 scenarios; only the evidence line differed (Table II). In C1 the VLM saw the image and the message and chose the class itself (C1-m for Moondream, C1-g for Granite). In C2 the VLM first described the face with a fixed prompt that did not include the message, and an LLM chose the class from the description and the message (C2-m, C2-g). C3 gave the LLM the FER label and probability distribution, C4 gave it a neutral human-written description of the face, written blind to class and pairing, and C5 gave it the message only. C4 estimates the ceiling with accurate perception, so C4 minus C2 isolates the cost of VLM perception, while C2 minus C1 isolates the cost of letting the VLM interpret.

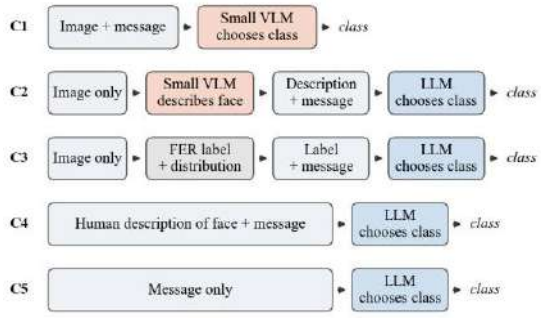



Fig. 1. Study 2 conditions. C1 and C2 are run with each small VLM; C2 to C5 are run with each LLM interpreter.

TABLE II. PROMPT COMPONENTS, FIXED BEFORE DATA COLLECTION

| Part | Text |
|---|---|
| Describe (C2) | Describe the person's face and facial expression in detail. |
| Question (all) | A person is chatting with an AI companion. {evidence} They sent this message: "{text}" Which option best describes the person's face together with their message? {options A to D} Answer with one letter only. |
| Evidence C1 | Look at the person's face in the photo. |
| Evidence C2, C4 | Here is a description of the person's face from their camera: "{description}" |
| Evidence C3 | Here is the output of a facial expression recognition sensor on their camera: {label, distribution} |
| Evidence C5 | No camera information is available. |
| Options | Happy or smiling face, and a positive message / Sad face, and a negative message / Happy or smiling face, but a negative message / Neutral face showing no clear emotion, but a negative message |

Because a pilot on five unused images showed a strong preference for option A, the option order was a seeded random permutation for each run and scenario, identical across conditions. Each condition ran three times (seeds 42 to 44) and the majority label per scenario was analysed. The primary outcome was macro-F1 over the four classes. Paired comparisons used the exact McNemar test on per-scenario correctness [18] and a bootstrap 95% confidence interval (10,000 resamples) of the macro-F1 difference. The hypotheses, prompts, exclusion rules and analysis scripts were fixed and version-controlled before any output was inspected, and every later change was logged with its reason. No scenario was excluded and no trial was missing.

### *D. Study 1: Response Quality*

For each scenario, Claude Sonnet 4.6 generated one reply per condition from the same short companion persona prompt and the same template, which differed only in one camera line: the Moondream description (D) or the FER label and distribution (F). Two independent raters, neither of them the author, rated all 168 replies in a browser tool that showed the photograph, the message and the two replies in a randomised A/B order, with item order shuffled per rater. Each reply was scored from 1 to 5 for emotional appropriateness (EA), specificity to the person and moment (SP, primary), natural language (NL) and visual accuracy (VA, or N/A when the reply made no visual claim). After rating, each rater guessed which reply in ten pairs came from the description condition, to test blinding. As a second check, Claude and DeepSeek judged all pairs with the same sheet in both A/B orders to cancel position bias [19]; since they cannot see images, the human-written description stood in for the photograph. The primary test was a one-sided Wilcoxon signed-rank test of SP(D) > SP(F) on per-scenario means across raters; EA, NL and the overall mean were tested two-sided with Holm correction. Agreement used Krippendorff's ordinal α [20].

## IV. RESULTS

### *A. Study 2: Describe-Then-Interpret Outperforms Direct Interpretation*

Table III gives macro-F1 for every condition. Letting the LLM interpret a VLM description outperformed letting the VLM interpret, for both VLMs and both interpreters. For Moondream, macro-F1 rose from 0.474 to 0.907 with DeepSeek and to 0.906 with Claude; 32 and 31 scenarios respectively were correct only under C2-m, against one and zero correct only under C1-m. For Granite, macro-F1 rose from 0.555 to 0.806 (DeepSeek) and 0.856 (Claude). Results were stable across runs: per-run macro-F1 ranged from 0.880 to 0.919 for C2-m and from 0.457 to 0.552 for C1-m.

TABLE III. STUDY 2 MACRO-F1 / ACCURACY (MAJORITY VOTE OVER THREE RUNS, N = 84)

| Condition | DeepSeek | Claude |
|---|---|---|
| C1-m: Moondream interprets | 0.474 / 0.536 (no LLM) | |
| C1-g: Granite interprets | 0.555 / 0.607 (no LLM) | |
| C2-m: Moondream describes | 0.907 / 0.905 | 0.906 / 0.905 |
| C2-g: Granite describes | 0.806 / 0.810 | 0.856 / 0.857 |
| C3: FER label | 0.879 / 0.881 | 0.880 / 0.881 |
| C4: Human description | 0.964 / 0.964 | 0.951 / 0.952 |
| C5: Message only | 0.412 / 0.452 | 0.370 / 0.488 |

Table IV reports the planned comparisons. The perception cost, C4 minus C2-m, was small (+0.057 and +0.045) and its confidence interval included zero for both interpreters. Vision contributed substantially over text alone (C2-m minus C5, about +0.50), as the design intends: with the message only, the LLM reached macro-F1 of 0.37 to 0.41. Moondream descriptions and FER labels did not differ (C2-m minus C3, +0.03, p = 0.77).

TABLE IV. STUDY 2 PLANNED COMPARISONS: MACRO-F1 DIFFERENCE [95% BOOTSTRAP CI], EXACT MCNEMAR P

| Comparison | DeepSeek | Claude |
|---|---|---|
| C2-m vs C1-m | +0.433 [0.331, 0.540] p < 0.001 | +0.433 [0.334, 0.536] p < 0.001 |
| C2-g vs C1-g | +0.251 [0.116, 0.389] p = 0.012 | +0.301 [0.172, 0.433] p < 0.001 |
| C4 vs C2-m | +0.057 [−0.000, 0.121] p = 0.125 | +0.045 [−0.018, 0.114] p = 0.289 |
| C2-m vs C5 | +0.495 [0.405, 0.581] p < 0.001 | +0.536 [0.464, 0.601] p < 0.001 |
| C2-m vs C3 | +0.028 [−0.052, 0.109] p = 0.774 | +0.027 [−0.053, 0.106] p = 0.774 |

Fig. 2 shows where the direct VLMs failed. Both handled congruent pairs well but almost never reported a mismatch. For a smiling face with a negative message, recall was 0.29 for both VLMs; Moondream labelled 13 and Granite 10 of these 21 scenarios as congruent positive, following the smile and

ignoring the message. For a neutral face with a negative message, recall was 0.10 for Moondream and 0.24 for Granite, and 16 and 13 scenarios were labelled congruent negative, following the message. When the same VLMs only described the face, the LLM recovered both incongruent classes, with recall of 0.86 to 0.91. The exception was Granite's description of sad faces: recall for the congruent negative class fell to 0.55 under C2-g, against 0.88 under C2-m, because several sad faces were described without any sign of sadness.

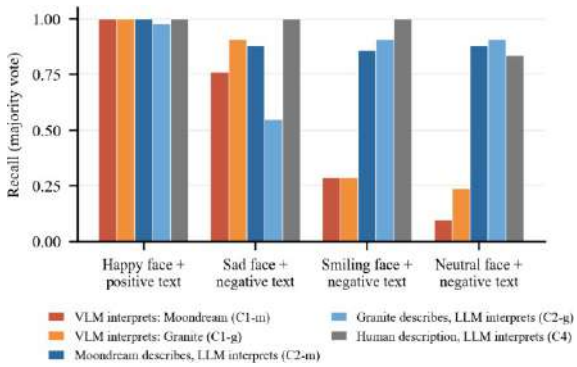



Fig. 2. Per-class recall in Study 2 (LLM conditions averaged over the two interpreters). Asked to interpret, both VLMs miss incongruent pairs; their descriptions let the LLM recover them.

Scenario s51 illustrates the difference. A face labelled neutral by all three annotators was paired with "Everything fell apart this week." Moondream, asked to interpret, chose the congruent negative class in two of three runs. Asked only to describe, it wrote that the woman "has a serious look on her face as she gazes directly into the camera", and from this description DeepSeek chose the correct incongruent neutral class in all three runs and Claude in two of three.

### *B. Study 1: Response Quality*

Table V summarises the human ratings. Replies built from the VLM description were not rated more specific than replies built from the FER label (4.29 vs 4.22, one-sided $p = 0.19$, rank-biserial $r = 0.13$), and EA, NL and the overall mean did not differ (all Holm-adjusted $p = 0.61$). Both LLM judges agreed: no criterion differed for either judge (all $p > 0.25$). The two raters agreed substantially (ordinal $\alpha$: EA 0.70, SP 0.73, NL 0.69, VA 0.66), and their scores correlated with the judges' most strongly on SP (Spearman $\rho = 0.46$ and 0.45). Blinding held: the raters identified the description reply in 4 and 5 of 10 pairs, which is chance. Description-based replies referred to the person's appearance almost twice as often (34 of 84 against 18 of 84), and none of these references was rated inaccurate (mean VA 4.47; no reply rated 2 or below).

TABLE V. STUDY 1 RATINGS (1 TO 5), MEAN OF TWO HUMAN RATERS, N = 84 REPLY PAIRS

| Criterion | D | F | D − F | p |
|---|---|---|---|---|
| Specificity (primary) | 4.29 | 4.22 | +0.07 | 0.191 (1-sided) |
| Emotional appropriateness | 4.39 | 4.45 | −0.06 | 0.605 (Holm) |
| Natural language | 4.45 | 4.56 | −0.11 | 0.605 (Holm) |
| Overall (EA, SP, NL) | 4.38 | 4.41 | −0.03 | 0.605 (Holm) |
| Visual claims (n of 84) | 34 | 18 | +16 | descriptive |
| Visual accuracy (VA) | 4.47 | 4.58 | −0.11 | descriptive |

## V. DISCUSSION

### *A. Interpretation, Not Perception, Is the Bottleneck*

The two studies support the separation principle for classification and show no cost to reply quality. The small VLMs perceived faces well enough: their descriptions, given to an LLM, came within about 0.05 macro-F1 of a human description. They failed when asked to interpret, and their errors were systematic. Instead of comparing face and message, they settled on one cue: a smile, when present, outweighed a negative message, and otherwise the message decided. This is the shortcut behaviour reported for VLMs in [9]–[11], where a salient, easily named cue replaces cross-modal comparison. Splitting the task moves the comparison to a model that reasons well over text and turns a weak interpreter into a useful sensor.

The design of the description step explains why the LLM succeeded. Because the description is written without the message, it cannot be pulled towards the text, so the LLM receives two independent pieces of evidence and compares them. This agrees with Mehra et al. [3], who found that LLMs interpret described facial expressions better than numeric affect values, and with description-mediated pipelines in vision [4], [5]. The quality of the describer still matters: Granite under-reported sadness, and that perception gap, not interpretation, accounts for most of its lower score. A practical consequence is that a describer should be validated on the emotions it must report, independently of the interpreter.

### *B. Descriptions, FER Labels and Reply Quality*

A good FER label was as useful as a VLM description for classification (C3 and C2-m did not differ), and it produced replies of the same rated quality. The value of the description is therefore not accuracy but content: it gives the generator concrete, checkable details, which description-based replies used twice as often and without error. In this sample that extra grounding did not raise specificity or appropriateness ratings, possibly because both conditions scored above 4 on a 5-point scale and left little room for a difference. A describer can therefore replace an FER module without loss, and it remains necessary when a system must reason about cues that FER categories do not cover, such as gaze, fatigue or context.

### *C. Threats to Validity*

*Construct.* The four classes are defined by the face alone, and the class definitions were given to every model in the prompt; real users may show mixed or ambiguous expressions.

The stimuli are 84 posed photographs rather than live webcam frames, and the messages are short English texts written for the study. *Internal.* The C4 descriptions were written by the author, blind to class and pairing, and the LLM judges in Study 1 saw a text description instead of the photograph. Claude ran at its default temperature, although its per-run results varied little. *External.* Only two small VLMs and two LLM interpreters were tested, no large end-to-end VLM was included as a baseline, and Study 1 used two raters. Whether telling raters about the camera changes how visual references are judged was not tested and is left for future work, together with video input, larger VLMs and larger rater panels.

## VI. Conclusion

Asked to judge a face and a message together, two small VLMs rarely detected a mismatch and reached macro-F1 of only 0.47 and 0.56. The same models, asked only to describe the face for a large LLM, supported macro-F1 of 0.81 to 0.91, close to the 0.95 to 0.96 reached with a human description. Replies built from VLM descriptions were rated as highly as replies built from FER labels and contained more frequent, accurate references to the person's face. For camera-aware conversational agents the evidence supports a clear division of labour: the small vision model should describe what it sees, and the language model should decide what it means.

## References


[1] L. Stacchio, A. Ubaldi, A. Galdelli, M. Mauri, E. Frontoni, and A. Gaggioli, "Empathic prompting: Non-verbal context integration for multimodal LLM conversations," arXiv:2510.20743, 2025.

[2] L. Fleig, "Towards multimodal affective intelligence in educational AI: Facial expression recognition for large language models," Undergraduate honors thesis, Dept. Cognitive Science, Univ. California, San Diego, CA, USA, 2025.

[3] V. Mehra, G. Laban, and H. Gunes, "Beyond vision: How large language models interpret facial expressions from valence-arousal values," arXiv:2502.06875, 2025.

[4] S. Menon and C. Vondrick, "Visual classification via description from large language models," in Proc. Int. Conf. Learn. Represent. (ICLR), 2023.

[5] Zeng et al., "Socratic models: Composing zero-shot multimodal reasoning with language," in Proc. Int. Conf. Learn. Represent. (ICLR), 2023.

[6] S. Bhattacharyya and J. Z. Wang, "Evaluating vision-language models for emotion recognition," in Findings Assoc. Comput. Linguist.: NAACL 2025, 2025, pp. 1798–1820.

[7] M. Agarwal, S. A. Tsaftaris, L. Sevilla-Lara, and S. McDonagh, "Why do vision language models struggle to recognize human emotions?" arXiv:2604.15280, 2026.

[8] Y. Fang, J. Liang, W. Huang, H. Li, K. Su, and M. Ye, "Catch your emotion: Sharpening emotion perception in multimodal large language models," in Proc. 42nd Int. Conf. Mach. Learn. (ICML), PMLR, vol. 267, 2025, pp. 16029–16039.

[9] S. Frank, E. Bugliarello, and D. Elliott, "Vision-and-language or vision-for-language? On cross-modal influence in multimodal transformers," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2021, pp. 9847–9857.

[10] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen, "Vision language models are blind," in Proc. Asian Conf. Comput. Vis. (ACCV), 2024, pp. 18–34.

[11] H. S. Shahgir et al., "VLMs need words: Vision language models ignore visual detail in favor of semantic anchors," arXiv:2604.02486, 2026.

[12] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in Proc. 57th Annu. Meeting Assoc. Comput. Linguist. (ACL), 2019, pp. 2506–2515.

[13] T. Yun, H. Lim, J. Lee, and M. Song, "TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. (NAACL), 2024, pp. 82–95.

[14] R. Janssens and T. Belpaeme, "Towards multimodal social conversations with robots: Using vision-language models," arXiv:2507.19196, 2025.

[15] "Moondream: A tiny vision language model," GitHub repository. [Online]. Available: https://github.com/vikhyat/moondream

[16] Granite Vision Team, IBM Research, "Granite Vision: A lightweight, open-source multimodal model for enterprise intelligence," arXiv:2502.09927, 2025.

[17] V. Mühler, "face-api.js: JavaScript API for face detection and face recognition in the browser," GitHub repository. [Online]. Available: https://github.com/justadudewhohacks/face-api.js

[18] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput., vol. 10, no. 7, pp. 1895–1923, 1998.

[19] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Datasets and Benchmarks Track, 2023.

[20] F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," Commun. Methods Meas., vol. 1, no. 1, pp. 77–89, 2007.

# Child-Aware Expected-Harm Manoeuvre Selection for Autonomous Vehicles: A Pre-Registered Paired Simulation Study

T. R. D. T. Dulshan
*Department of Software Engineering*
*NSBM Green University*
*Homagama, Sri Lanka*
rdtdulshan@students.nsbm.ac.lk

Anton Jayakody
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
anton.j@nsbm.ac.lk

***Abstract*—Autonomous-vehicle safety systems typically treat all pedestrians identically, yet children move less predictably than adults and are more severely injured at equal impact speed. We present a controller in which the detected pedestrian class conditions the worst-case reachable envelope, the injury-severity model, and the timing of intervention, inside a single expected-harm manoeuvre selector. We evaluate its decision layer in CARLA with ground-truth perception, under a pre-registered, seed-paired design (16,650 runs across 35 scenarios), against a braking-only AEB archetype (B0) and an ablation of the same system with child-awareness disabled (B1). On the pre-registered primary outcome, impact-speed reduction, the full controller improves on B0 by 0.192 m/s ($p < 0.001$), an effect that also holds with scenario as the unit of analysis ($p = 0.012$), although a braking artifact in the simulator staging, disclosed here, overstates its size. Against B1 it improves by 0.103 m/s under the registered pooled test (Bonferroni-corrected $p = 0.048$), but this child-tier effect is carried by a few class-sensitive scenarios and is not significant at scenario level ($p = 0.38$), so we claim it for the tested scenario set only. Secondary and exploratory results are directionally consistent: collision rates of 12.0% (B0), 8.0% (B1) and 6.9% (full), or 9.1%, 4.9% and 3.7% excluding an unavoidable-by-design control; zero false evasions, and no measured degradation for adult pedestrians. We further report two calibrated negative results: the worst-case-gated acceleration escape never fires under realistic authority (zero unsafe accelerations in all runs), and the harm minimiser never selects a lane change below the kinematic brake/steer crossover, consistent with established threat-assessment results.**

***Keywords*—*vulnerable road user protection, child pedestrian safety, autonomous vehicles, expected harm, pre-registered evaluation***

## I. Introduction

Child pedestrians are over-represented in road fatalities worldwide [1], and they are not small adults. Kinematically, children cross faster relative to their gait, accelerate in bursts, and reverse direction mid-crossing at rates adults do not [2]. Biomechanically, the fatality-risk curves that underpin automated emergency braking (AEB) calibration are fitted to casualties aged 15 and above [3], [4]; no widely adopted child-specific curve exists. Production AEB and most academic planners nonetheless treat "pedestrian" as a single class: one trigger, one motion model, one severity assumption.

This paper asks what changes when the pedestrian's class conditions the entire decision pipeline, and whether that measurably improves child protection without degrading anyone else's. The individual ingredients we build on are established: set-based pedestrian prediction [5], [6], occlusion-aware verification [7], fail-safe trajectory commitment [8], and risk-weighted harm costing [9], [10]. Our contributions are:

1) **An integrated child-aware controller and its empirical quantification:** to our knowledge the first end-to-end system in which pedestrian class conditions worst-case reachable envelopes *and* a class-graded severity model inside one expected-harm selector (with fail-safe classification bias, coasting lost tracks, and a worst-case-gated acceleration escape), evaluated at scale against a matched ablation that removes child-awareness as a whole.

2) **A pre-registered, seed-paired evaluation methodology for simulation-based AV safety claims:** design, sampling distributions, outcomes, and tests locked and dated before execution [11]; every random draw keyed only on (scenario, seed) so all controllers face bit-identical worlds; deviations dated and disclosed; negative and attenuated results reported alongside confirmed ones.

Two further findings are reported as findings, not contributions: the gated acceleration escape refuses to fire under realistic authority (Section V-D), and brake-first behaviour emerges from harm minimisation, consistent with classical brake/steer threat assessment [12].

## II. Related Work

*Pedestrian motion prediction.* Data-driven trajectory prediction is comprehensively surveyed in [13], with simple kinematic baselines remaining stubbornly competitive [14]; we adopt the set-based alternative because the safety requirement here is containment, not average accuracy.

*Set-based prediction and formal safety.* Reachability-based prediction of traffic participants [5], extended to pedestrians with occlusions and traffic rules [6], provides the containment guarantee our envelopes rely on; occlusion-aware pedestrian collision avoidance [7] and fail-safe planning with verified fallback trajectories and commitment timing [8] establish the safety architecture our controller instantiates. Against this line, our delta is not a new prediction or commitment method: it is

conditioning the envelope's kinematic bounds and the downstream severity model on the pedestrian's class, and measuring what that buys end-to-end. Responsibility-Sensitive Safety [15] formalises blame-free envelopes; our worst-case gates are in the same spirit but checked numerically per candidate rather than derived as closed-form rules.

*Risk- and harm-aware planning.* Ethical risk-distribution planners [9], [10] allocate expected risk across road users with explicit weights. We adopt the expected-harm framing of [10], replace probabilistic prediction with class-conditioned worst-case envelopes, and add a class-graded severity treatment with declared, inspectable parameters.

*AEB and evasive manoeuvring.* Consumer protocols exercise threshold-triggered braking on child-dummy targets [16]; combined automatic braking and evasive steering for pedestrians, and the stopping-versus-swerving kinematics underlying it, are established [12]; recent reviews catalogue the avoidance-manoeuvre space [17], and an acceleration escape appears in patent form [18]. Absent from this literature is a worst-case admissibility condition for accelerating near a child: the gate evaluated here.

*Child pedestrians.* Child crossing speeds and mid-crossing behaviours are documented empirically [2]; fatality-risk curves are adult-fitted [3], [4]. We use the adult curve as the severity backbone and encode child fragility as a declared shift (Section III-B), citing the absence of a widely adopted child curve as motivation rather than claiming to close that gap.

## III. The Child-Aware Controller

The controller is a pure function from a world state (ego kinematics, tracked agents with class labels and confidences, road context, friction estimate) to a control command, at 20 Hz (Fig. 1). Simulator access is confined to one adapter behind this interface; the safety core is unit-tested without the simulator (66 tests at the evaluated version).

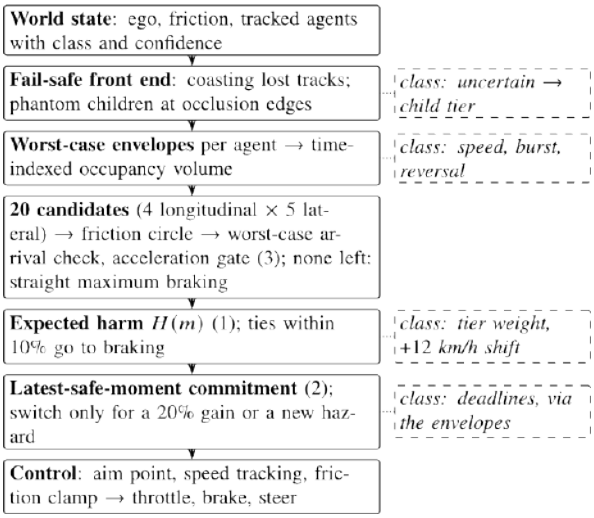



Fig. 1. One 20 Hz control cycle. Dashed tags mark where the pedestrian class enters the decision.

### A. Class-Conditioned Worst-Case Prediction

Each tracked agent receives a reachable envelope over a 2 s horizon: a time-indexed disc sequence containing the agent under worst-case kinematics for its class (we call each class's parameter set a *tier*). The child envelope is omnidirectional (children may reverse), growing at up to 3.5 m/s with 3.0 m/s$^2$ burst acceleration; the adult envelope is advected along current velocity without reversal, at 3.0 m/s and 2.0 m/s$^2$. The child speed bound is the top of the empirical crossing-speed range of primary-school children [2]; the accelerations and the adult bound are declared engineering values, not fitted. Unknown agents receive the most conservative (child-like) treatment. Envelopes merge into a time-indexed occupancy volume; escape candidates are checked against occupancy at their own arrival times (Fig. 2). Degraded perception is handled fail-safe: classification uncertainty is biased towards the child tier (boundary shift 0.15), because the two misclassification costs are asymmetric; a lost track is never deleted: it coasts with its envelope inflating 0.15 m per frame, so lost sight never reads as clear road; and occlusion edges spawn phantom child hypotheses [7] that drive anticipatory slowdown even when nothing is detected.

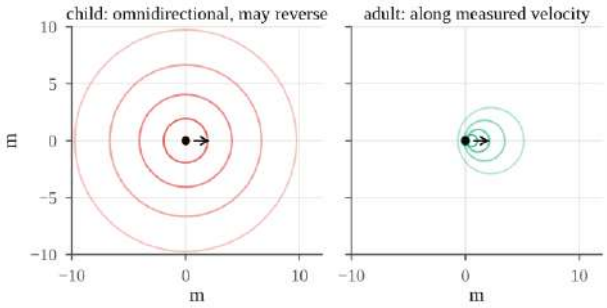



Fig. 2. Worst-case reachable envelopes for a child and an adult walking at 1.4 m/s, discs every 0.4 s to 1.6 s.

### B. Expected-Harm Selection

Each cycle the engine rolls out 20 candidates, four longitudinal profiles (maximum brake, comfort brake at 0.25 g, hold, capped accelerate) crossed with five lateral profiles (straight, in-lane offset left/right, lane change left/right), under a friction-circle constraint, discards the infeasible, and scores the survivors:

$$H(m) = \Sigma_a \, P(\text{collision} \mid m, a) \cdot S(\text{class}_a, v_{\text{res},a}) \qquad (1)$$

where the sum runs over tracked hazards $a$, $P$ grows with the candidate's deepest penetration of hazard $a$'s envelope scaled by track confidence (floored at 0.1; no contribution is discarded), and $v_{\text{res},a}$ is the ego speed at first penetration. Severity $S$ is a tier weight (child 1.6, adult 1.0, occupied vehicle 0.4, property 0.05) times the Rosén–Sander fatality logistic [3] evaluated at $v_{\text{res}}$, read +12 km/h higher for the child tier. The logistic reproduces the published adult curve (8.4% fatality risk at 50 km/h, 50% at 76 km/h, in line with [4]). No child curve exists, so the weight and the shift are declared design values: together they make a child contact four to five times as costly as an adult contact at 30–50 km/h. Child vulnerability is thus encoded twice, and Section VI reports a sensitivity sweep over both values. Because $S$ is convex in speed, braking is never

wasted; this property ultimately drives the brake-first result of Section V-D.

### C. Latest-Safe-Moment Commitment

Every candidate has an expiry time: the last moment it remains feasible under worst-case hazard motion. Hard commitment is deferred until

$$t_{\text{commit}} = t_{\text{expiry}} - (\tau_{\text{perc}} + \tau_{\text{replan}} + \tau_{\text{act}} + \tau_{\text{model}}) - \tau_{\mu} \quad (2)$$

with $\tau_{\text{perc}} = 0.08$ s, $\tau_{\text{replan}} = 0.05$ s, $\tau_{\text{act}} = 0.15$ s, $\tau_{\text{model}} = 0.10$ s, and $\tau_{\mu} = 0.30$ s added when friction < 0.6. While deferring, the controller emits comfort deceleration, which itself extends deadlines. Committed plans change only for a 20% score improvement or a new hazard; scores within 10% tie-break towards braking. A committed lane change that becomes infeasible aborts to straighten-and-brake. If no candidate passes its worst-case check, the controller degrades to straight-line maximum braking: impact-speed minimisation, never a gamble. This commitment architecture follows the fail-safe planning line of [8]; our addition is the class-conditioned inputs that set its deadlines.

### D. The Gated Acceleration Escape

Accelerating candidates (capped at +13 km/h) are admissible only if

$$t_{\text{exit}} + \text{buffer} + 0.3\text{ s} < t_{\text{arrival,worst}} \quad (3)$$

where $t_{\text{exit}}$ is the ego's worst-case corridor-exit time and $t_{\text{arrival,worst}}$ assumes the child sprints directly at the exit point regardless of current heading. The gate is barred outright for cyclists, at pedestrian crossings, and below 0.9 track confidence, and it covers every accelerating candidate including steer-and-accelerate combinations (a labelling loophole was found and closed before the evaluation ran; recorded as a dated pre-registration amendment).

## IV. Experimental Design

*Simulation.* CARLA 0.9.16 [19], Town10HD_Opt, ego blueprint *vehicle.audi.a2* (a compact car), synchronous mode at a fixed 0.05 s step: every episode is a deterministic function of its inputs, verified by replaying episodes and comparing full per-cycle traces. Scenarios place agents in the ego frame, so tested geometry is town-independent.

*Perception scope.* Perception is simulator ground truth (oracle), so that outcome differences are attributable to the decision layer. The oracle has no line-of-sight test: agents behind occluders are reported to the controller, so in the occlusion-emergence designs the hidden child is known before it emerges, and phantoms add to that knowledge rather than replace it. The oracle is therefore an upper bound on perception, and every result below is a decision-layer result under it.

*Scenarios and conditions.* 35 scenarios in eight families: crossings, darting/occlusion emergences, frozen and reversing pedestrians, groups, cyclists, dual-hazard and multi-lane trade-offs, perception-fault designs (night, glare, misclassification, track loss), and unavoidable-by-design controls. Thirty scenarios run three entry speeds (30/45/60 km/h) × two friction conditions (dry/wet); five condition-locked scenarios run only at their designed condition. Neither the wet condition nor the perception faults reached the simulator (Section VI). The controller was developed and debugged against pilot runs of this same scenario suite before the design was locked; the pre-registration therefore protects against outcome switching, distribution tuning, and analysis flexibility, not against controller familiarity with the scenario types, which we disclose and bound via the seed-sampled variations.

*Baselines.* B0 is a braking-only AEB archetype: full braking, straight, when any forward agent's time-to-collision falls below a uniform 1.6 s. It is parameterised at the generous end but is deliberately an archetype, not a production twin (no staged braking or trigger latching); we treat B0 as the "what exists" anchor and rest the child-awareness claim on B1. B1 runs identical machinery with the child tier's envelope and severity parameters set to their adult values, making it the most demanding available comparator: whatever separates it from the full controller is attributable to child-awareness alone. No published planner is re-implemented: a risk-distribution planner [10] or an RSS rule set [15] would change the planner and the child-awareness together, so B1, which holds the planner fixed, is the comparator that answers the child-awareness question.

*Sampling and pairing.* Seeds 0–29 per cell; seed 0 is the nominal design point, seeds 1–29 draw from a pre-registered registry (trigger distance ±25%; child crossing speed uniform on the empirical 1.0–3.5 m/s range [2], adults 0.8–1.6 m/s; mid-crossing reversal $p = 0.15$ and freeze $p = 0.10$ for children; entry-speed and friction jitter). Draws derive from (scenario, seed) only, never the controller: identical worlds by construction. Total 16,650 episodes, zero infrastructure failures. The design has a deliberate circularity: sampled child behaviour respects the kinematic bounds the child envelope encodes, so it cannot falsify those bounds, whose external validity rests on [2]. A post-evaluation audit found one deviation that breaks this in a single scenario: the harness did not normalise walker direction vectors, so 14 of the 35 scenarios scaled commanded walker speed by 0.5–1.6× (identically for all controllers, so pairing is intact). In s16, the most critical emergence design, the child ran at 1.6× the sampled speed, up to 5.3 m/s against the 3.5 m/s bound. We report the results as run; s16 is in effect an unplanned out-of-envelope test (Section V-D).

*Outcomes and analysis.* Primary confirmatory outcome: impact-speed reduction (ISR), defined per episode as approach speed minus impact speed (equal to approach speed for collision-free episodes), full vs. B0 and full vs. B1, Wilcoxon signed-rank (zero-split) on the 5,550 matched (scenario, condition, seed) pairs per comparison, Bonferroni-corrected across the two tests [20]. Because pairs cluster within scenarios, we additionally registered a cluster-aware sensitivity analysis: scenario-level paired aggregation tested across the 35 scenario units. The two answer different questions. The pooled test treats the 35 scenarios as fixed and generalises over seeded variation within them; the scenario-level test treats scenarios as a sample and generalises to new scenario types. We call an effect general only when both agree. Secondary outcomes (exact McNemar on the paired binary collision outcome [21]; minimum-TTC margin; false evasions; plan-switch stability) and all per-scenario results are exploratory, and each results subsection is

labelled accordingly. Five dated deviations are on record: the gate amendment and a harness checkpointing fix (both pre-execution), a post-hoc widening of the adult-parity scenario set (Section V-C), the walker-speed scaling above, and the unapplied conditions (Section VI).

## V. Results

### A. Primary Outcome and Its Unit of Analysis (Confirmatory)

On the pre-registered pooled analysis, the full controller improves ISR by +0.192 m/s [+0.166, +0.219], $p = 9.9\times10^{-30}$, against B0 and by +0.103 m/s [+0.077, +0.128] against B1 ($p = 0.024$, Bonferroni-corrected 0.048) (Table I). The B0 comparison is general in the sense above, although a staging artifact that favours early braking overstates its size (Section VI): it holds at scenario level ($p = 0.012$), with 14 scenarios improved, 3 worse and 18 unchanged on ISR, and 14 to 1 on collision rate ($p = 0.0016$). The B1 comparison is not ($p = 0.38$). The child tier alters ISR in only 10 of the 35 scenarios (6 better; 4 worse, by at most 0.09 m/s) and the collision rate in 7 (6 better, 1 worse); the other 25 produce the same outcome under both controllers. A post-hoc scenario-resampling bootstrap places the mean gain at [+0.017, +0.209] m/s; the interval excludes zero after removing any one scenario, but not after removing together the two close-range child crossings of Section V-B. We therefore state the child-tier claim narrowly: a reproducible effect on this scenario set, carried by a few class-sensitive designs, and not a demonstrated effect across scenario types. Its substantive support is the mechanism-level evidence of Section V-B.

TABLE I. Registered primary outcome (top) and secondary and exploratory metrics (bottom), 5,550 paired runs per controller. ISR = impact-speed reduction; plan switches = committed-plan changes per episode. B0 logs no TTC margin because its trigger is the TTC computation itself.

| Metric | Full | B1 | B0 |
|---|---|---|---|
| *Confirmatory: ISR gain of full over the baseline* | | | |
| Mean gain (m/s) | ref. | +0.103 | +0.192 |
| pooled $p$, 5,550 pairs (Bonf.) | | 0.048 | $2.0\times10^{-29}$ |
| scenario-level $p$, 35 units | | 0.38 | 0.012 |
| scenarios better/worse/same[a] | | 6/4/25 | 14/3/18 |
| *Secondary and exploratory* | | | |
| Collision rate (all contacts) | **6.9%** | 8.0% | 12.0% |
| excl. unavoidable control | **3.7%** | 4.9% | 9.1% |
| Mean impact speed (m/s) | **0.12** | 0.22 | 0.31 |
| Mean min-TTC margin (s) | **0.81** | 0.67 | n/a |
| False-evasion rate | **0.0%** | 0.0% | 3.2% |
| Plan switches per run | **1.08** | 1.48 | 7.36 |

Descriptive, post hoc; "same" means within 1 mm/s.

### B. Where Each Layer Wins (Exploratory)

The predicted ordering full < B1 < B0 holds on every safety metric in Table I, and paired collision outcomes disagree asymmetrically: among cells where the two controllers differ, B0 alone crashes in 304 and the full controller alone in 18 ($p = 3.3\times10^{-68}$); against B1 the split is 71 to 7 ($p = 1.9\times10^{-14}$).

*The adaptive layer (B1 vs. B0)* removes threshold-triggering failure modes: panic-braking into a crossing's arrival window (42.2% → 6.7% collisions on the pass-before-gap scenario), 100% false evasion on a rolling ball (→ 0%), and trigger chatter (7.36 → 1.48 plan switches). In the contested dual-hazard family, the full controller records 33 saves to 0 losses in discordant worlds against B0 ($p = 2.3\times10^{-10}$) and hits, when it must, at a quarter of the baseline's residual energy (0.22 vs. 0.93 m/s) (Fig. 4).

*The child tier (full vs. B1)* wins where child-like motion outruns an adult envelope at close range. In s26 and s28, children crossing 16–18 m ahead, it eliminates all collisions (0% vs. 12.2% and 0% vs. 13.3%). Both were designed as perception faults (glare; an adult misclassified as a child), but the oracle applies neither (Section VI): they run as correctly labelled child crossings, so they measure the envelope, not the classification fail-safe. On the crossing family the child envelope buys +0.285 s of minimum-TTC margin ($p = 6.0\times10^{-37}$) and cuts near-zero-margin runs from 29% to 11% (Fig. 3): the tier's everyday effect is removing near-misses, and its collision wins concentrate in the tail where that margin decides the outcome. All at zero false evasions.

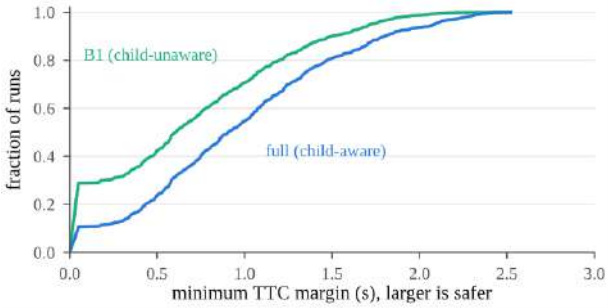



Fig. 3. Minimum-TTC-margin ECDF on the crossing family, full vs. B1: near-zero-margin runs drop from 29% to 11%.

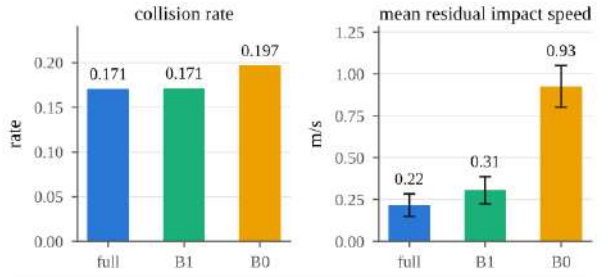



Fig. 4. Contested dual-hazard family (s09–s14, s34): collision rate (left) and mean residual impact speed with 95% CI (right).

### C. No Measured Cost to Adults (Exploratory)

On the registered adult-scenario subset (900 pairs), full and B1 produce identical outcomes in every pair. Under oracle classification this parity is expected by construction (adult-only scenarios exercise identical code paths), and we report it as verification that child-awareness is architecturally isolated, not as an empirical safety finding. The informative evidence is the exploratory widening to all 14 adult-target scenarios (2,220

pairs; a documented post-hoc deviation): 24 worlds where only B1 collides, zero conversely ($p = 1.2\times10^{-7}$): the shared machinery and the fail-safe bias actively help adults too.

*D. Calibrated Negative Results (Exploratory)*

*The darting-child hypothesis is favourable but unproven.* On the three near-unavoidable darting scenarios, collisions run 16.3% (full) vs. 17.6% (B1), $p = 0.17$, with 13 saves to 6 losses in 540 pairs. All of that discordance comes from s16, where the child outran the envelope bound (Section IV). Facing a child its envelope did not contain, the full controller still collided no more often than either baseline (27.8% vs. 31.7% B1 and 31.1% B0), at a third of B1's mean impact speed (0.18 vs. 0.57 m/s); in every arm most s16 contacts occur at near-zero speed (median below 0.04 m/s). This is one scenario, read post hoc: consistent, not proven.

*The acceleration escape never fires, by design logic rather than by accident.* Across all 16,650 runs the gate admitted zero unsafe accelerations and the manoeuvre never executed. A geometry sweep around the pass-before concept found no condition in which commanded acceleration under the +13 km/h cap is both worst-case-admissible per (3) and necessary: the cap buys 0.1–0.2 s of exit margin, and wherever passing satisfies the worst-case check, holding speed already achieves it with zero harm. We state this as an empirically verified refusal property over the tested speed range (30–60 km/h), not a theorem. The pass-before behaviour the manoeuvre targets is realised as a checked speed-hold (6.7% vs. B0's 42.2% collisions on that scenario).

*Lane changes never win at urban speeds, consistent with known kinematics.* A census over all 5,550 full-controller traces shows the lane-change manoeuvre never executed; the only lateral action used is a small in-lane offset while braking (dominantly threading between two children in adjacent lanes). Below the brake/steer crossover, stopping requires less distance than lane displacement [12], compounded here by worst-case costing of moving hazards and adjacent-lane occupancy. We read this as validation that the harm model is calibrated (it recovers established threat-assessment doctrine without hard-coding it), while the lateral repertoire remains available for regimes that warrant it.

## VI. Discussion and Limitations

*Perception.* Claims are bounded to the decision layer under clairvoyant ground truth (Section IV), in which the classification bias and track coasting are built but never exercised: every label arrives at full confidence and no track is lost. A real detector adds three error types the oracle hides. Missed and late detections shorten the time the envelope has to act. A confident child-as-adult error falls outside the fail-safe bias, which acts only on uncertain labels. Extra tracks cost progress rather than safety. A detector-in-the-loop evaluation is under way and will be reported separately.

*Unapplied conditions.* A post-evaluation audit found that two condition families never reached the simulator. "Wet" changed only the friction the controller is told ($\mu = 0.6$), on a road that still grips as dry, and "night" applied no lighting change. The oracle likewise injects none of the faults that four designs are named for (night, glare, misclassification, track loss), so they run as clean geometries. B0, which ignores $\mu$, is identical in all 2,700 paired dry/wet cells. A post-hoc spot-check on a road with the grip the controller assumes (14 wet cells per controller) raised collisions for every controller (full 1 → 5, B1 1 → 4, B0 1 → 9) with the ordering against B0 intact. The registered wet cells therefore understate collision risk, and no result here tests a physically wet road or degraded perception.

*Out-of-envelope behaviour.* Apart from the unplanned s16 case, sampled children respect the envelope bounds, so the evaluation quantifies those assumptions rather than testing them. A designed stress test with children beyond them (sprints above 3.5 m/s, accelerations above the burst bound, instantaneous reversals) is the right adversarial follow-up; no containment guarantee holds outside the declared kinematics.

*Mechanism attribution.* B1 sets envelope and severity to adult values together, so the full-vs-B1 contrast does not separate them. A registered sweep over the severity values (weight {1.3, 1.6, 2.0} × shift {+6, +12, +18} km/h, on a 13-scenario subset at bounding speeds, five seeds, 2,300 runs) found closed-loop outcomes essentially invariant: every grid point differs from the centre by at most one collision outcome in 230 paired cells, and mean ISR is unchanged to four significant figures. Under worst-case prediction the severity values are thus a near-inert closed-loop lever, which suggests that the child-tier benefit comes from envelope conditioning rather than from harm weighting; the classification fail-safe was never exercised (above). That is an inference from the sweep, not a measured ablation: envelope-only and severity-only arms, which the paired infrastructure supports, are needed to confirm it.

*Staging artifact.* Found after the evaluation and not corrected in it: every registered run sets the ego's entry speed over wheels at rest, so a brake applied before any throttle decelerates at a non-physical ≈3.9 g, while braking after throttle, as B0's fixed cruise throttle forces, reaches about 0.55 g; driven up to speed on its own drivetrain the car brakes at 8.5–9.3 m/s$^2$. This favours whichever controller brakes first, usually the full controller. A post-hoc re-run, with the ego driven up to speed, of the dry cells in which only B0 collided (87 of 148 could be staged; 60 km/h could not) kept that outcome in 53, removed B0's collision in 28, and added a full-controller collision in 6, four of them on the pass-before-gap scenario. The direction of the B0 comparison survives, but its size is overstated; the full-vs-B1 contrast was not re-screened.

*Generality and metric.* One town, one vehicle, and scripted (seed-randomised) pedestrian behaviour bound generality; because scenarios are ego-relative geometries, porting is cheap. The binary collision metric counts any ego contact, including being rear-ended while correctly stopped in the tailgater control (100% for every controller; Table I reports rates with and without it), which is why the continuous measure was registered as primary.

*Compute and availability.* The evaluation logs no compute time. Measured offline on a desktop CPU (Intel i5-14400F, single-threaded Python, up to nine agents and three phantoms), one control cycle takes a median of 1.3–2.6 ms and under 10 ms at worst, against the 50 ms cycle; automotive-grade hardware was not measured. The controller, scenario suite, harness, pre-

registration and unit tests are available from the authors on request.

## VII. Conclusion

Conditioning an expected-harm manoeuvre selector on pedestrian class (declared envelope and severity parameters inside class-identical machinery) improves protection over a braking-only AEB archetype at every level of analysis we ran, by a margin that a disclosed staging artifact overstates. The child-specific layer adds a pooled primary effect that is significant on the tested scenario set, with consistent, one-sided mechanism-level wins (close-range child crossings, crossing margins, near-miss removal); it is carried by a few class-sensitive scenarios and does not generalise at scenario level, which we report plainly. Adults measurably lose nothing. The negatives are informative: worst-case checking refuses the acceleration gamble a child's kinematics make unprovable in the tested regime, and harm minimisation recovers brake-first doctrine without hard-coding it. Future work: envelope-only and severity-only ablation arms, a designed out-of-envelope stress test, physically wet roads and injected perception faults, a registered detector-in-the-loop evaluation of the full comparison, comparison against published planners, and multi-town replication.

## References


[1] World Health Organization, "Global status report on road safety 2023," World Health Organization, Geneva, Tech. Rep., 2023.

[2] M. S. Riaz, A. Cuenen, E. Polders, M. B. Akram, M. Houda, D. Janssens, and M. Azab, "Child pedestrian safety: study of street-crossing behaviour of primary school children with adult supervision," *Sustainability*, vol. 14, no. 3, p. 1503, 2022.

[3] Rosén and U. Sander, "Pedestrian fatality risk as a function of car impact speed," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 536–542, 2009.

[4] Rosén, H. Stigson, and U. Sander, "Literature review of pedestrian fatality risk as a function of car impact speed," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 25–33, 2011.

[5] M. Althoff and S. Magdici, "Set-based prediction of traffic participants on arbitrary road networks," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 2, pp. 187–202, 2016.

[6] M. Koschi and M. Althoff, "Set-based prediction of traffic participants considering occlusions and traffic rules," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 249–265, 2021.

[7] M. Schratter, M. Bouton, M. J. Kochenderfer, and D. Watzenig, "Pedestrian collision avoidance system for scenarios with occlusions," in *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 1133–1138.

[8] C. Pek, S. Manzinger, M. Koschi, and M. Althoff, "Using online verification to prevent autonomous vehicles from causing accidents," *Nature Machine Intelligence*, vol. 2, pp. 518–528, 2020.

[9] M. Geisslinger, F. Poszler, J. Betz, C. Lütge, and M. Lienkamp, "Autonomous driving ethics: from trolley problem to ethics of risk," *Philosophy & Technology*, vol. 34, pp. 1033–1055, 2021.

[10] M. Geisslinger, F. Poszler, and M. Lienkamp, "An ethical trajectory planning algorithm for autonomous vehicles," *Nature Machine Intelligence*, vol. 5, pp. 137–144, 2023.

[11] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, "The preregistration revolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2600–2606, 2018.

[12] C. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1292–1304, 2011.

[13] Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: a survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[14] Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1696–1703, 2020.

[15] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[16] European New Car Assessment Programme, "Test protocol — AEB VRU systems, version 3.0.4," Euro NCAP, Tech. Rep., April 2021.

[17] T. Verstraete and N. Muhammad, "Pedestrian collision avoidance in autonomous vehicles: a review," *Computers*, vol. 13, no. 3, p. 78, 2024.

[18] Motional AD LLC, "Planning with dynamic state a trajectory of an autonomous vehicle," U.S. Patent 12,060,083 B2, 2024.

[19] Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Conference on Robot Learning (CoRL)*, 2017, pp. 1–16.

[20] Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[21] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

# Explainable Adaptive Conversational AI for Improving Spoken Technical Interview Responses

Weruni S Sathsarani
*NSBM Green University*
*Colombo, Sri Lanka*
savsathsarani@students.nsbm.ac.lk

***Abstract*—Many computing students understand technical concepts but struggle to explain them aloud in interviews. AI mock-interview systems can score such answers, yet a score does not show that a student's next explanation improves. This paper presents an explainable, adaptive interview framework, reports a usability evaluation with 30 students and a technical check of its scoring fallback, and sets out a randomized study to test that outcome. The framework extends DevMock, a browser-based prototype, with concept-level scoring, explainable feedback, knowledge-gap selection and targeted follow-up questions, and it gives the scoring formulas, thresholds, prompts and rubric. Because free LLM services can be unavailable, the paper also examines how far a local fallback can be trusted. The planned study randomly assigns computing students to static, explainable or adaptive-explainable conditions. To separate real learning from simply following hints, the main outcome is an unprompted post-test (no hints) scored by human raters who do not know each student's group. The study is limited to 60 students, so it can detect only large effects and will report results mainly as estimates with confidence intervals. In the usability evaluation, 30 final-year IT students gave the system a mean rating of 4.3/5, with privacy and trust rated highest (4.5/5). These ratings show that students accepted the system, not that their explanations improved. A technical check on 30 test answers shows the limit of word matching: the fallbacks keep the system running, but neither scored a correct answer written in different words above the matching wrong answer that reused its keywords. The controlled study with students has not yet been conducted, and no hypothesis outcomes are reported.**

***Keywords—conversational AI, explainable AI, adaptive questioning, technical interview preparation, spoken language assessment, usability evaluation***

## I. Introduction

Technical interviews increasingly require candidates to explain their reasoning aloud in real time. Computing students with sound conceptual knowledge can underperform in this spoken setting, and mock interviews run by lecturers, career advisers or mentors remain limited by availability, scalability, cost and consistency.

Advances in speech recognition and large language models (LLMs) now allow an AI system to converse with a student, transcribe a spoken answer and assign a score. Prior work has examined AI-mediated interviewing through recruitment automation, algorithmic fairness and dynamic LLM evaluation [1], [2], [3], [4], [5], [6], [7]. However, a plausible score or a realistic follow-up question is not evidence that a student becomes better able to explain a concept. This paper is organized around one reframing of the problem:

*"The objective is not merely to determine whether AI can evaluate a student's technical interview answer, but to determine whether an explainable and adaptive AI interviewer can use that evaluation to help the student produce a better technical explanation on the next attempt."*

This paper reports a system design, a usability evaluation with 30 students, a technical check of the scoring fallback and a complete plan for a controlled experiment. It does not report results from that experiment, which has not been conducted. Implemented components are distinguished throughout from components specified for the study.

## II. Related Work

### *A. Review Method*

This is a targeted narrative review, not a systematic review. Sources came from IEEE Xplore, the ACM Digital Library, the ACL Anthology, ScienceDirect, SpringerLink and arXiv, using terms such as "AI interview," "mock interview," "technical interview," "explainable feedback," "adaptive questioning" and "fairness." Works were included if they were in English, published from 2013 to 2025, and addressed interview assessment, AI interview practice, feedback or fairness, with priority given to evaluated studies. Works on resume screening alone, or without any evaluation, were excluded unless they were reviews.

### *B. AI-Based Interview Assessment*

Automated systems have long predicted interview ratings from verbal and nonverbal behavior [8], and automatic short-answer grading has a long research history [9]. LLM scoring can come close to human raters on some data sets, but it remains inconsistent, especially for short responses [10]. Fabeyo [1] reviews transparency methods in hiring and finds that explainable AI can aid understanding, while technical complexity and limited human interpretability restrict its effect. This work establishes that automated assessment is feasible, but it is oriented toward selection rather than learning.

### C. Algorithmic Fairness

Köchling and Wehner [2] review discrimination in automated hiring. Speech recognition can also be unequal: five commercial systems made about twice as many word errors for Black speakers as for white speakers [11]. In two experiments (N = 113 and N = 206), Xu et al. [6] find that AI literacy, human involvement and positive outcomes improve fairness perceptions, while interactivity and detailed explanations can make the process feel fairer yet make the final decision feel less fair. Explanations therefore cannot be assumed to increase

perceived fairness, which motivates measuring fairness across conditions. An accurate technical explanation is also not always clear to its user, so this study judges explainability from the student's side.

### *D. Dynamic and Voice-Based Interview Systems*

Kim et al. [3] show that an LLM can conduct multi-turn interviews with feedback and adaptive follow-up questions, but they evaluate the model, not a human learner. Allbert et al. [4] compare five speech, LLM and text-to-speech configurations and report that most automated quality scores correlate only weakly with user ratings (below 0.11), which motivates reporting technical quality and subjective experience separately.

### *E. AI Interview Practice and Learning*

MACH [12] is the closest experimental precedent: in a weeklong trial with 90 undergraduates, human experts rated its users as improved relative to controls. Its feedback, however, addressed nonverbal behavior rather than technical content. Conversate [13], a think-aloud tool [14] and an AI mock interviewer [5] were each evaluated qualitatively with 17 to 20 participants and report perceived benefits rather than objective change. Zara [7] reports satisfaction with LLM feedback but not later performance.

Learning research adds further constraints. Feedback has a medium average effect on learning (d = 0.48) that varies by feedback type [15], yet over one third of feedback interventions in an earlier meta-analysis reduced performance [16]. Feedback works best when it is specific and focused [17]. Assistance can raise immediate performance without producing durable learning [18], while recalling knowledge without hints improves later retention [19]. An adaptive interviewer must therefore be tested on what students can do after the assistance is removed.

### *F. Synthesis of Literature*

Table I compares the closest studies with this paper. None combines spoken technical content, concept-level feedback, adaptive questioning and an objective outcome measured without hints.

The reviewed work establishes that automated evaluation is feasible [8], [4], that explanation shapes fairness perceptions in complex ways [2], [6], that adaptive questioning is achievable [3], and that AI practice is well received [13], [14], [5], [7]. Whether these components, combined in one loop, improve a later spoken explanation remains largely unexamined.

TABLE I. COMPARISON OF EXISTING RESEARCH AND BASELINE APPROACHES

| Study | Focus and method | Main finding | Spoken input | Objective learning outcome | Limitation for this question |
|---|---|---|---|---|---|
| Kim et al., 2025 [3] | LLM evaluation; multi-turn interviews | Follow-up questions based on feedback are feasible | No | No | Evaluates the LLM, not a human |
| Hoque et al., 2013 [12] | MACH; controlled trial, N = 90 | Expert-rated interview improvement over controls | Yes | Yes | Nonverbal focus; no technical content |
| Daryanto et al., 2025 [13] | Conversate; qualitative study, N = 19 | Two-way feedback is valued for reflection | Yes | No | Perceptions only |
| Gomez et al., 2025 [5] | AI mock interview; formative study, N = 20 | Perceived realism, usefulness and confidence | Yes | No | Self-report; no comparison group |
| Yazdani et al., 2025 [7] | Zara; deployed feedback system | Feedback is feasible; satisfaction 4.37/5 | Yes | No | Later performance not measured |
| This paper | Usability evaluation, N = 30; randomized study planned | Usability rated 4.3/5; learning study not yet conducted | Yes | Planned | Single setting; self-report so far |

## III. RESEARCH GAP AND CONTRIBUTIONS

Existing work is spread across separate research areas. No identified study integrates spoken technical-response assessment, concept-level evaluation, explainable feedback, adaptive questioning and an objective learning outcome while also measuring trust, perceived fairness and privacy. The research gap is therefore limited empirical evidence on whether an explainable, adaptive AI interviewer can use its evaluation of a spoken response to improve the student's later explanation.

The paper makes three contributions: (1) a complete feedback loop with clear scoring rules that others can repeat; (2) a usability evaluation with 30 students and a controlled study design that separates learning from following hints; and (3) a clear account of what privacy protection and a local fallback can and cannot do, with a repeatable technical check. Fig. 1 shows the conceptual model.

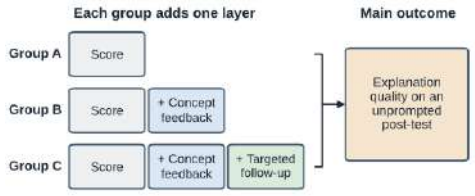



Fig. 1. Conceptual model: each condition adds one layer to the one before it.

## IV. Proposed Framework

### A. System Architecture and Implementation Status

DevMock is a React, Node.js/Express and MongoDB web application. It implements speech capture, transcript correction, scoring and per-question feedback. It does not implement knowledge-gap selection or adaptive follow-up; both are specified here and will be built for the study. Fig. 2 shows the complete loop and its implementation status.

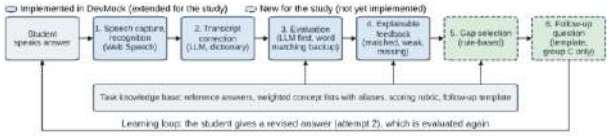



Fig. 2. Architecture and learning loop of the proposed interviewer, showing implementation status.

### B. Speech Processing

Answers are transcribed with the browser Web Speech API (en-US). Each segment is corrected by an LLM request (llama-3.1-8b-instant, temperature 0.1), with a technical-term dictionary as fallback. The en-US setting may mishear Sri Lankan English, so the pilot will compare it with the nearest available English setting, such as en-IN, and keep the one with the lowest word error rate. Each participant will also read a fixed passage, which gives a word error rate for each student that can be used in the analysis.

### C. Evaluation: Existing Prototype

Each transcript is sent with the reference answer and core keywords in one LLM request (llama-3.3-70b-versatile, temperature 0.1, 500 tokens), which returns a score from 0 to 10, matched and missing keywords, strengths and improvements. Before this step, answers under five words score 2, and answers containing a phrase such as "I don't know" score 1. The LLM score and the word-matching score are never combined. The word-matching score is used only when the LLM is unavailable:

where K is the percentage of core keywords found and Sim is the similarity between the answer and the reference answer:

$$\Sigma\mu = 0.7 \cdot J + 0.3 \cdot B \quad (2)$$

Here J is the percentage of words shared with the reference answer, counting words longer than three characters, and B measures how well each answer sentence matches its closest reference sentence. Sim is kept between 15 and 100. The final score is rounded to a 0–10 scale, and a session passes at 60%.

The code uses word overlap rather than TF–IDF cosine similarity, so it cannot give credit for a correct answer written in different words.

### D. Evaluation: Study Specification

For the study, scoring becomes concept-level, with weights w = 2 for core and w = 1 for supporting concepts. Using the prompt in Appendix A, the LLM returns a credit $a \in \{0, 0.5, 1\}$ and an exact evidence quote for each concept; credit is reset to 0 if the quote does not occur in the transcript. Concept coverage is (3) Here w is the concept's weight and a is its credit. If the LLM fails after one retry, a = 0.5 for any concept named in the transcript and 0 otherwise, and the response is flagged. Naming a concept never earns full credit, as in the rubric. Weights and credits are frozen after the pilot. CC drives feedback only, so an evaluator error cannot directly alter the main outcome, which is scored by humans. An LLM may still give credit for answers that use the right terms but explain them wrongly, so this risk is tested directly (Section V-E). The check for "I don't know" answers will match whole answers only, because it now treats answers such as "pass by reference" as non-answers (Section VI-B). The prototype's two fallbacks will be replaced by this rule.

### E. Explainable Feedback and Adaptive Questioning

Feedback lists concepts as matched (a = 1), weak (a = 0.5) or missing (a = 0), with a short suggestion for each concept below full credit. The model answer is not shown, so feedback points to gaps without supplying the answer. A concept is a gap if a < 1, and the follow-up targets the gap with the highest priority:

Ties are broken by list order. The follow-up uses a fixed template: "You have not fully explained {concept}. In your own words, how does {concept} apply to {topic}?" If no gap exists, the fixed re-prompt in Table II is used. This fixed rule can be reproduced exactly.

### F. Privacy-Protecting Design

The prototype stores no audio or video. In the study, audio will be kept only with separate consent, encrypted, for human rating of spoken delivery, and deleted afterward. Collecting less data is not the same as processing data locally, however: browser speech recognition may send audio to the browser vendor's service [20], and transcripts are sent to the LLM provider. Only transcripts, scores and questionnaire data stored under code numbers are kept. The in-browser eye-contact and posture overlay is disabled in the study.

### G. Fallback Design

DevMock uses the free tier of the Groq API. Free services can stop responding or reject requests during heavy use, and rate limits were reached during development testing. A practice session should not stop when this happens. The prototype switches to a local check that works without the LLM when the LLM returns an error, times out, reaches a rate limit or sends an invalid reply. Fig. 3 shows the planned study design, which first tries the LLM once more. In development testing, the local check took about 0.5 s, compared with about 1.2 s for the LLM.

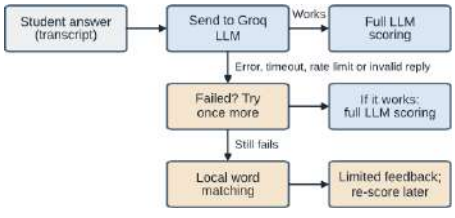



Fig. 3. Planned fallback design when the Groq LLM is unavailable.

The fallback keeps the session running, but it only matches words and cannot tell whether a student understands a concept

(Section VI-B). The study therefore treats it as a backup, not as a second evaluator. Students are told that the full review is unavailable, and no pass or fail result is shown. The answer is flagged and scored again by the LLM when the service returns. How often the fallback runs will be reported by group, and the main analysis will be repeated without participants who received fallback feedback.

## V. Planned Controlled Study

### *A. Design, Participants and Sample Size*

The study is a small randomized controlled experiment with 60 computing students from one university. They must be aged 18 or over, have completed a programming course and not have joined the usability evaluation. Six students will take part in a pilot that checks timing, tasks, the speech setting and scoring. The other 54 will be randomly assigned to three equal groups of 18. A computer will assign students in random blocks, balanced by program level and prior interview experience, and the group will stay hidden until after the pre-test. Group characteristics, including first language and accent region, will be reported.

Allowing for 10% dropout, 16 students per group are expected to complete the study. The sample size is set by the students available, not by a power target. With 16 per group, $\alpha = 0.025$ and a pre-test to post-test correlation of 0.5, the study has 80% power only for a very large effect ($d = 0.98$). The value is 0.81 if the correlation is 0.7 and 1.08 if it is 0.3. The value of 0.025 is the conservative case: under the Holm correction it applies to the smaller p-value, and the second comparison is tested at 0.05. Power for a medium effect ($d = 0.5$), close to the average feedback effect ($d = 0.48$) [15], is only 26%. The study is therefore treated as an early randomized study. A non-significant result will not be read as proof of no effect.

### *B. Experimental Conditions*

Table II defines the conditions. All groups use the same hidden evaluator and the same number of spoken turns, so they differ only in what the student sees and hears. B vs. A isolates explainability; C vs. B isolates adaptivity.

TABLE II. Experimental Conditions

| Feature | Group A: static | Group B: explainable | Group C: adaptive |
|---|---|---|---|
| Evaluator | Same (hidden) | Same | Same |
| Shown after attempt 1 | Overall score | Score and concept feedback | Score and concept feedback |
| Prompt for attempt 2 | Fixed re-prompt* | Fixed re-prompt* | Gap-targeted follow-up |
| Spoken turns per task | 2 | 2 | 2 |
| Model answer shown | No | No | No |

*"Please explain {topic} again, as completely as you can."

### *C. Procedure and Control of Cueing Bias*

In a session of about 45 minutes (Fig. 4), students complete attempt 1 on four tasks, receive their condition and give an assisted attempt 2 on each task. Gap-targeted prompts can raise attempt 2 simply by telling students what to say. This is called cueing bias, so attempt 2 is not used as the learning outcome. After a 10-minute unrelated filler task, students complete an unprompted post-test with no feedback, hints or follow-ups. To reduce memorization, the post-test covers the same concepts in new wording: "A junior developer asks what {topic} means. Explain it." All groups have the same practice, so any practice effect should not favor one group. Because the reworded post-test stays close to the practice questions, the four transfer tasks at the 7-day session are the key secondary outcome: they use new scenarios and show whether learning carries over. The 7-day and 30-day sessions also repeat the reworded post-test to measure what students retain, although 30 days is still not long-term retention. Task order will be rotated across students. Sessions will follow a written script, and a researcher will check a random 10% of feedback for accuracy.

### *D. Hypotheses and Research Questions*

H1 and H2 are the main hypotheses and use the unprompted post-test. H1: explainable feedback (B) produces higher post-test quality than score-only feedback (A). H2: adaptive follow-up (C) produces higher post-test quality than explainable feedback alone (B). Four exploratory research questions follow. RQ1: Do trust ratings differ across conditions? RQ2: Do fairness and privacy ratings differ across conditions? RQ3: Is improvement related to change in confidence? RQ4: Are technical-quality measures related to satisfaction?

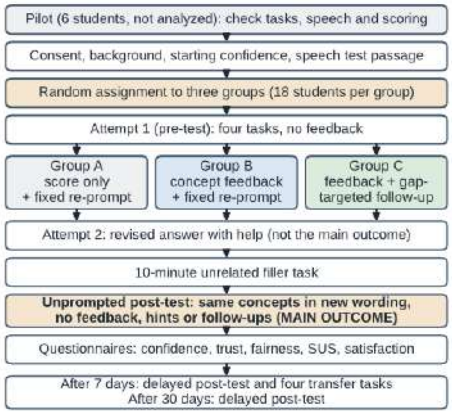



Fig. 4. Study procedure and control of cueing bias.

### *E. Human Scoring, AI Validation and Baselines*

Two trained raters will score every attempt 1, attempt 2, post-test, delayed and transfer response with the rubric in Appendix B, without knowing the condition, attempt or automated score. Raters will practice on pilot responses first. A third rater resolves differences of more than one credit level; otherwise, the two scores are averaged. Besides CC, raters give an explanation quality score for logical reasoning, links between concepts, technical depth and clarity. For students who consent to audio storage, they also rate pronunciation, fluency, intelligibility and delivery from the recording, because a transcript cannot show these.

The evaluator will be tested in two ways before the main study. First, it will score pilot responses and a keyword-trap set of 40 test answers, in which half use the right terms but explain them wrongly or incompletely. It passes only if weighted $\kappa$

against the agreed human scores is at least 0.70 and false credit on trap answers is at most 10%; otherwise the prompt will be revised and tested on new answers. Second, it will be compared with all human scores from the main study, by concept and by topic, with concept-level precision and recall. Three automated baselines will be compared on the same answers: the word-matching score in (1), the original prototype prompt and Sentence-BERT similarity [21].

### *F. Data Collection*

The system logs transcripts, concept credits, prompts, response times and fallback events. Confidence is measured before and after the session. The questionnaire uses the 10-item System Usability Scale (SUS) [22], items adapted from published trust [23] and fairness [24] scales, and the single items for confidence, transparency, usefulness, satisfaction and privacy in Appendix C. Prompts, rubric and concept lists are in the appendix, and the prototype source code is available at https://github.com/satharasingha/devmock-interview-coach.

### *G. Statistical Analysis*

The main outcome is the human-rated post-test CC averaged across the four tasks, analyzed in the group each student was assigned to. Groups will be compared with analysis of covariance (ANCOVA) [25], which adjusts post-test scores for pre-test score, program level and interview experience. Its assumptions (normal residuals, equal variance and similar pre-test slopes) will be checked; if they are not met, bootstrapped intervals will be used. The two planned comparisons are Group B vs. Group A (H1) and Group C vs. Group B (H2), with the Holm correction [26]. Transfer-task CC, the key secondary outcome, uses the same model and comparisons. Every result will be reported as a score difference with a 95% confidence interval and a standardized effect size. With 16 per group, these intervals are expected to be about ±0.62 standard deviations wide, so results will be estimates rather than firm conclusions.

Agreement between raters will be reported with a two-way random-effects, absolute-agreement, average-measures intraclass correlation, ICC(2,k) [27], because the outcome uses the average of two raters. An ICC of 0.75 or above is treated as good. Weighted κ is reported by concept, and both have 95% confidence intervals. Missing outcomes will be reported by group and handled with multiple imputation (20 data sets, using pre-test score, group and strata), alongside results from complete cases. A mixed-effects model will repeat the analysis with task and word error rate included, and results will be shown for each topic.

Hint dependence shows how much a student's human-rated score falls when help is removed:

RQ1 and RQ2 compare questionnaire scores across groups, and RQ3 and RQ4 use correlation or regression. All of these are exploratory.

### *H. Fairness and Reporting*

Word error rate and the difference between automated and human scores will be reported by first language and accent region, with 95% confidence intervals. A difference of 5 CC points or more between groups will be flagged for review. With 60 students, these groups will be small, so this check can show warning signs but cannot prove fairness. The study plan will be registered before data collection, and results will follow the CONSORT reporting guideline [28], with attempt 2 reported separately as an assisted measure and any changes disclosed. Instruments, code and anonymized data will be shared where consent allows.

## VI. Current Evaluation Results

### *A. Usability Evaluation with 30 Students*

Thirty final-year undergraduate IT students at one university (15 women and 15 men) used DevMock in a structured interview session and then rated it on five items from 1 to 5, adapted from the SUS. Fig. 5 shows the results. Every area averaged 4.1 or above, and the overall mean was 4.3 (SD 0.6). Privacy and trust scored highest (4.5, SD 0.4), which suggests that the no-storage design was well received. Willingness to recommend scored lowest and varied most (4.1, SD 0.8). Table III gives the context for these results.

These ratings show that students accepted the system. They are self-reported, have no control group and no pre-use measure, and so they do not show that explanations improved or test H1 or H2. Age and language background are not available. Two percentages in the project record, human–AI agreement and the share of participants reporting increased confidence, are not reported because they cannot be reconciled with the recorded sample sizes. The record also gives a Cohen's κ of 0.76, but it does not say whether κ was weighted or how the three raters were combined, so the value is treated as unverified.

TABLE III. Usability Evaluation and System Test Context

| Metric | Value | Basis |
|---|---|---|
| κ, AI vs. humans (unverified) | 0.76 | 20 answers, 5 questions, 3 raters; LLM mode; method not documented |
| Overall usability | 4.3/5 (SD 0.6) | N = 30 (15 women, 15 men); five items adapted from SUS, not a SUS score |
| LLM / fallback evaluation time | ≈ 1.2 s / ≈ 0.5 s | Development testing |
| Concurrent load | 10 users, stable | Development testing |

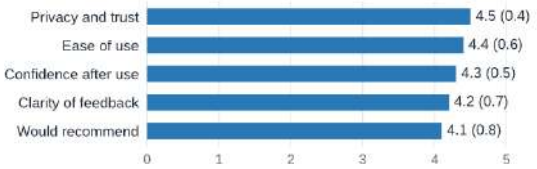



Fig. 5. Mean usability ratings from 30 students on a 1–5 scale (SD in brackets).

### *B. Technical Check of the Word-Matching Fallback*

To test the design without participants, the two word-matching fallbacks in the prototype's public code (commit a109930) were run on the ten questions in its sample question bank. Three answer sets were fixed before scoring: the reference answer, a correct answer written in different words, and an incorrect answer that reuses the core keywords. Fig. 6 shows all 30 answers on the application's 0–100 scale, where 60 is the pass mark. The correct and incorrect test answers were written

for this check and are not student responses, and LLM scoring was not tested.

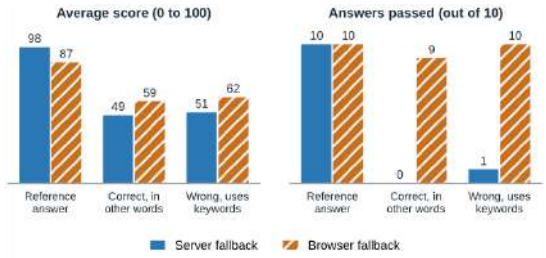



Fig. 6. Results of the fallback check on 30 test answers.

Both fallbacks gave the reference answers high scores, but they could not separate the other two sets. On neither fallback did a correct answer score above the matching incorrect answer: the server version tied on 9 of 10 questions and the browser (client) version on 8, and the incorrect answer scored higher on the rest. The server version failed all ten correct answers, while the browser version passed all ten incorrect answers. In a separate check of six correct technical sentences, four were scored as non-answers (score 1) because they contained the letters "pass" or "skip". These results cover only test answers and one code version, but they support three design decisions: scoring concepts rather than word overlap, keeping word-matching scores out of the main outcome, and fixing the "I don't know" check.

## VII. Discussion

The framework applies the multi-turn interview approach of Kim et al. [3] to a human learner and adds objective measurement to qualitative AI practice studies [13], [14], [5]. If Group C outperforms Group B on attempt 2 but not on the post-test, adaptive prompts would be aiding performance rather than learning, a problem known as the assistance dilemma [18]. Because explanations can have mixed fairness effects [6], RQ2 could also reveal a cost that a learning-only evaluation would miss. The technical check adds a practical caution: a fallback can keep a system available and still fail to separate correct from incorrect answers. Apart from that check, these points are implications of the design, not findings.

## VIII. Ethical and Privacy Considerations

Recruitment will begin only after institutional ethics approval, with written informed consent and the right to withdraw without academic consequence. Consent materials will state that speech is processed by the browser's recognition service, that transcripts are sent to an external LLM provider, and that audio storage is optional. Transcripts will be stored under code numbers, and individual data will not be shared with instructors. The system is a practice tool, not a hiring or selection system; other uses would need separate validation and governance.

## IX. Limitations

No experimental evidence yet shows that the system improves interview performance. With 60 students from one university, the study can detect only large effects, and subgroup results will be uncertain. The four topics cover only part of computing, and results will apply only to these tasks, not to real interviews or to other subjects. Practice effects may remain, and students could discuss the conditions. CC and the quality score may still miss some strengths of a good explanation, and delivery can be rated only for students who consent to audio storage. Speech recognition may still treat some accents unequally. A 30-day test does not show long-term retention.

## X. Conclusion

AI interview research should move beyond whether a system can produce plausible questions or accurate scores toward whether it helps a student explain better. This paper specifies an evaluate–explain–adapt–practice loop on the DevMock prototype, reports strong usability ratings from 30 students and sets out a randomized study to test its effect with independent human scoring and an unprompted post-test. These ratings show acceptance, not learning. The technical check shows that a local fallback keeps practice running but should not be trusted to judge understanding. Future work will build the new components, improve the fallback, complete the pilot, obtain ethics approval and report results from real data, followed by a larger study across more topics and universities.

## Appendix

*A. LLM evaluation prompt. System: "You score a spoken technical answer against a fixed concept list. Use only the transcript. It may contain speech-recognition errors; do not penalize these if the meaning is clear. Return only valid JSON." User: "QUESTION: {q}. REFERENCE ANSWER: {ref}. CONCEPTS: {id, name, aliases}. TRANSCRIPT: {t}. For each concept return {id, credit, evidence, suggestion}. credit: 0 = absent or incorrect; 0.5 = named or vague; 1 = correctly explained. evidence: an exact quote from the transcript, or an empty string. suggestion: one sentence on what to explain better, without giving the answer, or an empty string if credit is 1." The final model identifier will be reported.*

*B. Human scoring rubric. For each concept, score 0 if absent or incorrect; 1 if named, vague or partly correct; 2 if correctly explained in the speaker's own words. A definition without an explanation of the concept's role scores 1. Ignore grammar, accent and recognition errors when the meaning is clear. Scores map to a = 0, 0.5 and 1 in (3). Raters also score logical reasoning, links between concepts, technical depth and clarity from 0 (weak) to 2 (strong), and, from consented audio, pronunciation, fluency, intelligibility and delivery on the same scale.*

*C. Questionnaire items (1 = strongly disagree, 5 = strongly agree). Confidence: "I can explain technical concepts clearly in an interview." Transparency: "I understand why I received this feedback." Usefulness: "The feedback helped me improve my answer." Satisfaction: "Overall, I am satisfied with this practice session." Privacy: "I am comfortable with how this system handles my data."*

*D. Tasks and reference concept lists. Table IV extends the DevMock question bank. Each transfer prompt uses the same concepts in a new scenario and the same list.*

TABLE IV. INTERVIEW TASKS AND REFERENCE CONCEPT LISTS

| Task prompt (topic) | Core concepts (w = 2) | Supporting concepts (w = 1) | Transfer prompt |
|---|---|---|---|
| What is a REST API? (REST) | Stateless requests; HTTP methods; resources identified by URIs; client–server separation | JSON representations; status codes | Why can a stateless API scale behind a load balancer? |
| What is a database index? (Indexing) | Separate lookup structure; faster reads; avoids full table scan; write and storage overhead | B-tree; query optimization | Why might many indexes slow an order-entry system? |
| What is multithreading? (Threads) | Multiple threads in a process; shared memory; concurrent execution; synchronization need | Parallelism; race condition | Two threads increment a shared counter and the total is wrong. Why? |
| What is a deadlock? (Deadlock) | Mutual exclusion; hold and wait; no preemption; circular wait | Lock ordering; resource allocation | Two transfers each lock one account and wait for the other. Explain and fix this. |

## REFERENCES


[1] S. Fabeyo, "Explainable AI in employment decision-making: A systematic review of transparency methods in hiring algorithms," Issues in Information Systems, vol. 26, no. 3, pp. 127–135, 2025, doi: 10.48009/3_iis_2025_2025_110.

[2] Köchling and M. C. Wehner, "Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development," Business Research, vol. 13, pp. 795–848, 2020, doi: 10.1007/s40685-020-00134-w.

[3] E. Kim, J. Suk, S. Kim, N. Muennighoff, D. Kim, and A. Oh, "LLM-as-an-Interviewer: Beyond static testing through dynamic LLM evaluation," in Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 26456–26493, doi: 10.18653/v1/2025.findings-acl.1357.

[4] R. Allbert, N. Yazdani, A. Ansari, A. Mahajan, A. Afsharrad, and S. S. Mousavi, "Evaluating speech-to-text × LLM × text-to-speech combinations for AI interview systems," arXiv:2507.16835, 2025.

[5] N. Gomez, S. S. Batham, M. Volonte, and T. D. Do, "Virtual interviewers, real results: Exploring AI-driven mock technical interviews on student readiness and confidence," in Companion Publ. 2025 Conf. Computer-Supported Cooperative Work and Social Computing, 2025, pp. 209–213, doi: 10.1145/3715070.3749227.

[6] Y. Xu, Z. Chen, and M. Dong, "Shaping the fairness journey: The roles of AI literacy, explanation, and interpersonal interaction in AI interviews," Int. J. Human-Computer Studies, vol. 205, art. 103629, 2025, doi: 10.1016/j.ijhcs.2025.103629.

[7] N. Yazdani, A. Mahajan, and A. Ansari, "Zara: An LLM-based candidate interview feedback system," arXiv:2507.02869, 2025, doi: 10.48550/arXiv.2507.02869.

[8] Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated analysis and prediction of job interview performance," IEEE Trans. Affective Comput., vol. 9, no. 2, pp. 191–204, 2018, doi: 10.1109/TAFFC.2016.2614299.

[9] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," Int. J. Artif. Intell. Educ., vol. 25, no. 1, pp. 60–117, 2015, doi: 10.1007/s40593-014-0026-8.

[10] M. D. Shermis, "Using ChatGPT to score essays and short-form constructed responses," Assessing Writing, vol. 66, art. 100988, 2025, doi: 10.1016/j.asw.2025.100988.

[11] Koenecke et al., "Racial disparities in automated speech recognition," Proc. Natl. Acad. Sci. USA, vol. 117, no. 14, pp. 7684–7689, 2020, doi: 10.1073/pnas.1915768117.

[12] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing, 2013, pp. 697–706, doi: 10.1145/2493432.2493502.

[13] T. Daryanto, X. Ding, L. T. Wilhelm, S. Stil, K. M. Knutsen, and E. H. Rho, "Conversate: Supporting reflective learning in interview practice through interactive simulation and dialogic feedback," Proc. ACM Hum.-Comput. Interact., vol. 9, art. GROUP009, 2025, doi: 10.1145/3701188.

[14] T. Daryanto et al., "Designing conversational AI to support think-aloud practice in technical interview preparation for CS students," in Proc. 2025 IEEE Symp. Visual Languages and Human-Centric Computing, 2025, pp. 191–203, doi: 10.1109/VL-HCC65237.2025.00029.

[15] Wisniewski, K. Zierer, and J. Hattie, "The power of feedback revisited: A meta-analysis of educational feedback research," Frontiers in Psychology, vol. 10, art. 3087, 2020, doi: 10.3389/fpsyg.2019.03087.

[16] N. Kluger and A. DeNisi, "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory," Psychological Bulletin, vol. 119, no. 2, pp. 254–284, 1996, doi: 10.1037/0033-2909.119.2.254.

[17] V. J. Shute, "Focus on formative feedback," Rev. Educ. Research, vol. 78, no. 1, pp. 153–189, 2008, doi: 10.3102/0034654307313795.

[18] R. Koedinger and V. Aleven, "Exploring the assistance dilemma in experiments with cognitive tutors," Educ. Psychol. Rev., vol. 19, no. 3, pp. 239–264, 2007, doi: 10.1007/s10648-007-9049-0.

[19] H. L. Roediger and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," Psychological Science, vol. 17, no. 3, pp. 249–255, 2006, doi: 10.1111/j.1467-9280.2006.01693.x.

[20] Mozilla, "Web Speech API," MDN Web Docs. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API. Accessed: Nov. 2, 2026.

[21] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

[22] Brooke, "SUS: A 'quick and dirty' usability scale," in Usability Evaluation in Industry, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

[23] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," Int. J. Cognitive Ergonomics, vol. 4, no. 1, pp. 53–71, 2000, doi: 10.1207/S15327566IJCE0401_04.

[24] J. A. Colquitt, "On the dimensionality of organizational justice: A construct validation of a measure," J. Applied Psychology, vol. 86, no. 3, pp. 386–400, 2001, doi: 10.1037/0021-9010.86.3.386.

[25] J. Vickers and D. G. Altman, "Analysing controlled trials with baseline and follow up measurements," BMJ, vol. 323, no. 7321, pp. 1123–1124, 2001, doi: 10.1136/bmj.323.7321.1123.

[26] S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Statistics, vol. 6, no. 2, pp. 65–70, 1979.

[27] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," J. Chiropractic Medicine, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

[28] F. Schulz, D. G. Altman, and D. Moher, "CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials," BMJ, vol. 340, art. c332, 2010, doi: 10.1136/bmj.c332.

# Smart Urban Waste Collection System

Hansika Subhashini Keeragala
*Department of Software Engineering*
*NSBM Green University*
*Colombo, Sri Lanka*
hansikasubhashini50@gmail.com

G. Perera
*Department of Computer and Data Science*
*NSBM Green University*
*Colombo, Sri Lanka*
gayanp@nsbm.ac.lk

***Abstract*—Missed garbage collection is a common problem in urban areas such as Kalubowila, Wellawatta, Maharagama and Kottawa. Field observations and discussions with residents and garbage collectors showed that one main problem is poor communication. Residents often do not know when the garbage collection vehicle will arrive, which can result in waste being left on roadsides and in canals. Most existing waste management systems focus on operational efficiency and give residents little information. This research aims to develop a resident-centred Smart Urban Waste Collection System to improve communication between residents, garbage collectors and the municipal council. The system includes a Flutter mobile application, a Firebase backend, a municipal dashboard, GPS based vehicle tracking, geo-fencing push notifications, a community incident-reporting module and an estimated time of arrival (ETA) service. The ETA service uses a Random Forest model to estimate when the vehicle will reach a resident's location, with a distance and speed calculation as a fallback. The study follows a Pragmatist research approach and Design Science Research methodology, with Agile Scrum used for development. Requirements were collected through a resident survey, interviews with collectors and discussions with the municipal council. The prototype was evaluated through API, backend, Firebase, front-end and user acceptance testing. The ETA model was compared with a traditional distance and speed approach using MAE and RMSE. Due to project limitations, live GPS tracking was not completed, so simulated coordinates were used for location-based features. The ETA model was trained using limited real data supported by synthetic data. The developed system aims to reduce missed waste collections, improve communication and support better waste management in urban Sri Lanka. The Maharagama Municipal Council has accepted the project with some improvements and expressed interest in supporting its future development.***

***Keywords—Smart waste management, resident communication, GPS tracking, geo-fencing, push notifications, ETA prediction, municipal waste management***

## I. Introduction

Urbanization has significantly increased municipal solid waste generation worldwide, creating major waste management challenges [1], [9]. In Sri Lanka, approximately 7,000 to 10,000 metric tons of municipal solid waste are generated daily, while only half is collected by local authorities [13]. The Western Province accounts for over 59% of the country's total waste generation, with Colombo facing significant challenges due to population growth and irregular waste collection [13]. Unpredictable collection schedules leave residents uninformed about vehicle arrival times, resulting in improper waste disposal along roadsides and canals. This contributes to environmental pollution, public health risks, and urban flooding [9], [13]. Although existing studies focus on GPS tracking, route optimization, and machine learning based ETA prediction [2][4], limited attention has been given to real-time communication with residents. To address this gap, this research proposes a smart waste collection system that uses machine learning to predict waste collection vehicle arrival times and provides real-time notifications to residents, improving communication and collection efficiency.

### *A. Research Aim and Objectives*

This study proposes the design, development and evaluation of intelligent real-time waste collection tracking and notification system (WCTNS) which would enhance the reliability of waste collection and minimize environmental pollution in urban residential area such as Colombo city of Sri Lanka by incorporating GPS tracking, mobile application and machine learning arrival-time model. To support the objectives, the following will be done:

- Identify communication barriers that lead to missed collections and data inputs required to arrive-time predict.
- Compare and analyze current route coverage and compare a Random Forest regression model and a rule-based ETA estimator by comparing their MAE and RMSE values.
- Design and develop a mobile app for the residents, a geo-fencing notification system and a model using Random Forest.
- Utilize technical performance testing and user satisfaction/residence surveys to review the system.

### *B. Research Questions*

- RQ1: Is real-time GPS tracking of collection vehicles able to reduce the number of missed waste collection in urban residential areas?
- RQ2: Does early push notification improve resident participation and timeliness for waste collection?
- RQ3: What is the accuracy of the Random Forest model in predicting the arrival time of vehicles compared to the rule-based approach?
- RQ4: Can improved communication between residents and collectors be measured as a decrease in roadside and canal pollution?

### *C. Research Motivation*

This study is inspired by the regular perceptible littering on the streets, overfilled canals and foul smell in the surroundings while driving through the Hathbodhiya Road area of Wellawatta and Wellawatta through Maharagama and Kottawa. The primary problem was found to be a lack of communication, something that was noted in interviews with residents, the

boarding house owners and the three-wheel drivers. Residents also recommend that a truck's current location be displayed, like ride-hailing apps do for the location of a driver, to address much of the uncertainty about collection times. This lived experience, rather than an abstract technology gap, motivates a resident-centred rather than a purely municipal-efficiency-centred system design.

## II. Problem Statement

While waste collection services are established in urban areas, there is often a delay in the collection of garbage due to the unpredictability of the arrival of vehicles and lack of real time communication. The existing systems are based on route optimization and lack of accurate ETA prediction and resident notification, which results in improper waste disposal and environmental pollution.

## III. Literature Review

### *A. IoT-Based Smart Bin Systems*

Ghahramani et al. proposed a cloud-based IoT waste management system that uses bin sensors and AI-driven adaptive routing to generate efficient collection routes, but it did not include any resident-facing communication features [1]. Neema and Gor designed an energy-efficient smart-bin system using a nearest-neighbour routing algorithm that reduced energy consumption by roughly 34 percent, again without any resident notification capability [2]. Both studies highlight operational efficiency for municipalities while leaving the resident communication gap unaddressed.

### *B. GPS Tracking and Mobile Notification*

GPS-based fleet tracking is well established for monitoring vehicle location, speed and route coverage in logistics and public service delivery. Rathod and Agarwal combined augmented reality with IoT and AI so that waste-management staff could view bin conditions and plan routes in real time, improving operational awareness but not providing predictive alerts to residents [3]. In ride-hailing platforms, combining live GPS tracking with ETA prediction and push notification has been shown to substantially reduce user uncertainty about waiting times [4], a pattern that has not yet been widely applied to municipal waste collection.

### *C. Machine Learning for Arrival Time Prediction*

Sulistio and Ahmad applied IoT and machine learning to forecast waste generation and improve collection scheduling, but their model predicted quantities rather than vehicle arrival time [5]. In the transportation domain, Random Forest regression has been widely used to predict travel time and estimated arrival time because it captures non-linear relationships between speed, distance, time of day and road conditions even with modest dataset sizes [6], [7]. Random Forest itself was introduced by Breiman as an ensemble of decision trees trained on random samples of data and features, which reduces overfitting and improves generalisation [8].

### *D. Smart Waste Management Initiatives in Sri Lanka*

Existing studies on IoT-enabled waste management mainly focus on bin monitoring and collection-route optimization rather than resident communication [9], [10]. In Sri Lanka, Hemapriya et al. introduced a rewards-based smart bin application [11], while Baddegama et al. developed a web-based platform for waste hotspot mapping and dynamic collection scheduling [12]. However, neither provides real-time vehicle tracking for residents. Furthermore, Fernando identified inconsistent waste collection and poor resident communication as key challenges among local governments in Sri Lanka's Western Province [13]. These limitations highlight the need for a resident-focused waste collection system.

### *E. Research Gap*

While IoT-based bin monitoring, GPS fleet tracking and machine learning arrival-time prediction have each been studied individually, no existing system combines real-time GPS tracking of waste vehicles with direct resident notification, and no prior study has compared a Random Forest ETA model against a rule-based baseline for waste-collection vehicles in the Sri Lankan urban context. This research addresses both gaps by developing a resident-centred tracking and notification system supported by a data-driven arrival-time model.

## IV. Methodology

The study is conducted in a Pragmatist research paradigm, which is a research paradigm that determines the best solution by its practical consequences, rather than by whether it conforms to any single approach to research or philosophy; the study's objective is to design, construct and validate a system that addresses a real-world observed problem, not to prove an isolated theory. The study flows from a human-centred approach that begins with observation of the streets in Kalubowila village, Wellawatta, Kottawa and Maharagama, and a machine learning approach based on deductive reasoning, where the testable hypothesis is that a Random Forest model accurately predicts the time of arrival as compared to a rule-based formula. Design Science Research (DSR) is the overall strategy adopted and builds in an iterative and multi-component fashion using Agile.

### *A. System Architecture*

The proposed artefact follows a three layer architecture. A field layer consisting of a GPS-enabled device on each collection vehicle, a backend layer that ingests location data, runs the Random Forest ETA model and manages geo-fence triggers, and an output layer consisting of a resident-facing mobile application that displays live vehicle location and delivers push notifications shortly before a vehicle reaches a resident's street.

### *B. Data Preprocessing*

To train the model, GPS and routing information need to be preprocessed. This includes:

- Handling of missing values: Fill in short segments of GPS signal with linear interpolation between the last and next GPS point.
- Feature engineering: transforming raw GPS coordinates to a "distance to destination" feature with Haversine formula and a "time since last stop" feature from timestamps.
- Normalization: Rescaling numerical features (speed, distance, historical travel time) so they are in the same range to avoid that one feature dominates the training of the model.

Train/test split: The 80/20 split is suggested and additional validation subset (10% of the training set) is suggested for hyperparameter tuning.

### C. Machine Learning Model

The prediction task in ETA is a supervised regression problem. Two important factors are considered: First, the limited size of the initial dataset, particularly during the pilot stage and second, the need to analyze the non-linear relationships between traffic conditions and travel time, which makes Random Forest and XGBoost promising models for the limited tabular, feature-based regression task at hand. The features of the input are:

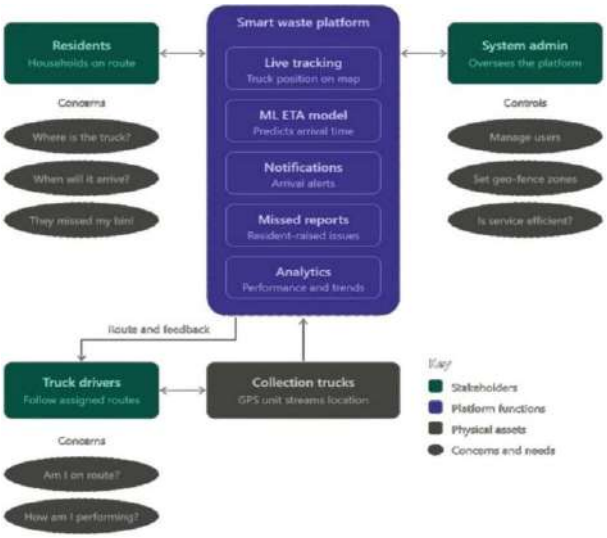



Fig. 1. Rich picture of the proposed Smart Urban Waste Collection System.

- Vehicle speed
- Distance to destination
- Historical average travel time for the route segment
- Time of day and day of week
- Traffic condition indicator

The predicted ETA is computed as:

$$\text{ETA} = T_{\text{current}} + (D_{\text{remaining}} / V_{\text{avg}}) + T_{\text{traffic}} \quad (1)$$

where $T_{\text{current}}$ is the current timestamp, $D_{\text{remaining}}$ is the remaining distance to the destination, $V_{\text{avg}}$ is the average vehicle speed for the current route segment and $T_{\text{traffic}}$ is a learned traffic delay adjustment produced by the model rather than a fixed constant.

### D. Notification System

The notification system is based on the concept of 'geofencing' [14], in which a virtual radius of R is created around each resident's registered address. If the GPS position that the vehicle is sending is within this boundary, a notification will be generated:

$$d(\text{vehicle, resident}) \leq R \quad (2)$$

where $d$ is the real-time distance between the vehicle and the resident's location. The three types of notifications proposed are: advance ETA notification (e.g. "arriving in 10 minutes"), arrival notification (when the vehicle is in the geofence) and collection reminder notification (just before the time for collection).

### E. Research Paradigm

Ontologically, the study treats reality as both objective and socially constructed: GPS coordinates, arrival-time error and notification-delivery speed are objective and measurable, while the frustration of a missed collection and the decision to dump waste in a canal are shaped by human behaviour and social context. Epistemologically, knowledge is generated both by measurement, through testing the system against real GPS and survey data, and through direct engagement with residents and collectors, which is consistent with the Pragmatist position adopted for this study.

### F. Architecture Diagram

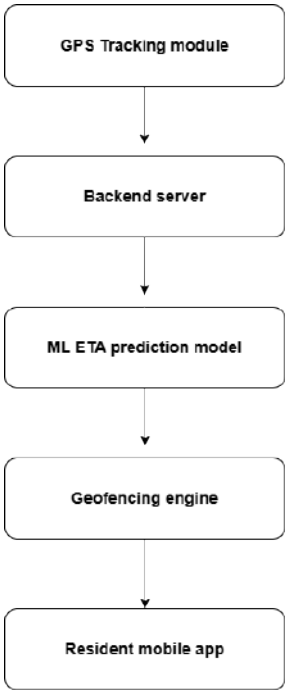



Fig. 2. Architecture diagram.

### G. Design Science Research Stages

The study follows the six DSR stages summarised in Table I, moving from problem identification on the streets of Kalubowila and Wellawatta through to communication of the findings to academic supervisors and local stakeholders.

### H. Project Management Approach

An Agile approach is used to manage development because the system is made up of several distinct components, the GPS tracking module, the Random Forest model, the geo-fence notification engine and the resident mobile application, that each need their own build-test-review cycle. Each sprint begins with a specific goal and ends with a reviewable increment, so that feedback from one component, for example a geo-fence trigger distance found to be too short during testing, can be fed into the next sprint rather than waiting until the end of the project.

*TABLE I*

*DESIGN SCIENCE RESEARCH STAGES APPLIED IN THIS STUDY*

| Stage | Activity in This Study |
|---|---|
| 1. Problem identification | Field observation and resident/collector conversations in the target areas |
| 2. Define objectives | Real-time visibility, timely notification, ML-based arrival prediction |
| 3. Design and development | Mobile app, backend, Random Forest model, geo-fence engine |
| 4. Demonstration | Deployment and trial runs in target streets |
| 5. Evaluation | GPS accuracy, notification speed, MAE/RMSE, usability testing |
| 6. Communication | This report and dissemination to supervisors and stakeholders |

### *I. Geo-Fence Design Considerations*

Shevchenko and Reips found that geofencing accuracy and notification sensitivity vary depending on the configured radius, smartphone model, and operating system [14]. This highlights the need to carefully select geo-fence radii to account for GPS inaccuracies while ensuring residents receive timely alerts as collection vehicles approach.

## V. DATA COLLECTION AND MODEL DEVELOPMENT

### *A. Field Data Collection*

Three methods were employed to gain insight into the problem from various points of view prior to development. A structured 20 question resident survey was sent out using Google Forms, to the residents of Kalubowila and Wellawatta, and included questions regarding their experience with collection, the frequency of missed collections, their attitudes towards real time tracking and notifications, and their comfort with technology. Informal interviews conducted with garbage collectors on the target routes to gain insight into the reasons why some streets are skipped and what a driver facing app would entail for their work. Maharagama Municipal Council also provided official collection schedules and route records to provide a foundation for the study in the framework of the official service. Also, continued discussion with the relevant people in the municipal council to support the development of the proposed system. This is the end of data collection phase.

### *B. Data Sources and Synthetic Trip Data Generation*

The dataset preparation process involved the use of actual waste collection route assignment information and synthetically generated vehicle trip records. The route assignment information provided by the relevant data source included truck identifiers, driver details, street names and collection schedules. These records were used to establish the route structure and operational context of the proposed system.

As real-time GPS trip logs containing actual vehicle movement and arrival-time records were not yet available, synthetic trip data were generated to simulate vehicle movements and travel durations along the selected collection routes. The generated dataset was structured to represent the input characteristics required for the Random Forest ETA prediction model.

The synthetic dataset was used to demonstrate the model development, training and testing workflow. Therefore, the experimental results represent preliminary model performance under simulated conditions rather than validated real-world vehicle arrival-time accuracy.

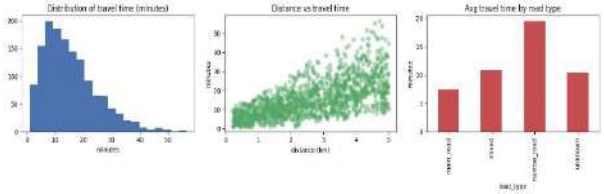



Fig. 3. Travel time visualizations.

### *C. Random Forest ETA Model*

A simple rule-based ETA estimate divides distance by average speed and ignores traffic, time-of-day variation and stop time, which typically produces large errors in complex urban environments. The Random Forest model instead builds multiple decision trees on random samples of the collected data and combines their outputs [8], allowing it to capture non-linear patterns such as narrow streets taking longer to service on weekday mornings. Once trained, the two approaches will be compared using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), as summarised in Table II.

*TABLE II*

*RULE-BASED VS. RANDOM FOREST ETA ESTIMATION*

| Criterion | Rule-Based Model | Random Forest Model |
|---|---|---|
| Handles non-linear patterns | No | Yes |
| Learns from historical data | No | Yes |
| Accounts for time-of-day variation | No | Yes |
| Suitable for small datasets (20–400 records) | N/A | Yes |
| Explainable to examiner | Yes | Yes |
| Expected accuracy (MAE) | High error (5–8 min) | Lower error (2–3 min, target) |

### *D. Data Preprocessing*

Raw GPS trip logs are cleaned before use: duplicate or stationary pings are removed, timestamps are aligned to a common time zone, and trips with clearly erroneous coordinates (for example, points falling outside the target service areas) are discarded. Distance to the target street is computed from consecutive GPS points, and each record is labelled with the time of day, day of week and street type described in Section IV-C.

### *E. Random Forest Model Training and Testing*

The Random Forest Regressor was implemented using Python and Scikit-learn, with the synthetic trip dataset divided into training and testing subsets. The model was trained using selected transportation and temporal features to predict vehicle travel duration. Its predictive performance was evaluated by

comparing predicted and simulated actual durations using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). This experiment provides a preliminary evaluation of the proposed ETA prediction approach, while real-world validation using GPS-based vehicle movement and arrival-time data remains necessary.

Fig. 4. Training random forest models using synthetic dataset.

## VI. System Implementation

The resident-facing mobile application allows a user to see the live position of collection vehicles on a map and receive a push notification when a vehicle enters the geo-fenced zone around their street. The following figures illustrate the current implementation and will be updated as development progresses.

### *A. System Workflow*

• The GPS device on the collection vehicle reports its location to the backend server at short, regular intervals.

• The backend server checks the vehicle's position against the geo-fence boundary defined for each street on its route.

• When the vehicle enters a resident's geo-fence, the backend triggers a push notification through Firebase Cloud Messaging.

• In parallel, the Random Forest model uses the vehicle's current speed, distance to the target street, time of day, day of week and street type to refine the estimated arrival time shown in the app.

• The resident's mobile application updates the vehicle's live position on the map and displays the notification together with the current estimated arrival time.

## VII. Evaluation Plan

### *A. Technical Evaluation*

Technical performance will be assessed through GPS tracking accuracy, notification delivery speed and the MAE and RMSE of the Random Forest arrival-time model compared against the rule-based baseline described in Section V-C.

### *B. User Evaluation*

Resident satisfaction and usability will be assessed through follow-up surveys and field usability testing with residents in Kalubowila, Wellawatta, Kottawa and Maharagama, focusing on whether real-time tracking and notifications reduce missed collections and improve awareness of collection timing.

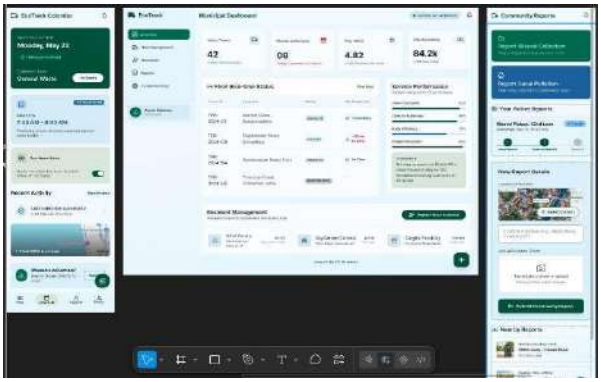


Fig. 5. Resident mobile application: live tracking and notification screens.

*TABLE III*

*ACTUAL USER EVALUATION RESULTS*

| Statement | Mean (1 to 5) | Agree or strongly agree |
|---|---|---|
| The app is easy to use | 4.35 | 90% |
| Knowing the expected arrival time would help me avoid missed collections | 4.48 | 92% |
| The arrival time display is easy to understand | 4.30 | 87.5% |
| Overall, I am satisfied with the app | 4.75 | 95% |

## VIII. Ethical Considerations

All participants are informed about the research and participate voluntarily, with the freedom to withdraw at any time. Personal information is securely stored on encrypted cloud servers, kept confidential, and used only for academic purposes. Official information obtained from the Maharagama Municipal Council is used with proper authorization and solely for research purposes.

## IX. Limitations and Assumptions

### *A. Limitations*

• Limited GPS data coverage and density may affect the accuracy of Random Forest ETA predictions, particularly for streets with insufficient data.

• The trained model may require further training and validation before being applied to other waste collection routes and areas.

### *B. Assumptions*

• Residents have smartphones with internet access to receive notifications.

• Collection vehicles have functioning GPS-enabled devices.

• Collection schedules and street routes remain relatively consistent during the study period.

## X. Conclusion and Future Work

This paper presents a Smart Urban Waste Collection System for Colombo, integrating a GPS-based vehicle tracking system,

Random Forest based ETA prediction, and geo-fencing push notifications to reduce missed waste collections. Future work includes model evaluation, real-time system deployment in target areas, and improving notification timing based on resident feedback.

## ACKNOWLEDGMENT

The author would like to thank the Maharagama municipal council and residents, boarding house owners, three-wheel drivers and garbage collectors of Kalubowila, Wellawatta, Kottawa and Maharagama who generously shared their time and experience for this study and the Faculty of Computing, NSBM Green University, for its guidance and support.

## REFERENCES

[1] M. Ghahramani, M. Zhou, and C. T. Hon, "Toward cloud-based smart waste management systems driven by IoT and AI for smart cities," *IEEE Access*, vol. 8, pp. 175966–175977, 2020.

[2] M. N. Neema and R. Gor, "Smart waste management system using IoT," *International Journal of Advanced Research in Computer Science*, vol. 12, no. 3, pp. 45–52, 2021.

[3] D. Rathod and A. Agarwal, "AR and IoT-based smart waste management system," *International Journal of Engineering Research and Technology*, vol. 10, no. 7, pp. 123–128, 2021.

[4] M. Chen, Q. Liu, X. Song, and C. Wang, "Ride-sharing services and real-time ETA notification systems: a user behaviour analysis," *Transportation Research Part C*, vol. 112, pp. 220–235, 2020.

[5] A. Sulistio and R. Ahmad, "Predictive waste collection using machine learning and IoT," *Journal of Cleaner Production*, vol. 284, pp. 124–135, 2021.

[6] Y. Zhang, R. Liu, and J. Xie, "Random forest-based travel time prediction for urban roads," *Transportation Research Record*, vol. 2674, no. 8, pp. 735–746, 2020.

[7] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Proc. 23rd ACM SIGSPATIAL Int. Conf.*, 2021, pp. 1–10.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] I. Sosunova and J. Porras, "IoT-enabled smart waste management systems for smart cities: a systematic review," *IEEE Access*, vol. 10, pp. 73326–73363, 2022.

[10] M. Ghahramani, M. Zhou, A. Molter, and F. Pilla, "IoT-based route recommendation for an intelligent waste management system," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 11883–11892, Jul. 2022.

[11] H. P. B. D. Hemapriya, D. S. Hewaranaweera, N. T. M. Sajith, and M. Aashiq, "IoT based smart waste bins for smart waste management in Sri Lanka," in *Proc. IEEE Conference*, Feb. 2025.

[12] T. Baddegama, H. Ariyasena, S. Wijethunga, M. Bowaththa, D. Nawinna, and B. Attanayake, "Solid-waste management system for urban Sri Lanka using IoT and machine learning," in *Proc. 2022 4th Int. Conf. Advancements in Computing (ICAC)*, Dec. 2022.

[13] R. L. S. Fernando, "Solid waste management of local governments in the Western Province of Sri Lanka: an implementation analysis," *Waste Management*, vol. 84, pp. 194–203, 2019.

[14] Y. Shevchenko and U.-D. Reips, "Geofencing in location-based behavioral research: methodology, challenges, and implementation," *Behavior Research Methods*, 2023.

[15] J. Brooke, "SUS: a quick and dirty usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

# Few Shot Gemstone Classification: Comparing Prototypical Networks Against Standard CNNs on Scarce Mineral Image Data

S. I. M. S. Wimalasiri
*Department of Computer Science and Data Science*
*NSBM Green University*
*Sri Lanka*
imswimalasiri@students.nsbm.ac.lk

YKA Rathnasiri
*Department of Computer Science and Data Science*
*NSBM Green University*
*Sri Lanka*
ykarathnasiri@students.nsbm.ac.lk

Thilini Bakmeedeniya
*Department of Computer Science and Data Science*
*NSBM Green University*
*Sri Lanka*
thilini.b@nsbm.ac.lk

***Abstract*—Gemstone image classification remains tough when labelled training data are scarce, particularly for visually comparable groups. This research studies Prototypical Networks for few-shot gemstone categorization and investigates their capacity to detect previously undiscovered types. Experiments utilize the Gemstones Images Expanded dataset, including virtually 4,400 photos spanning 88 gemstone categories. The dataset is partitioned into 60 establish classes for training and 28 unique classes for assessment. A typical Convolutional Neural Network (CNN) obtains 68.83% validation accuracy on the establish classes. The Prototypical Network is meta-trained utilizing a 5-way, 5-shot, 5-query episodic architecture with Euclidean distance for prototype-based categorization. Throughout 200 assessment sessions, it achieves 82.94% mean accuracy-with a claimed 95% confidence interval of ±1.32%, compared with the 20% random-chance baseline. These results highlight the potential of Prototypical Networks for recognizing recent gemstone categories from few tagged instances. Nevertheless, additional tests utilizing identical CNN-based few-shot baselines and numerous training sessions are essential to demonstrate comparative performance and generalizability.**

***Keywords—Gemstone Classification, Few-Shot Learning, Prototypical Networks, Convolutional Neural Networks, Metric Learning***

## I. Introduction

Gemstone classification is an essential undertaking in mineral identification and the gemstone business. However, discriminating between superficially similar gemstone groups may be problematic, especially when only a limited number of annotated photos are available. Conventional Convolutional Neural Networks (CNNs) may train excellent picture representations but often need adequate labelled input and are intended to identify predetermined categories. This inhibits their capacity to recognize new gemstone classifications without extra training. Few-shot learning (FSL) provides an alternative by allowing models to identify new categories using just a limited number of annotated samples. Prototypical Networks, developed by Snell et al. [1] , do this by building class prototypes from support-set embeddings and categorizing query pictures according to their distance from these prototypes. This research Prototypical Networks for gemstone image categorization utilizing the Gemstones pictures Expanded dataset, encompassing roughly 4,400 pictures over 88 categories.

Sixty categories are employed as base classes, while the remaining 28 are retained for novel-class assessment. The paper assesses the model using a 5-way, 5-shot episodic context and investigates its capacity for classification under low labelled data settings. It also emphasizes the necessity of identical assessment techniques when comparing few-shot methods with standard CNN-based systems..

## II. Literature Review

### *A. Deep Learning for Image Classification*

Deep learning has become an essential method to image classification because it allows models to automatically learn usable representations from picture data. Convolutional Neural Networks (CNNs) are frequently utilized because its convolutional layers understand hierarchical visual information, including edges, textures, colors and complicated patterns. These qualities make CNNs useful for applications requiring visually related categories and detailed picture properties. Deep learning has also been utilized to gemstone identification and mineral-related studies. GEMTELLIGENCE researched deep learning for gemstone classification utilizing gemological information [1] . CNN-based gemstone categorization was investigated in [2], while a machine-learning system for detecting gemstones from corundum, topaz and chrysoberyl families was reported in [3]. These works highlight the usefulness of computer approaches to gemstone identification and automated analysis. Deep learning has also been researched in larger geological applications. Research on mineral prospectivity mapping addressed the use of deep learning for evaluating geological information[4]. In addition, Query Plot examined the production of geological evidence layers using natural-language queries for mineral exploration [5] . Although these works do not directly address gemstone image categorization, they highlight the expanding use of intelligent computational tools in mineral-related research.

### *B. CNN-Based Image Classification*

CNN-based image classification normally comprises of a feature extraction network followed by a classification layer that assigns an input picture to one of the predetermined categories. As supervised training, the network learns discriminative representations from tagged pictures. But typical CNN classifiers are normally intended for closed-set classification-

where the output categories are confined to the classes represented as training.

Consequently, incorporating fresh gemstone categories may need extra labelled pictures and further model training. Transfer learning presents an alternative by leveraging pretrained CNN networks to extract reusable visual representations. These representations may aid classification when the target dataset comprises a limited number of tagged samples. Nevertheless, transfer-learning success relies on the quality of the retrieved features and their fit for the destination domain. Research on few-shot classification has demonstrated that feature representation quality may a bunch impact classification performance [6], [7]. See, in instance, simple classification algorithms employing useful feature embeddings may yield competitive results compared with extra complicated meta-learning systems [9]. These results display the need of analyzing the contribution of the feature extractor apart from the learning way. So, comparisons amid classic CNNs and few-shot techniques should employ identical assessment methodologies to guarantee that performance differences are understood quiet.

*C. Few-Shot Learning*

Few-Shot Learning (FSL) seeks to detect previously unknown categories using just a limited number of tagged samples. Unlike traditional supervised classification, which often needs adequate tagged pictures for every category, FSL aims to transfer information learnt from available classes to new categories. A typical few-shot classification framework separates a dataset into base classes and new classes. Base classes are utilized during model training, while novel classes are designated for testing the model's ability to identify categories that were not seen during training. Evaluation generally follows an episodic structure consisting of a support set and a query set. The support set provides a restricted number of labelled examples, whereas the query set contains pictures used to test classification performance.

Several strategies have been developed for few-shot categorization. Matching Networks employ attention processes to categories query samples based on tagged support examples [8]. Siamese Neural Networks build similarity associations between picture pairs for one-shot recognition [9]. Relation Networks develop connections between support and query representations to accomplish categorization [10]. Model-Agnostic Meta-Learning (MAML) learns model initializations that may be modified to different tasks using a limited number of instances [11]. Research surveys have evaluated few-shot and meta-learning strategies, including metric-based, optimization-based and model-based approaches [12], [13]. Few-shot learning has also been researched in specialist image-recognition areas. Research on remote-sensing scene identification has addressed feature improvement and classification with minimal labelled data[14], [15]. Meta-learning algorithms have also been used to ultrasound breast cancer picture categorisation [16]. These examples highlight the significance of few-shot approaches in sectors where gathering and annotating huge datasets might be problematic.

*D. Prototypical Networks*

Prototypical Networks were presented as a metric-based method to few-shot categorization [3]. The approach learns an embedding function that maps input pictures into a feature space. During each training episode, a prototype is calculated for each class by taking the mean of the embeddings of its labelled support samples. Query photos are then sorted according to their distances from the class prototypes.

The prototype for class (k) is defined as,

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\theta(x_i)$$

where (S_k) represents the support examples belonging to class (k) and (f_\theta) denotes the learnt embedding function. Classification is achieved by comparing the distance between a query embedding and each accessible class prototype. Euclidean distance is widely employed to locate the closest prototype [17]. Further research has studied enhancements to prototype-based categorization. ProtoDiff examines task-guided diffusion for learning Prototypical Networks [18]. Prototype Former analyses links between class prototypes utilizing transformer-based representations [19]. Proto-CLIP integrates vision-language representations with prototype-based few-shot categorization [20]. Other study has studied progressive multi-scale relation learning for few-shot picture categorization [20]. This research reflects continuous interest in refining prototype representations and interactions between classes. However, the efficacy of Prototypical Networks relies on the quality of the learnt embedding space and the representativeness of the support instances. Their performance should thus be tested against adequate few-shot baselines under comparable experimental settings. Research has demonstrated that well chosen feature extractors and baseline techniques may yield competitive few-shot performance [15], [16].

*E. Existing Research Gaps*

The studied literature reveals two related research areas deep learning for gemstone classification and few-shot learning for picture recognition under limited-data settings. Existing gemstone research have studied CNN-based and machine-learning algorithms to gemstone identification [1], [3]. Meanwhile, few-shot learning research has presented approaches for detecting unseen categories using a small number of tagged instances [4], [14]. Further application-specific study is required to examine the applicability of Prototypical Networks for gemstone image categorization. In particular, it is critical to evaluate if a model trained on basic gemstone categories can detect new categories using just a limited labelled support set. Fair experimental comparison is another key concern. A standard CNN trained as a 60-class closed-set classifier and a Prototypical Network assessed using a 5-way episodic approach execute distinct classification tasks. Their raw accuracy scores consequently cannot be compared directly to prove that one approach is better. A frozen CNN feature extractor paired with nearest-centroid classification delivers a more acceptable baseline when assessed using the identical support and query episodes [13], [17]. Accordingly, this research analyses Prototypical Networks utilizing the Gemstones Images Expanded dataset, which covers 88 gemstone types [6]. The dataset is partitioned into 60 base classes and 28 new classes. The Prototypical Network is examined utilizing a 5-way, 5-shot episodic setup to measure its

capacity to detect unique gemstone categories from few tagged samples. The work focuses on the use and assessment of metric-based few-shot learning for gemstone image classification, while noting the necessity for equal baselines and additional experimental validation.

## III. Methodology

### *A. Dataset Description*

The studies employ the Gemstones pictures Expanded dataset, which has 4,400 RGB pictures covering 88 gemstone classifications. Each category has roughly 50 photos. The dataset is organized into base and novel classes to examine the capacity of few-shot learning to detect previously unknown gemstone groups. Sixty classes are given to the base set, comprising roughly 3,000 photos, while the remaining 28 classes constitute the novel set, including around 1,400 images.

### *B. Data Preprocessing and Augmentation*

All input photos are reduced to 64 × 64 pixels to preserve a constant input dimension and decrease computing needs. Data augmentation is done during training to boost the variety of visual representations. The augmentation methods include random horizontal flips, rotations of up to 15 degrees and color jittering with brightness, contrast and saturation fluctuations of ±0.2. Images are transformed into tensors and normalized using ImageNet channel means of (0.485, 0.456, 0.406) and standard deviations of (0.229, 0.224, 0.225). A batch size of 16 is utilized for traditional CNN training. The evaluation transformation eliminates random augmentation.

### *C. Dataset Splitting (Base and Novel Classes)*

A class-level split is implemented to differentiate the training and evaluation categories. The 60 base classes are utilized for CNN training and Prototypical Network meta-training, whilst the 28 unique classes are saved for few-shot testing. For standard CNN training, the base-class pictures are separated into training and validation subsets using an 80:20 split, resulting in roughly 2,400 training images and 600 validation images. The innovative courses are omitted from the base-class training procedure. This separation permits the measurement of classification performance on categories not encountered during meta-training.

### *D. CNN Architecture*

A typical CNN is implemented as a closed-set classification baseline. Its architecture comprises of three convolutional blocks incorporating convolutional layers, batch normalization, ReLU activation and max-pooling operations. The final convolutional block provides 128 feature channels, followed by adaptive average pooling to a 4 x 4 feature map. The collected features are flattened and sent through a fully connected layer with 256 units, ReLU activation and dropout of 0.4. A final 60-unit output layer depicts the basic gemstone types. The network is trained using cross-entropy loss and the Adam optimizer with a learning rate of 0.001 and weight decay of 0.0001. A StepLR scheduler slows the learning rate every 10 epochs by a factor of 0.5. Training has a maximum limit of 50 epochs, with early terminating based on validation loss and a patience of seven epochs.

### *E. Prototypical Network Architecture*

The Prototypical Network employs a different three-block convolutional embedding architecture to translate gemstone pictures into feature representations. The network utilizes convolutional feature extraction followed by global average pooling to build 128-dimensional embeddings. Unlike the standard CNN, it does not employ a fixed 60-class classification output layer. Instead, classification is achieved by creating class prototypes from tagged support samples. The network is trained using the Adam optimizer with a learning rate of 0.001. A StepLR scheduler is applied every 200 episodes with a decay factor of 0.5.

### *F. Episodic Training Strategy*

Meta-training is done across 600 episodes utilizing a 5-way, 5-shot, 5-query design. In each episode, five classes are picked from the 60 foundation classes. Five tagged photos per class are allocated to the support set, while five extra images per class constitute the query set. Therefore, each episode has 25 support pictures and 25 inquiry photos. The network learns to discriminate between classes via recurrent episodic training. Validation is conducted every 100 episodes using 100 sampled episodes and the best-performing checkpoint is preserved. The implementation utilizes random seed 42 for Python, NumPy and PyTorch.

### *G. Prototype Calculation and Distance Measurement*

For each chosen class, a prototype is calculated by taking the mean of the 128-dimensional embeddings of its support pictures. The Euclidean distance between each query embedding and the five class prototypes is then determined. Negative distances are employed as classification logits, enabling the model to assign each query picture to its closest prototype. Cross-entropy loss is employed to optimize the model during episodic training.

### *H. Evaluation Metrics*

The Prototypical Network is assessed using 200 randomly picked 5-way, 5-shot test episodes from the 28 novel types. Classification accuracy is determined for each episode and the mean accuracy is presented over all test episodes. A 95% confidence interval is also computed to indicate variance across episode sampling. The random-chance baseline for a five-class classification test is 20%. For the standard CNN, training and validation accuracy and loss are utilized to assess closed-set classification performance. Since the CNN and Prototypical Network utilize distinct evaluation tasks, their initial accuracy results are understood individually rather than as a direct performance comparison.

## IV. Results and Discussion

### *A. CNN Training and Validation Results*

The standard CNN was trained using 60 basic gemstone classes, with roughly 2,400 pictures for training and 600 for validation. The model attained a final training accuracy of 68.00% and a maximum validation accuracy of 68.83% at epoch 29. The lowest measured validation loss was roughly 0.9947 at

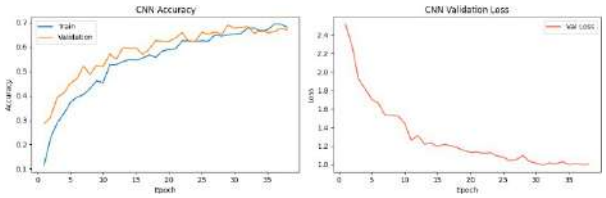



*Figure 1 CNN training performance*

epoch 31 and training terminated at epoch 38 by early stopping.

The findings imply that the CNN learnt effective visual representations for the basic gemstone classifications. However, random augmentation was also applied to the validation photos in the present implementation. Therefore, a deterministic validation experiment is required to achieve a more trustworthy evaluation.

### *B. Prototypical Network Results*

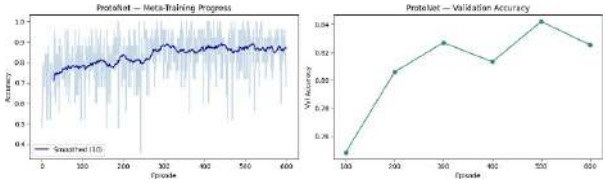



*Figure 2 Prototypical Network training progress and validation accuracy across 600 episodes*

The Prototypical Network was meta-trained for 600 episodes using a 5-way, 5-shot, 5-query setup. Each episode comprised 25 support pictures and 25 question images taken from five base classes. The model obtained 86.60% training accuracy throughout the last 100 episodes. Its greatest validation accuracy was 84.20% at episode 500 and the best checkpoint was picked for testing.

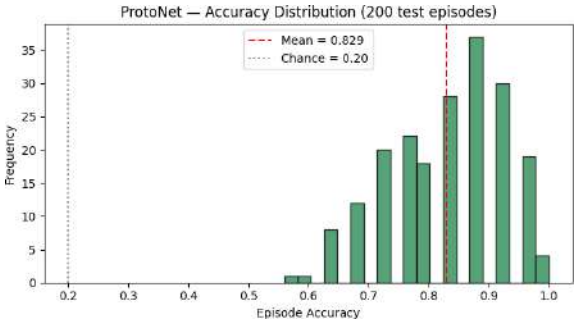



*Figure 3 Distribution of Prototypical Network accuracy across 200 novel-class test episodes*

These findings imply that the embedding network developed effective representations for prototype-based categorization. However, validation episodes were picked from the same base-class image pool used during meta-training, which reduces validation independence.

### *C. Novel-Class Classification Performance*

The trained Prototypical Network was assessed on the 28 unique gemstone classes using 200 randomly picked 5-way, 5-shot, 5-query episodes. It attained a mean accuracy of 82.94%, with a stated 95% confidence range of ±1.32%. The random-chance reference for five-way categorization was 20%. Most episodes attained accuracy above the random-chance reference, revealing the ability of the learnt embeddings to discriminate chosen unseen gemstone types. The confidence interval reflects episode-sampling variance for one trained model and does not include variability across separate training runs.

### *D. Experimental Comparison*

The CNN and Prototypical Network accomplish many classification jobs. The CNN recognizes 60 predefined base classes, whereas ProtoNet discovers five undiscovered classes using tagged support samples. Therefore, their accuracy figures cannot establish direct performance advantage. The reported 0.71% CNN novel-class accuracy is removed since the CNN output indices and novel-class label indices apply different mappings. A fair comparison entails removing the CNN classification head and performing nearest-centroid classification on the same support and query episodes used for ProtoNet. This experiment continues to be concluded.

*TABLE 1*
*SUMMARY OF EXPERIMENTAL RESULTS*

| Model | Accuracy | Evaluation |
|---|---|---|
| Standard CNN | 68.83% | Base-class validation |
| Prototypical Network | 82.94% ± 1.32% | 5-way, 5-shot novel classes |
| Random chance | 20.00% | Five-way reference |

### *E. Feature Visualisation and Discussion*

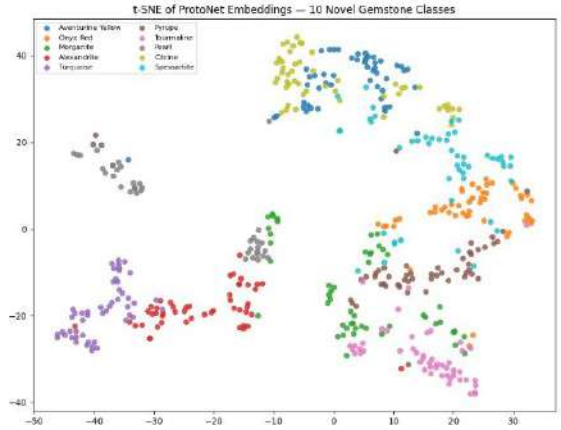



*Figure 4 t-SNE visualization of Prototypical Network embeddings for ten novel gemstone categories*

The learnt embeddings were visualised using t-distributed Stochastic Neighbour Embedding (t-SNE). Ten unique gemstone categories were projected onto a two-dimensional area to analyse their feature distributions.

The visualization displays identifiable clusters for certain categories, while others exhibit overlapping distributions. This gives qualitative evidence of the learnt feature representations but does not independently verify classification accuracy.

*TABLE 2*
*ADDITIONAL EXPERIMENTS REQUIRED*

| Experiment | Status |
|---|---|
| Frozen-CNN nearest-centroid baseline | Not performed |
| Multiple training seeds | Not performed |
| 1-shot and 10-shot evaluation | Not performed |
| Class-wise accuracy and confusion matrix | Not performed |
| Duplicate-image audit | Not verified |

Overall, the findings reveal that ProtoNet can detect chosen unique gemstone types above random chance. However, additional tests are need to verify its comparative efficacy and robustness.

## V. LIMITATIONS

### *A. Dataset Limitations*

The research employs a single dataset including around 4,400 photos spanning 88 gemstone categories. The restricted amount of photographs per category may restrict visual variation. Duplicate and near-duplicate photos have not yet been independently audited.

### *B. Limited Evaluation Configurations*

The assessment predominantly employs a 5-way, 5-shot, 5-query arrangement. Other scenarios, like 1-shot, 10-shot and higher-way categorization, have not been examined. Equivalent CNN-based few-shot baselines are likewise absent.

### *C. Model Generalisation Limitations*

The models employ relatively shallow convolutional architectures and analyses pictures at 64 × 64 pixels. Only one training run is presented, limiting the evaluation of random-seed variability and model resilience.

### *D. Lack of Cross-Dataset Evaluation*

The suggested technique has not been validated using independent gemstone datasets or photos acquired under diverse lighting and camera settings. Furthermore, gemstone authenticity is established using visual information without physical, chemical, or spectral verification.

## VI. CONCLUSION

This research studied the use of Prototypical Networks for gemstone picture classification under low labelled-data settings using the Gemstones Images Expanded dataset. The dataset comprised of 88 gemstone categories, which were separated into 60 base classes and 28 innovative classes. This class-level separation allowed an experimental scenario for investigating the capacity of a few-shot learning model to detect gemstone types that were not seen during meta-training.

A typical Convolutional Neural Network (CNN) was developed as a closed-set classification model utilizing the basic gemstone categories. In contrast, the Prototypical Network adopted an episodic learning technique with a 5-way, 5-shot, 5-query design. This strategy enables the model to create class prototypes from a minimal number of annotated support pictures and categories query images according to their distances from the associated prototypes. The experimental findings indicated that the Prototypical Network obtained a mean classification accuracy of 82.94%, with a stated 95% confidence interval of ±1.32%, over 200 novel-class test episodes. This performance exceeds the 20% random-chance reference associated with five-way categorization. The findings reveal that the learnt embedding representations provide useful visual information for identifying chosen new gemstone categories using just five annotated support pictures per class.

Furthermore, the CNN training and validation results indicated that traditional supervised learning may give appropriate classification representations for the chosen base categories. The t-SNE visualization gave further qualitative insight into the distribution of the learnt feature representations across 10 unique gemstone categories. However, certain overlapping feature distributions show that differentiating visually identical gemstone types remains problematic. The major contribution of this research is an application-oriented assessment of Prototypical Networks for gemstone picture categorization under low labelled-data settings. The work highlights the possibility of metric-based few-shot learning in identifying chosen unknown gemstone categories without needing a separate fixed classification output for every unique category.

Nevertheless, the experimental results should be interpreted within the restrictions of the present assessment. The CNN and Prototypical Network were examined under distinct categorization methodologies and consequently their stated accuracy ratings cannot show direct comparative superiority. Additionally, the trials were done using a single dataset and one reported training run. Consequently, additional work is necessary to verify the robustness, repeatability and generalization capabilities of the suggested technique.

## VII. FUTURE WORK

Future research might expand this study by refining the experimental assessment and examining the practical usefulness of Prototypical Networks for gemstone picture categorization. First, the suggested model should be tested using several independent training runs with distinct random seeds. Reporting the mean accuracy, standard deviation and confidence intervals over these runs would offer a more trustworthy evaluation of model stability and repeatability.

Furthermore, alternative few-shot learning configurations, including 1-shot, 10-shot and higher-way classification, should be examined to understand how the model performs under varied degrees of labelled data availability and classification complexity. A fair comparison with a frozen CNN feature extractor paired with a nearest-centroid classifier should also be undertaken utilizing similar support and query episodes. Other recognized few-shot learning systems, such as Matching Networks and Relation Networks, might be added to offer a wider performance comparison.

Another key trend is to examine more sophisticated feature extraction architectures, including pretrained ResNet and EfficientNet models. These structures may increase feature representation and assist identify visually related gemstone groups. In addition, class-wise accuracy, recall, F1-score and

confusion matrix analysis might give additional insights into the categories that are challenging to categories.

Future research should also examine broadening the assessment utilizing more gemstone image datasets acquired under diverse lighting settings, backdrops and picture gathering environments. This would assist evaluate the model's generalization capacity beyond the present dataset. Duplicate and near-duplicate image detection should also be incorporated to minimize potential data leakage.

Finally, the suggested technique might be evolved into a realistic gemstone identification application that delivers anticipated gemstone classifications and confidence levels. Integrating explainable AI strategies may further boost transparency and consumer confidence. These advances might enable the development of more dependable and adaptive gemstone categorization systems..

## REFERENCES


[1] J. Jaykumar, K. Palanisamy, Y.-W. Chao, X. Du, and Y. Xiang, "Proto-CLIP: Vision-Language Prototypical Network for Few-Shot Learning." [Online]. Available: https://irvlutd.github.io/Proto-CLIP

[2] Y. Du, Z. Xiao, S. Liao, and C. G. M. Snoek, "ProtoDiff: Learning to Learn Prototypical Networks by Task-Guided Diffusion."

[3] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking Generalization in Few-Shot Classification," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2206.07267

[4] J. Y. Lim, K. M. Lim, C. P. Lee, and Y. X. Tan, "A review of few-shot image classification: Approaches, datasets and research trends," Oct. 07, 2025, *Elsevier B.V.* doi: 10.1016/j.neucom.2025.130774.

[5] "Gemstone Identification using Image Classification," 2023.

[6] F. Sung Yongxin Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning."

[7] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition."

[8] J. Snell, K. Swersky, and T. R. Zemel, "Prototypical Networks for Few-shot Learning."

[9] W.-Y. Chen *et al.*, "A CLOSER LOOK AT FEW-SHOT CLASSIFICATION." [Online]. Available: https://github.com/wyharveychen/CloserLookFewShot

[10] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, P. Isola, and M. Csail, "Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need?" [Online]. Available: http://github.com/WangYueFt/

[11] E. Pachetti and S. Colantonio, "A systematic review of few-shot learning in medical imaging," Oct. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.artmed.2024.102949.

[12] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1606.04080

[13] M. Su, F. He, and F. Li, "Highlights PrototypeFormer: Learning to Explore Prototype Relationships for Few-shot Image Classi-fication PrototypeFormer: Learning to Explore Prototype Relationships for Few-shot Image Classification ⋆," 2025.

[14] B. H. Y. Chow and C. C. Reyes-Aldasoro, "Automatic gemstone classification using computer vision," *Minerals*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/min12010060.

[15] T. Bendinelli *et al.*, "GEMTELLIGENCE: Accelerating gemstone classification with deep learning," *Communications Engineering*, vol. 3, no. 1, Dec. 2024, doi: 10.1038/s44172-024-00252-x.

[16] X. Luo, H. Wu, J. Zhang, L. Gao, J. Xu, and J. Song, "A Closer Look at Few-shot Classification Again." [Online]. Available: https://github.com/

[17] J. Ren, C. Li, Y. An, W. Zhang, and C. Sun, "Few-Shot Fine-Grained Image Classification: A Comprehensive Review," Mar. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai5010020.

[18] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," 2017.

[19] Z. Wu, Y. Li, L. Guo, and K. Jia, "PARN: Position-Aware Relation Networks for Few-Shot Learning."

[20] G. Işık and İ. Paçal, "Few-shot classification of ultrasound breast cancer images using meta-learning algorithms," *Neural Comput. Appl.*, vol. 36, no. 20, pp. 12047–12059, Jul. 2024, doi: 10.1007/s00521-024-09767-y.

# Learning Collector Orchestration with Proximal Policy Optimization in a Multi-Agent Requirements Engineering Pipeline

Sujair Ibrahim
*Department of Software Engineering*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
ssujair@gmail.com

Vadivel Abishethvarman
*Department of Computing and Information Systems*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
abishethvarman@gmail.com

B. T. G. S. Kumara
*Department of Software Engineering*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
kumara@foc.sab.ac.lk

***Abstract*—Requirements Engineering (RE) is a sequential decision problem, yet existing multi-agent LLM pipelines such as MARE execute their constituent agents under a fixed action schedule with a hand-coded quality-gated iteration loop, with no mechanism to adapt strategy to the evolving state of the workspace. This static orchestration is a poor fit for RE, where the right next action depends on how much of a project has already been elicited, not on a fixed position in a pipeline. Proximal Policy Optimization (PPO) is a reinforcement learning algorithm that updates policies in small, stable steps, making it well suited to the noisy, expensive rewards typical of large language model driven agents. This paper isolates and evaluates the core orchestration mechanism of REMARL: a PPO policy trained to replace the fixed schedule for the Collector role, the agent responsible for turning stakeholder dialogue into formal requirement statements. The Collector observes a 1544-dimensional semantic embedding of the shared workspace and selects among four discrete actions, with immediate and Oracle-scored terminal rewards driving policy improvement over 50 training episodes. Evaluated on 22 paired synthetic scenarios held out from training, with results averaged over two independent seeds, against a faithful re-implementation of a fixed-schedule baseline, the PPOtrained Collector achieves a large improvement in overall oracle reward (Cohen's d = 1.109, p = 3.7 × 10−5) and a very large improvement in Oracle-measured requirement precision (Cohen's d = 2.119, 22 of 22 episodes won), while coverage remains statistically unchanged. Training diagnostics indicate stable, nondivergent optimisation. These results are an initial, controlled proof-of-concept that a learned orchestration policy can raise Oracle-measured precision without loss of coverage on synthetic scenarios, at a modest training cost; independent evaluation, heldout domains and additional baselines are left to future work.**

***Index Terms*—Large language models, Multi-agent systems, Proximal Policy Optimization, Reinforcement learning, Requirements engineering***

## I. INTRODUCTION

Requirements Engineering (RE) is among the most consequential phases of software development: requirements defects are well documented to cost substantially more to correct after deployment than during elicitation [1]. Recent multiagent LLM systems automate software engineering tasks by assigning specialized roles to separate LLM-backed agents that communicate through a shared workspace or dialogue history [2]. Within RE specifically, this pattern has been shown to produce requirements models competitive with prior automated approaches [3], but the framework that demonstrated this, along with its successors, executes agents under a fixed action schedule with a hand-coded quality-gated iteration loop, regardless of what the workspace actually contains. A project that is already well elicited and one that is barely started proceed through the identical sequence of agent calls. This is a poor fit for RE, where the right next action, whether to ask another clarifying question, draft a requirement, or flag missing coverage, depends on the current state of the work, not on a fixed position in a pipeline. This limitation is consistent with broader surveys of LLM based multi-agent systems for software engineering, which identify static orchestration as a recurring constraint across the field [4].

This paper addresses that gap for a single, well-scoped slice of the broader REMARL system: the Collector role, which converts stakeholder dialogue into formal requirement drafts. We replace its fixed schedule with a Proximal Policy Optimization (PPO) policy [5] that observes the workspace state and learns which of four candidate actions to take at each step. We deliberately scope this paper to the Collector- PPO mechanism in isolation; the remaining agent roles operate under a fixed policy throughout, so that any measured difference is attributable to the orchestration mechanism rather than to unrelated architectural changes.

The contributions of this paper are as follows. First, we formulate Collector-role requirements drafting as a Markov Decision Process compatible with standard policy-gradient reinforcement learning. Second, we introduce RESimEnv1, a Gymnasium-compatible [6] benchmark environment over synthetic RE scenarios used to train and evaluate this policy. Third, we report an empirical comparison, on 22 paired held-out scenarios and averaged over two training seeds, of the PPO-trained Collector against a faithful re-implementation of a fixed-schedule Collector, including effect sizes for total reward, precision, and coverage. We frame these results as an initial, controlled proof-of-concept rather than as evidence of generalisable improvement.

1Dataset available at https://doi.org/10.5281/zenodo.21952436

## II. RELATED WORK

Beyond requirements engineering specifically, several frameworks have explored multi-agent collaboration among

large language models for broader software development tasks. ChatDev simulates a virtual software company in which agents take on designer, programmer, and tester roles, communicating through structured dialogue to produce working code from a single natural-language idea [7]. MetaGPT encodes standardised operating procedures, borrowed from human software engineering practice, directly into agent prompts, and enforces structured intermediate outputs such as design documents and interface definitions to reduce hallucination during multi-agent collaboration [8]. Generative agents equipped with memory and reflection mechanisms have also been shown to produce believable, coordinated behaviour over long horizons in openended simulated environments [9]. These systems, together with general-purpose multi-agent conversation frameworks [2], establish that role specialisation and structured communication among LLM agents is an effective general pattern. REMARL adopts this pattern for requirements engineering specifically, but diverges from all of the above by making the sequencing of agent actions a learned decision rather than a hand-authored one.

Reinforcement learning has been applied to language models mainly for output alignment rather than for controlling multi-agent orchestration. Reinforcement learning from human feedback fine-tunes a language model's output distribution using PPO [5] against a learned reward model trained on human preference comparisons, and underlies most modern instruction-following LLMs [10]. This body of work, together with reliable open-source implementations of PPO such as Stable-Baselines3 [11], establishes PPO as a dependable choice for the noisy, expensive reward signals typical of LLMadjacent training loops, which is a central motivation for using it in this paper as well. The two applications differ substantially, however: RLHF trains the parameters of the language model itself, whereas the approach in this paper trains a separate, lightweight policy that decides which preexisting agent action to invoke next, leaving the underlying LLMs unchanged. This orchestration-level use of reinforcement learning is closer in spirit to classical sequential decisionmaking [12] than to LLM fine-tuning, and remains largely unexplored for multi-agent requirements engineering specifically.

Automating parts of the requirements engineering process predates large language models by decades. Classical natural language processing approaches to RE have addressed tasks such as requirement classification, ambiguity detection, and traceability recovery using rule-based and statistical methods; a systematic mapping study catalogs this line of work across more than one hundred studies and notes that most techniques target isolated sub-tasks rather than end-to-end elicitation and drafting [13]. Foundational surveys of the field identify the management of evolving, incomplete, and conflicting stakeholder input as a persistent open challenge [14], one that motivates the state-conditioned Collector policy evaluated in this paper. The IEEE 830 standard remains a common structural target for automatically generated software requirements specifications [15]. Relative to this classical NLP-for-RE literature, the work in this paper differs in scope, using LLM agents for open-ended drafting rather than closedset classification, and in mechanism, training an orchestration policy rather than a single-task classifier.

## III. BACKGROUND: PROXIMAL POLICY OPTIMIZATION

PPO is an on-policy, actor-critic reinforcement learning algorithm [12] that learns a stochastic policy π(a|s) and a value function V (s) through repeated interaction with an environment. Its central idea is a clipped surrogate objective that penalizes policy updates that move too far from the previous policy in a single step [5]:

$$L(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (1)$$

where rt(θ) is the probability ratio between the new and old policies, ˆAt is the estimated advantage, and ϵ is a clip range (0.20 in this work). This clipping is what makes PPO attractive for an environment like RESimEnv: each environment step triggers a real LLM call, so a single episode is expensive and slow. A policy-gradient method without PPO's trustregion-like clipping risks a destructive update that discards the value of costly, already-collected experience. PPO's stability under limited, expensive samples is the primary reason it was selected over vanilla policy gradient or value-based alternatives for this problem.

We use the Stable-Baselines3 implementation of PPO [11], with separate actor and critic networks, each a two-layer multilayer perceptron with 256 hidden units per layer. Full hyperparameters are given in Table I.

TABLE I
PPO TRAINING HYPERPARAMETERS

| Hyperparameter | Value |
|---|---|
| Actor / Critic architecture | MLP [256, 256] (separate) |
| Discount factor ($\gamma$) | 0.95 |
| GAE lambda ($\lambda$) | 0.95 |
| Clip range ($\epsilon$) | 0.20 |
| Training episodes per role | 50 |
| State dimensionality | 1544 |
| Action space | Discrete(4) |
| Embedding model | all-MiniLM-L6-v2 [16] |
| Coverage threshold (Oracle) | 0.65 cosine similarity |

All training and evaluation in this paper uses hosted large language models accessed through the NVIDIA Build API, replacing an earlier Ollama-based configuration used during initial development. The Collector policy is trained against meta/llama-3.2-3b-instruct, while the fixed-schedule baseline and terminal Oracle scoring use meta/llama-3.1-8b-instruct, so that evaluation and reward computation are not conditioned on the same model family as the trained agent. This separation reduces the risk that the policy learns to exploit idiosyncrasies of a single model's output distribution rather than genuinely improving requirement quality. The two Collectors therefore draft with different underlying models (llama-3.2-3b for PPO, llama-3.1-8b for the baseline); we discuss this difference in Sec. VII.

## IV. PROBLEM FORMULATION

REMARL retains a six-agent architecture, Stakeholder, Collector, Modeler, Checker, Negotiator, and Documenter, communicating through a single SharedWorkspace. In this paper, only the Collector's action selection is governed by a

learned policy; the remaining five roles operate under a fixed schedule for both the baseline and the PPO condition, isolating the Collector-orchestration effect.

The Collector's decision problem is formulated as an MDP over four discrete actions: propose question (elicit more information from the simulated stakeholder), write req draft (commit a new formal requirement), refine req draft (revise an existing requirement), and flag missing coverage (mark a gap for downstream agents). At each step, the current SharedWorkspace is encoded into a 1544-dimensional observation: four text fields (user stories, req draft, req model, error report) are each embedded with all-MiniLM-L6-v2 (384 dimensions each), concatenated with a 5-dimensional pipelinephase one-hot vector and 3 scalar progress indicators.

Reward is computed at two timescales. An immediate, perstep reward combines clarity, consistency, and coverage-delta signals (weights 0.30 / 0.40 / 0.30) to give the policy dense feedback without waiting for episode completion. A terminal reward, computed once per episode by the Oracle against ground-truth requirements at a 0.65 cosine-similarity threshold, combines coverage, precision, conflict, and NFR subscores (weights 0.50 / 0.20 / 0.15 / 0.15). Fig. 1 summarizes the full loop.

RESimEnv scenarios are fully synthetic. Requirements documents exist for real projects, but the complete elicitation flow, meaning stakeholder dialogue followed by requirement drafts, is not available in a form usable by a reinforcement learning environment, so we constructed scenarios that reproduce this flow. Scenario themes and stakeholder needs were drafted from ideas gathered in discussion with two business analysts and cover 14 domain templates. Each scenario has 11 to 13 written ground-truth requirements (mean 12, functional and non-functional combined), phrased in "shall" form, and each episode is limited to 18 steps because of resource constraints. Because the ground-truth phrasing may favour formal statements, Sec. VII discusses the possibility that the policy exploits surface form.

## V. TRAINING SETUP AND CONVERGENCE

Framing Collector orchestration as an MDP lets the system learn, from experience, when to keep questioning a stakeholder versus when to commit a requirement to the workspace, a context-sensitive trade-off that a fixed schedule cannot represent by construction. PPO is well suited to this setting for a concrete, non-generic reason: with roughly 30–35 seconds per LLM-backed environment step, each training episode is expensive, so the learner cannot afford the instability of unclipped policy-gradient updates discarding costly experience. The clip range of $\epsilon = 0.20$ keeps successive policies close, trading a slower learning rate for reliable, near-monotonic improvement. Whether the learned policy genuinely conditions on the workspace state, rather than simply writing requirements earlier, is not isolated by the present experiments (see Sec. VII).

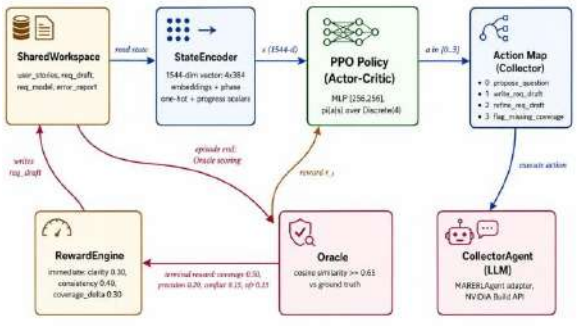



Fig. 1. Collector-role Markov Decision Process: state encoding, PPO action selection, LLM execution, and reward computation.

Table II reports PPO's internal training diagnostics after 50 episodes of Collector-role training. An explained variance of 0.399 is a meaningful result in this setting: standard RL benchmarks with deterministic, microsecond-scale steps typically reach explained variance above 0.8 quickly, whereas each REMARL step is a stochastic LLM call with substantially higher reward variance per state. A clip fraction of 7.1% and an approx kl of 0.0076 together indicate the policy updated steadily without divergence throughout training. These diagnostics measure optimisation stability only; they cannot by themselves rule out reward exploitation, that is, the policy learning the Oracle's preferred surface form (see Sec. VII).

TABLE II
PPO TRAINING DIAGNOSTICS AFTER 50 COLLECTOR EPISODES

| Metric | Value | Interpretation |
|---|---|---|
| explained_variance | 0.399 | Critic recovering reward structure (0=random, 1=perfect) |
| clip_fraction | 0.071 | 7.1% of updates clipped; healthy 5–15% range |
| approx_kl | 0.0076 | Near-zero KL; stable, non-divergent updates |
| policy_gradient_loss | −0.0119 | Actor improving in the correct direction |
| entropy_loss | −1.300 | Action diversity retained; exploration not collapsed |

## VI. RESULTS: COVERAGE AND PRECISION

We compare the PPO-trained Collector against a faithful re-implementation of MARE's per-phase action schedule operating over the same agent population and shared workspace, so that the comparison isolates the orchestration mechanism apartfromthedifferenceindraftingmodelnotedinSec. refsec : limits), across22pairedsyntheticscenariosdrawnfrom14syntheticdomains.The22evaluationscenarioswerestrictlydisjoi samplemeans, standarddeviations, effectsizes(Cohen′sd), andtwo− sidedt−testp−valuesfortotalreward, precision, andcoverage; per− episodewincountsarealsoreported.

TABLE III
COLLECTOR (PPO) VS. FIXED-SCHEDULE BASELINE, N = 22 PAIRED HELD-OUT SCENARIOS (MEAN OF TWO SEEDS)

| Metric | Baseline (mean ± SD) | PPO (mean ± SD) | Cohen's $d$ ($p$-value) | Wins |
|---|---|---|---|---|
| Total reward | 0.5735 ±0.0432 | 0.6300 ±0.0337 | 1.109 ($3.72 \times 10^{-5}$) | 19/22 |
| Precision | 0.1832 ±0.0974 | 0.4525 ±0.0752 | 2.119 ($2.15 \times 10^{-9}$) | 22/22 |
| Coverage | 0.8959 ±0.0135 | 0.8943 ±0.0198 | −0.068 (n.s.) (0.753) | 11/22 |

Precision is the standout result: an absolute improvement from 0.183 to 0.453 (Cohen's d = 2.119, p = 2.15 × 10−9), with the PPO Collector winning all 22 of 22 paired episodes. This gap is consistent with a qualitative pattern in the generated text: the PPO policy converges toward committing formal "shall"-style requirement statements early, which score well against ground-truth requirements, whereas the fixedschedule baseline spends more of its early steps in exploratory question-asking text that does not match the same formal ground truth.

TABLE IV
PER-SEED RESULTS (SEEDS 42 AND 44) AND MEAN ± SD ACROSS SEEDS

| Metric | System | Seed 42 | Seed 44 | Mean ± SD |
|---|---|---|---|---|
| Total reward | Baseline | 0.5712 | 0.5758 | 0.5735 ± 0.0033 |
| Total reward | PPO | 0.6351 | 0.6249 | 0.6300 ± 0.0072 |
| Precision | Baseline | 0.1807 | 0.1857 | 0.1832 ± 0.0035 |
| Precision | PPO | 0.4594 | 0.4456 | 0.4525 ± 0.0098 |
| Coverage | Baseline | 0.8984 | 0.8934 | 0.8959 ± 0.0035 |
| Coverage | PPO | 0.8917 | 0.8969 | 0.8943 ± 0.0037 |

This pattern is equally consistent with the policy having learned the Oracle's preferred surface form rather than producing higher-quality requirements, and the present evidence cannot distinguish the two explanations (Sec. VII). Total reward also improves with a large effect (d = 1.109, p = 3.7×10−5, 19/22 wins), driven primarily by the precision component of the terminal reward. PPO outperformed the baseline on total reward and precision in both seeds (Table IV): the total-reward gains were +0.064 and +0.049 and the precision gains +0.279 and +0.260 for seeds 42 and 44, respectively, while the coverage difference changed sign across seeds (−0.007 and +0.004), consistent with no coverage effect. With only two seeds we describe this as consistent across both seeds, not as an estimate of seed variance.

Coverage, by contrast, is statistically unchanged (d = −0.068, p = 0.753, 11/22 wins), the PPO Collector maintains, rather than improves, elicitation coverage relative to the baseline. We report this as a genuine null result rather than omit it: at 50 training episodes, the policy appears to have converged toward a write-heavy strategy that trades a marginal, nonsignificant coverage difference for a large precision gain, and has not yet learned to selectively favor elicitation in scenarios where it would raise coverage further.

From a practitioner's perspective, the size of the precision effect is arguably more actionable than its statistical significance alone. A precision improvement from 0.183 to 0.453 means that the share of generated statements with no groundtruth match falls from about 82% to about 55%. Whether this reduces the effort of a human reviewer has not been measured. Total training cost for this result was modest: 50 episodes of Collector-role training, at roughly 30 to 35 seconds per LLM-backed environment step, corresponds to a few hours of wall-clock time on commodity hardware, well within reach of a small research group rather than requiring large-scale compute infrastructure. This cost profile suggests that learned orchestration policies of this kind are a practical addition to existing multi-agent RE pipelines, rather than a purely academic exercise reserved for well-resourced teams.

## VII. LIMITATIONS AND THREATS TO VALIDITY

Nine limitations bound the claims of this paper. First, the Negotiator agent is integrated into the framework's action space; empirical evaluation under single-role training does not exercise its action set, so no claim is made here about conflict-resolution quality. Second, the baseline is a faithful re-implementation of MARE's per-phase action schedule operating over the same agent population and shared workspace, not the original published MARE system on its original LLM backend; the comparison therefore isolates the orchestration mechanism rather than establishing an absolute comparison against published MARE results. Third, the same

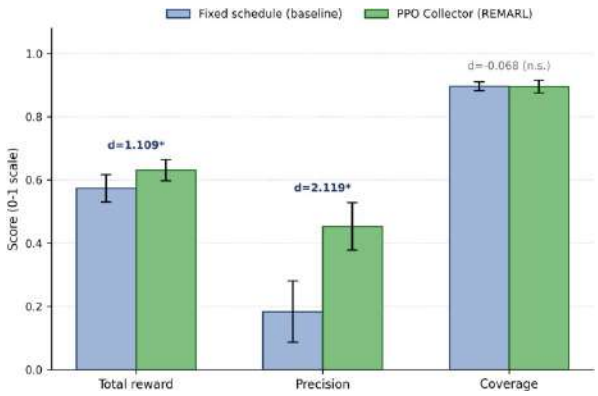



Fig. 2. Total reward, precision, and coverage for the fixed-schedule baseline versus the PPO Collector, n = 22 paired scenarios. Error bars show ±1 SD. Values are means over two seeds.

Oracle model that scores terminal reward during training also scores evaluation, so coverage and precision numbers reflect an oracle's dual role in training and evaluation rather than an independent human-judged ground truth; no independent evaluator or human rating was applied. Fourth, all 22 evaluation scenarios are drawn from a synthetic benchmark scope of 14 domain templates; generalization to real, messy client requirements documents remains untested in this paper. Fifth, results are averaged over only two training seeds (42 and 44), which is too few to estimate seed variance reliably. Sixth, although the 22 evaluation scenarios were strictly held out from training, all scenarios come from the same synthetic benchmark of 14 domain templates, so generalisation to unseen domains is untested, and we did not record paired train and test curves over the course of training. Seventh, we did not evaluate simple heuristic baselines such as an alwayswrite policy, or a state-blind policy, so the gain cannot yet be attributed to state-conditioned learning rather than to writing requirements earlier. Eighth, the precision gain may partly reflect the policy learning the Oracle's preferred surface form (early, formal "shall"-style statements) rather than higherquality requirements; we did not

test this with an independent evaluator or human expert rating, and training diagnostics cannot rule it out. Ninth, the baseline Collector drafts with llama-3.1-8b whereas the PPO Collector uses llama-3.2-3b, so the comparison is not purely a comparison of orchestration; we did not run a baseline with the 3b model.

## VIII. CONCLUSION

Replacing a fixed action schedule with a PPO-trained policy for the Collector role produces a large improvement in Oraclemeasured requirement precision and overall oracle reward on held-out synthetic scenarios, consistent across two training seeds, while preserving coverage, using only 50 episodes of training. We present these results as an initial, controlled proof-of-concept. PPO's clipped update rule is well matched to the expensive, high-variance nature of LLM-backed RL environments, and its training diagnostics indicate stable, nondivergent optimisation, although they cannot rule out reward exploitation.

Future work follows three steps. First, train the Modeler and Checker roles the same way, since they use the same actionspace design as the Collector. Second, combine all roles into joint training; an early pilot showed the policy's exploration collapsing too early when trained jointly, so this issue needs to be fixed first. Third, test the trained policy on a different LLM family (mistralai/mistral-7b-instruct-v0.3) and on heldout domains, to confirm the results generalise beyond the current setup. Beyond these steps, the most important next work addresses the limitations above: (i) evaluation on heldout domains with 3–5 independent seeds and paired train and test curves; (ii) always-write, write-first and random heuristic baselines, a state-blind PPO ablation, and the per-phase action distribution; and (iii) an independent LLM judge from a separate model family together with blind human expert rating of a subset of requirements.

## DATA AVAILABILITY

The RESimEnv scenario corpus used for training and evaluation in this paper is openly available on Zenodo [17].

## REFERENCES

[1] B. W. Boehm and P. N. Papaccio, "Understanding and controlling software costs," IEEE Trans. Softw. Eng., vol. 14, no. 10, pp. 1462– 1477, Oct. 1988.

[2] Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework," arXiv preprint arXiv:2308.08155, 2023.

[3] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang, "MARE: Multi-agents collaboration framework for requirements engineering," arXiv preprint arXiv:2405.03256, 2024.

[4] J. He, C. Treude, and D. Lo, "LLM-based multi-agent systems for software engineering: Literature review, vision, and the road ahead," ACM Trans. Softw. Eng. Methodol., 2025.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[6] M. Towers et al., "Gymnasium: A standard interface for reinforcement learning environments," arXiv preprint arXiv:2407.17032, 2024.

[7] C. Qian et al., "ChatDev: Communicative agents for software development," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics, 2024, pp. 15174–15186.

[8] S. Hong et al., "MetaGPT: Meta programming for a multi-agent collaborative framework," in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[9] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in Proc. 36th Annu. ACM Symp. User Interface Software and Technology (UIST), 2023, pp. 1–22.

[10] L. Ouyang et al., "Training language models to follow instructions with human feedback," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, 2022, pp. 27730–27744.

[11] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," J. Mach. Learn. Res., vol. 22, no. 268, pp. 1–8, 2021.

[12] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA: MIT Press, 2018.

[13] L. Zhao et al., "Natural language processing for requirements engineering: A systematic mapping study," ACM Comput. Surv., vol. 54, no. 3, pp. 1–41, 2021.

[14] B. H. C. Cheng and J. M. Atlee, "Research directions in requirements engineering," in Proc. Future of Software Engineering (FOSE), 2007, pp. 285–303.

[15] IEEE Computer Society, "IEEE recommended practice for software requirements specifications," IEEE Std 830-1998, 1998.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.

[17] S. Ibrahim, V. Abishethvarman, and B. T. G. S. Kumara, "RESimEnv," Zenodo, Aug. 15, 2026. doi: 10.5281/zenodo.21952436.

# SolarisPredict-SL: A Weather-Driven Machine Learning Framework for Retrospective Net-Load Proxy Modelling and Indicative Dispatch Support in the Sri Lankan Power Grid

Geemal Muthugala
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
geemalmuthugala@gmail.com

***Abstract*—Sri Lanka's rapid solar photovoltaic expansion increases the importance of accurate generation-mix reconstruction and grid observability. SolarisPredict-SL combines 181 days of 15-minute National System Operator (NSO) records with historical weather features, gradient-boosted models and an indicative dispatch advisory. The study distinguishes the reported generation-mix total G, its observable non-SCADA-solar residual R, the NSO photovoltaic estimate P and the constructed index Q = G − P; Q is not independently validated physical net load. After timestamp-aligned lag reconstruction, a chronological 21-day retrospective holdout (1,974 eligible records per target) produced XGBoost mean absolute errors of 49.79, 46.89, 23.57 and 57.11 MW for G, R, P and Q, respectively. Weather/calendar-only ablations yielded 105.56, 108.71, 52.85 and 141.10 MW. Against paired 15-minute persistence, G was worse (50.08 versus 46.89 MW), R improved only marginally (47.22 versus 48.76 MW), and P and Q improved (23.81 versus 31.33 MW and 57.50 versus 75.35 MW). These results measure retrospective reconstruction with archived target-time weather, not performance of operationally issued forecasts. A descriptive SCADA-to-NSO solar-estimate ratio and an illustrative dispatch-advisory prototype complement the evaluation; the public software release excludes third-party observation-level records.**

***Index Terms*—net-load proxy, solar forecasting, grid operations, gradient boosting, Sri Lanka**

## I. INTRODUCTION

Sri Lanka has pursued an accelerated transition toward renewable electricity generation, with renewable sources accounting for a substantial portion of installed capacity by the mid-2020s and rooftop solar photovoltaic capacity growing rapidly [1]. This growth has increased the importance of generation visibility and forecasting infrastructure available to the National System Operator (NSO). Reported solar-curtailment practices expanded beyond weekends amid operational concerns over midday generation conditions and evening ramp requirements [2]. The present study does not establish a causal explanation for any particular grid disturbance.

Three strands of prior Sri Lankan research are directly relevant but do not, individually, close this operational gap. Amarasinghe and Abeygunawardane applied machine learning algorithms to forecast solar power output at the Buruthakanda solar park, showing that tree-based ensemble methods outperform a smart-persistence baseline [3]. Abeysingha et al. applied deep learning techniques to short- and long-term electricity demand forecasting using NSO demand data, finding that calendar effects—including Poya (full-moon) holidays and the Sinhalese and Tamil New Year period—are important predictors [4]. Premathilaka et al. estimated floating photovoltaic solar potential across Sri Lankan reservoirs using satellite-derived irradiance and time-series models [5]. Forecasting, however, is only one component of curtailment management, alongside plant flexibility, transmission constraints, and grid stability; the present study therefore avoids attributing curtailment solely to insufficient forecasting.

The contribution of this paper is threefold. First, we document integration of fifteen-minute NSO generation-mix and photovoltaic-estimate records with weather and calendar covariates, including an explicit accounting boundary for grid-relevant residuals. Second, we evaluate timestamp-aligned XGBoost retrospective reconstruction for G, R, P and the constructed index Q, with paired persistence comparisons at 15-minute, one-hour, daily and weekly offsets, and a weather/calendar-only ablation. Third, we describe an indicative merit-order dispatch-advisory layer and a descriptive SCADA-to-NSO solar-estimate ratio. These are engineering and analytical contributions, not evidence of physically validated net-load forecasts or operational dispatch optimization.

The remainder of this paper is organized as follows. Section II reviews related work; Section III describes the system architecture; Section IV presents the forecasting methodology; Section V describes the dataset; Section VI reports results and an operating-day case study; Section VII discusses limitations; and Section VIII concludes.

## II. RELATED WORK

### *A. Solar power forecasting*

Solar power forecasting methods range from persistence and smart-persistence baselines through numerical-weather-prediction-driven statistical models to machine-learning approaches [3]. For the Sri Lankan context specifically,

Amarasinghe and Abeygunawardane compared deep belief networks, support-vector regression, and random-forest regression against a smart-persistence baseline at the Buruthakanda solar park, finding that a tree-based ensemble achieved the lowest root-mean-square error, particularly under partly cloudy conditions where forecast variability is highest [3]. This finding motivates the use of gradient-boosted trees in the present work. Gradient boosting has been substantially advanced through XGBoost [6] and LightGBM [7], which complement the original random-forest formulation [8] for tabular, feature-engineered time series. Prophet's decomposable time-series model [9] and classical seasonal time-series methods have also been applied to solar-resource analysis at reservoir scale [5].

### *B. Electricity demand and net-load forecasting*

Electricity demand forecasting is conventionally divided into short-, medium-, and long-term horizons [4]. Deep recurrent architectures, including long short-term memory (LSTM) networks [10], have been shown to outperform classical regression for Sri Lankan demand forecasting, with calendar features—day type, Poya holidays and the New Year period—found to be informative in a tropical climate with relatively limited seasonal temperature variation [4]. More recent architectures such as the Temporal Fusion Transformer [11] extend multi-horizon forecasting through attention over static, known-future and observed inputs, while probabilistic forecasting provides prediction intervals rather than point estimates [12]. Weron provides a broader statistical treatment of electricity load and price forecasting [13]. A fair comparison must also specify forecast origin, horizon and accessible covariates [12], [14].

### *C. Renewable integration and curtailment*

The operational challenge addressed in this paper is closely related to the "duck curve" phenomenon documented for the California grid, in which rising solar penetration produces a midday residual-load dip followed by a fast evening ramp, creating operational challenges where flexibility is insufficient [15]. Denholm et al. discuss options for mitigating overgeneration through flexibility and integration measures [15]. For Sri Lanka, a comparable net-load framing is potentially useful, but a physical residual-demand calculation requires a consistent metering boundary and verified solar accounting. The NSO generation-mix sum in this study is not a separately metered gross-demand series; the present experimental proxy is therefore distinguished from validated physical net load.

## III. SYSTEM ARCHITECTURE

Fig. 1 shows the system architecture. The Open-Meteo Forecast and Archive APIs provide prospective and historical weather inputs, respectively [16]. For the retrospective national evaluation, hourly archived weather is interpolated to 15-minute timestamps. The NASA POWER archive provides satellite/reanalysis-derived zonal irradiance climatology, not independently measured PV generation [17]. Historical generation-mix and solar-estimate records are drawn from the publicly accessible NSO generation-summary interface [18]. These sources form a feature store supporting prototype zonal solar models and national generation-mix, solar-estimate and constructed-proxy models. Model outputs feed the deterministic merit-order advisory and web dashboard; no connection to the NSO energy-management system is implied.

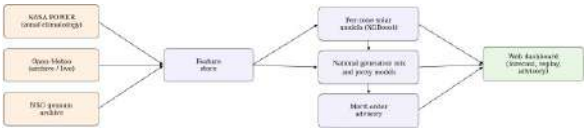


Fig. 1. SolarisPredict-SL architecture. Archived and live weather inputs are distinguished; NSO source records support national generation-mix and constructed-proxy modelling, while the advisory feeds a prototype dashboard.

## IV. METHODOLOGY

### *A. Problem formulation*

Let $G_t$ denote the sum of NSO-reported load-curve generation sources, including the SCADA-listed solar component $S_t^{sc}$. Let $P_t$ be NSO pvEstimate, an estimate of system photovoltaic output rather than an independent power-meter measurement. The implemented demand_mw field is $G_t$ (generation mix), not a separately metered gross load $D_t$. Thus the original relation $N_t = D_t - S_t$ is a conceptual physical balance only; it is not established by substituting the available NSO fields.

$$R(t) = G(t) - Ssc(t); \quad Q(t) = G(t) - P(t) \qquad (1)$$

$R_t$ is an observable non-SCADA-solar generation-mix residual, whereas $Q_t$ is the target named net_load_mw in the implementation. These differ from independently established physical net load because $G_t$ already includes SCADA-listed solar and may reflect unmetered behind-the-meter solar through changes in grid-served load. Subtracting $P_t$ from $G_t$ can therefore double-count photovoltaic contribution. We retain $Q_t$ only as an explicitly constructed index and do not interpret its negative values as validated negative demand or dispatch requirement.

$$\hat{y}(t) = f[x(t)] \qquad (2)$$

The reported holdout estimates targets at timestamp t using archived weather at timestamp t and earlier observations; it does not evaluate predictions genuinely issued h hours in advance. Accordingly, reported accuracy is termed retrospective conditional reconstruction, while operational lead-time performance remains unverified.

### *B. Solar forecasting model*

Following [3], we employ gradient-boosted regression [6]. The national model fits the NSO pvEstimate target in MW, using hour/day encodings, cloud, temperature, humidity, solar irradiance (W/m²), wind, and lagged NSO-estimate/SCADA signals. The implementation excludes the contemporaneous pvEstimate from its live feature set, preventing direct target copying. The separate prototype zonal irradiance model estimates all-sky irradiance (W/m²); panel area and efficiency yield indicative power as

$$Psolar = A \cdot G_eff \cdot \eta_eff \qquad (3)$$

where A is panel area (m²), $G_eff$ is effective irradiance (W/m²), and $\eta_eff$ is dimensionless efficiency. Consequently Psolar has units W and requires conversion to MW for power-system reporting. NASA POWER irradiance is not

independently measured photovoltaic plant output [17]. The reported national solar-model MAE (MW) and zonal irradiance MAE (W/m²) evaluate different targets and must not be combined.

*C. Demand and net-load forecasting model*

Calendar covariates include hour, weekday, month, Poya-day and holiday indicators with cyclical encodings. XGBoost uses 400 trees, maximum depth 6, learning rate 0.05, row/column subsampling of 0.85, minimum child weight 4 and random seed 42. Autoregressive features are generated after reindexing the source series to a full local-time 15-minute grid: lag offsets of four, 96 and 672 slots therefore correspond to exactly one hour, one day and seven days; unavailable observations remain missing rather than shifting adjacent calendar times. For the chronological 20 July–9 August 2026 holdout, the corrected feature filters yield 14,382 training rows for G and 14,946 for each of R, P and Q, with 1,974 test rows per target. A separate weather/calendar-only XGBoost ablation, omitting target lag and rolling-load features, is evaluated on the same test period.

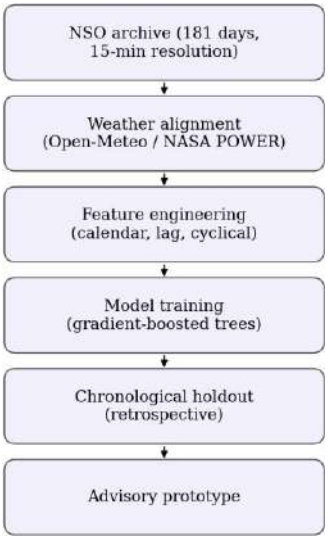



Fig. 2. Data integration and model-training workflow. The chronological holdout tests retrospective conditional reconstruction; it is not a verification of forecast-issue-time accuracy.

*D. Baselines*

Persistence is evaluated against the identical eligible target timestamps as each model, using exactly 15-minute, one-hour, 24-hour and seven-day clock-time offsets rather than row-position shifts. The 15-minute comparator retains 1,953 paired rows; other comparisons retain 1,974 paired rows. Each reported gain is the persistence mean absolute error (MAE) minus model MAE recalculated on the same paired rows; positive values favour XGBoost. The 15-minute reference uses a more recent observation than the one-hour lag available in the fitted model and is therefore a demanding but not forecast-origin-equivalent benchmark. Historical weather values remain retrospectively aligned with target timestamps, so none of these comparisons establishes issued-forecast skill.

$$\hat{y}(t;\,24\text{ h}) = y(t - 24\text{ h}) \qquad (4)$$

*E. Evaluation metrics*

Evaluation summarizes mean absolute error (MAE) and root mean square error (RMSE), both in MW for national power-series targets. Percent-based errors are not emphasized because irradiance and solar series approach zero at night.

$$\text{MAE} = (1/n)\ \Sigma_i\ |y_i - \hat{y}_i| \qquad (5)$$

$$\text{RMSE} = \sqrt{[(1/n)\ \Sigma_i\ (y_i - \hat{y}_i)^2]} \qquad (6)$$

*F. SCADA-to-NSO solar-estimate ratio*

A descriptive SCADA-to-NSO solar-estimate ratio is computed from the NSO generation-mix solar component and the separately published NSO pvEstimate series over intervals in which $P_t$ exceeds a selected threshold:

$$C(\theta) = 100 \times [\Sigma\ \text{Ssc}(t)] / [\Sigma\ P(t)],\ \ P(t) > \theta \qquad (7)$$

Both sums in (7) use identical timestamps satisfying $P(t) > \theta$. Since observations represent equal-duration 15-minute intervals, the common time multiplier cancels. $C(\theta)$ describes a ratio of two NSO-reported series, not a directly measured fraction of rooftop photovoltaic output visible to grid operators. The holdout sensitivity calculation considers thresholds of 25, 50, 75, 100 and 200 MW; the archive-wide 50-MW result is reported separately.

## V. DATASET

The operational archive covers 10 February–9 August 2026 (181 days). NSO generation-mix components comprise wind, SPP biomass, thermal oil, major hydro, coal, SCADA-listed solar and SPP mini-hydro. The solar target P is the separately published NSO pvEstimate (P50), not an independently measured national PV-output series. Table I summarizes the archive. The generation curve contains 17,195 observations and omits 181 midnight 15-minute slots over the period; the 96-slot daily indexing and missing-value treatment avoid mistaking row shifts for elapsed time. The NSO solar-estimate series contains 17,376 observations. Model feature completeness yields 1,974 eligible holdout rows for each target. National historical weather uses a Sri Lankan geographic proxy (7.8731° N, 80.7718° E) and hourly Open-Meteo Archive values interpolated to local quarter-hours; separate NASA POWER irradiance data support zonal climatology. The evaluated weather inputs are historical target-time values rather than forecasts archived at issuance [16], [17].

TABLE I
OPERATIONAL DATASET SUMMARY

| Record type | Rows | Interval |
|---|---|---|
| Generation (by source) | 17,195 | 15 min |
| Solar estimate (NSO P50) | 17,376 | 15 min |
| Energy summary (daily) | 1,980 | daily |
| Peak / minimum demand | 542 | daily |
| Reservoir status | 2,715 | daily |
| Night-peak generation mix | 2,172 | daily |

*A. Data provenance, access and availability*

The study used NSO/EDL generation-summary observations obtained through the publicly accessible source interface [18]. The accompanying software release [19] supplies

retrieval and normalization code, feature definitions and model configurations; the present manuscript reports aggregate evaluation outcomes without distributing the third-party raw JSON responses, processed NSO observation-level tables or merged training observations. The author reports removing these record-level files from the publication branch. No written authorization to redistribute underlying NSO/EDL observations is asserted. Independent researchers must acquire any source records separately, subject to the provider's applicable permissions and conditions.

## VI. RESULTS

### *A. Forecast accuracy*

Tables II and III present the timestamp-corrected chronological holdout for all four evaluated targets. The holdout extends from 20 July to 9 August 2026, with 1,974 eligible instances per model. Table II reports XGBoost MAE, RMSE and the no-lag weather/calendar-only ablation; Table III reports persistence MAE on paired eligible timestamps. The corrected models reconstruct the NSO source-sum G, the non-SCADA-solar residual R, the NSO estimated PV series P and the constructed index Q, respectively; none is presented as measured gross electricity demand or independently metered physical net load.

TABLE II
TIMESTAMP-CORRECTED XGBOOST HOLDOUT AND NO-LAG ABLATION (MW)

| Target | Training rows | XGB MAE (RMSE) | No-lag MAE |
|---|---|---|---|
| G | 14,382 | 49.79 (72.80) | 105.56 |
| R | 14,946 | 46.89 (72.30) | 108.71 |
| P | 14,946 | 23.57 (45.14) | 52.85 |
| Q | 14,946 | 57.11 (85.69) | 141.10 |

TABLE III
PAIRED PERSISTENCE MAE (MW; N = 1,953 FOR 15 MIN; N = 1,974 OTHERWISE)

| Target | 15 min | 1 hour | 24 hours | 7 days |
|---|---|---|---|---|
| G | 46.89 | 167.85 | 182.41 | 132.96 |
| R | 48.76 | 177.18 | 172.77 | 137.78 |
| P | 31.33 | 120.06 | 56.34 | 88.60 |
| Q | 75.35 | 283.13 | 216.28 | 200.65 |

The corrected models improve materially over weather/calendar-only ablations: for G, R, P and Q, ablation MAEs are 105.56, 108.71, 52.85 and 141.10 MW, respectively. Nevertheless, improvements are not uniform against recent-observation persistence. For G, 15-minute persistence is more accurate than XGBoost on paired timestamps (46.89 versus 50.08 MW; gain −3.19 MW, reported 95% CI [−14.10, 6.33]). For R, the corresponding gain is only 1.53 MW (95% CI [−9.97, 11.99]). P improves over 15-minute persistence by 7.52 MW (95% CI [0.39, 13.81]), while Q improves by 17.84 MW (95% CI [5.36, 28.96]); the latter remains a constructed, non-physical target. Against one-hour persistence, the paired gains are 118.05 MW for G, 130.29 MW for R, 96.49 MW for P and 226.02 MW for Q. These are retrospective error comparisons under archived target-time weather, not validated 1-hour-ahead forecasts. The NSO P50 record is the model target, so treating it as an independent perfect-reference forecast would be circular. A separate zonal Hambantota irradiance evaluation yielded 15.02 W/m² MAE and 35.04 W/m² RMSE against NASA POWER-derived irradiance; those units cannot be compared directly with national MW errors.

### *B. SCADA-to-NSO solar-estimate ratio*

The saved full-period training summary reports C(50) = 12.03%; the distinct 21-day holdout gives 11.15%, 11.13%, 11.10%, 11.09% and 10.99% for thresholds of 25, 50, 75, 100 and 200 MW, respectively (950, 933, 913, 902 and 841 paired intervals). These different values relate to different observation windows and need not agree. The stability across thresholds is descriptive, not an independent measurement of physically visible rooftop PV. Installed capacity [1] provides contextual scale rather than interval-metered validation.

### *C. Case study: operating-day reconstruction*

Fig. 3 reconstructs the 2 August 2026 operating day using hourly observations at :15 sampled from the 15-minute source records. It plots generation-mix total G, NSO photovoltaic estimate P, constructed index Q = G − P and observable generation residual R = G − SCADA solar. G reaches approximately 2,453 MW at 19:15 and P approaches 1,240 MW around midday. Q becomes negative for some daylight observations, illustrating why it cannot be interpreted as validated physical system net load. R increases from approximately 823.46 MW at 13:15 to 2,453.00 MW at 19:15: a change of 1,629.55 MW over six hours, equivalent to an interval-average slope of 271.59 MW/h. This figure does not report the maximum rolling one-hour or three-hour ramp. The series can reflect actual operator curtailment, for which interval-specific flags were unavailable.

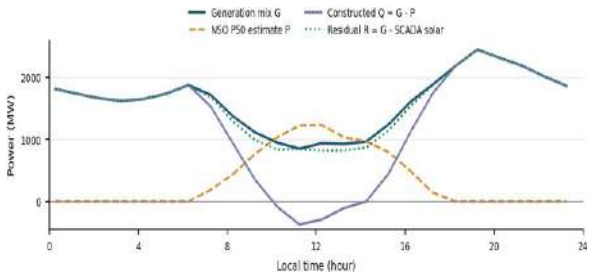



Fig. 3. Operating-day reconstruction for 2 August 2026. Generation mix G, NSO P50 estimate P, constructed proxy Q = G − P, and observable non-SCADA-solar residual R = G − Ssc; MW values are plotted at hourly :15 observations.

### *D. Merit-order dispatch advisory*

The dispatch-advisory component is an illustrative deterministic post-processor, not a validated dispatch optimization or a connection to the NSO energy-management system. Its configurable rule metadata assigns wind, solar and hydro to must-run classes, assumes a coal minimum-stable fraction of 45%, and treats oil generation as relatively flexible. When a low-residual flag or a proxy below 0.92 times a configurable threshold is detected, the code may recommend reducing up to 35% of nameplate capacity for each of at most three larger oil units, followed by standby instructions where indicated. Other rules recommend preserving roughly 90–150

MWh of hydro energy for evening demand and scheduling fast reserve. A 100-MW oil unit with a requested reduction of at least 35 MW would produce the illustrative schedule-line structure "−35 MW at [time]". This is an example of rule logic, not a recorded dispatch execution for 2 August 2026. Plant-specific ramp limits, reserve margins, transmission security and economic feasibility are not evaluated; neither recommendations nor heuristic confidence scores should be interpreted as operational dispatch instructions.

## VII. DISCUSSION AND LIMITATIONS

The corrected-lag evaluation addresses the 95-observation daily archive and extends the earlier study with four matched persistence offsets and no-lag ablations. However, the reported 1,974-instance test window remains a single chronological holdout within February–August 2026 and does not establish performance across a complete annual cycle or changing monsoon and curtailment regimes. Rolling-origin experiments and additional seasons would improve assessment of variability across training origins. The paired confidence intervals supplied with the evaluation exports describe uncertainty in relative errors; their construction and full underlying fold outputs should accompany any independent statistical reproduction.

Second, all numerical results in Tables II and III remain retrospective conditional reconstructions: the models use historical weather aligned to the target interval and autoregressive observations. Hourly weather interpolation may incorporate information unavailable at the forecast issue time. The reported superiority over one-hour, daily or weekly persistence must therefore not be described as validated lead-time forecasting skill. A prospective comparison would require fixed forecast origins, issue-time weather products and covariates available at each origin. Historical-forecast services may support such a study where coverage and access rights permit [16].Third, neither Q nor the SCADA-to-NSO solar-estimate ratio has independent meter-based ground truth. The holdout threshold experiment measures sensitivity to the daytime inclusion rule, not physical accuracy. Installed rooftop capacity does not establish actual 15-minute PV production. A capacity-consistency assessment using irradiance, uncertainty intervals and available inverter data would be necessary before making claims about untelemetered rooftop generation.Fourth, the advisory is an indicative deterministic heuristic, not a verified unit-commitment or security-constrained dispatch system. It has not been validated against plant ramp-rate constraints, reserve limits, transmission constraints or dispatch economics. Curtailment-affected intervals may reflect past operator decisions rather than unconstrained solar availability. The publication release intentionally excludes redistribution of source-provider record-level observations and does not claim rights beyond the applicable provider terms (Section V-A).

## VIII. CONCLUSION

This paper presented SolarisPredict-SL as an integrated research framework for retrospective reconstruction of Sri Lankan NSO generation-mix and solar-estimate signals, with an indicative dispatch-advisory prototype. Following calendar-time lag correction, XGBoost achieved MAEs of 49.79 MW for G, 46.89 MW for the observable residual R, 23.57 MW for estimated solar P and 57.11 MW for constructed Q on a 21-day holdout. Weather/calendar-only ablations produced substantially larger errors, but 15-minute persistence outperformed the G model and yielded an uncertain, small difference for R. These results support the value of lagged information in retrospective reconstruction without demonstrating operational forecast-origin skill or validating physical net load. The SCADA-to-estimate ratio and 2 August case study provide descriptive system context, while actual dispatch feasibility, independent PV measurements and prospective forecast validation remain outside the demonstrated evidence.

## REFERENCES


[1] Ceylon Electricity Board, Statistical Digest 2025. Colombo, Sri Lanka, 2025.

[2] EconomyNext, "Sri Lanka solar firms say expanding 'Sunny Sunday' curtailment has hit debt service," Jan. 22, 2026. [Online]. Available: https://economynext.com/sri-lanka-solar-firms-say-expanding-sunny-sunday-curtailment-has-hit-debt-service-258528/. Accessed: Nov. 8,

[3] P. A. G. M. Amarasinghe and S. K. Abeygunawardane, "Application of machine learning algorithms for solar power forecasting in Sri Lanka," in Proc. IEEE Int. Conf. Electrical Engineering (EECon), Colombo, Sri Lanka, Sep. 2018.

[4] A. A. K. U. Abeysingha, K. Ahilan, A. S. Sritharan, D. H. G. A. E. Jayasinghe, and R. Valluvan, "Electricity load/demand forecasting in Sri Lanka using deep learning techniques," in Proc. Int. Conf. Information and Automation for Sustainability (ICIAFS), 2021.

[5] S. D. A. V. S. Premathilaka, R. D. Yapa, and R. Punchi-Manage, "Uncovering the solar energy potential of reservoirs in Sri Lanka: A time series and machine learning approach," in Proc. Peradeniya Univ. Int. Research Symp. and Exposition (iPURSE), vol. 26, Aug. 2025, p. 334.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[7] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32

[9] S. J. Taylor and B. Letham, "Forecasting at scale," The American Statistician, vol. 72, no. 1, pp. 37–45, 2018.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748–1764, 2021.

[12] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," International Journal of Forecasting, vol. 32, no. 3, pp. 914–938, 2016.

[13] R. Weron, Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. Chichester, U.K.: Wiley, 2006.

[14] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.

[15] P. Denholm, M. O'Connell, G. Brinkman, and J. Jorgenson, "Overgeneration from solar energy in California: A field guide to the duck chart," National Renewable Energy Laboratory, Tech. Rep. NREL/TP-6A20-65023, Nov. 2015.

[16] Open-Meteo, "Historical Weather API; Historical Forecast API; Previous Runs API." https://open-meteo.com/en/docs/historical-forecast-api; [17] NASA Langley Research Center, "Prediction of Worldwide Energy Resources (POWER)."

[18] Electricity Distribution Lanka, "EDL Care generation summary." [Online].

[19] G. Muthugala, "SolarisPredict-SL: Source implementation and evaluation procedures," GitHub repository, commit cf0abb84fbf9a440e1e0caa19dbe1e0a90c0d577, Nov. 2026. [Online].

# Explainable AI for Academic Article Classification To Support Research and Learning

S. A. G. K. Gunasekara
*Department of Computing and Information Systems*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
sagkgunasekara@std.appsc.sab.ac.lk

P. K. D. K. Kaushalya
*Department of Computing and Information Systems*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
kaushalya@foc.sab.ac.lk

***Abstract*—With the rapid growth of scholarly publications, automated academic article classification has become an important tool for researchers. However, deep learning classifiers are typically black boxes that provide little insight into their predictions, which can reduce user trust. This article proposes an explainable framework that combines a fine-tuned SciBERT classifier with a consensus fusion method based on SHAP and Integrated Gradients (IG), and converts the resulting keywords into template-based natural-language explanations. A pilot survey of 100 participants motivated the need for explainable classification: 83% found it difficult to identify the correct research category of an article, and 86% were interested in a system that explains its decisions. A SciBERT model trained on 35,000 balanced arXiv abstracts from seven computer science fields achieved an accuracy of 90% and a weighted F1-score of 0.90. On 500 held-out papers, the consensus method was compared with five baselines (random selection, LIME, SHAP, IG and rank-average fusion) using four fidelity measures. It achieved the best value on all four (sufficiency 86.2%, comprehensiveness 0.319, flip rate 60.8%, stability 93.5%), with the largest gain in flip rate. A within-subjects user study with 300 participants showed that providing explanations alongside predictions significantly increased trust, perceived reliability, efficiency, adoption intent and understanding, and decreased suspicion (all $p < .001$ after Bonferroni correction). These findings highlight the potential of explainable classification to support researchers.**

***Keywords—Explainable AI, SciBERT, SHAP, Integrated Gradients, Academic Article Classification***

## I. INTRODUCTION

The volume of scholarly publications is growing at a rapid rate in nearly every discipline, and thousands of new articles are released to researchers every year in digital repositories like arXiv. Automated academic article classification, which categorizes research articles into predefined sub-areas based on their textual descriptions, has emerged as a crucial tool for literature retrieval, digital library management, and research discovery. Previous feature-based methods have shown significantly lower classification accuracy than transformerbased language models, with SciBERT [1], a model pretrained on scientific text, demonstrating particularly strong results. These models are, however, largely black boxes: they are accurate but provide little insight into the reasoning behind a specific prediction. This opacity may reduce trust and slow adoption in academic research contexts, where users rely on classification results to find relevant literature. Existing work has three common shortcomings: (i) most studies employ only one explanation method and are therefore subject to the limitations and assumptions of that method; (ii) few studies assess the quality of explanations using both quantitative fidelity measures and user-centred studies [8], [9], [10]; and (iii) explanations are often presented as raw importance scores rather than converted into a form that non-experts can read. A pilot survey of 100 computer science researchers and students, conducted before the system was developed, confirmed that

this is a real rather than assumed problem: 83% of respondents reported difficulty determining an article's subject area, and 86% expressed interest in a classification system that explains its decisions.Motivated by these findings, this paper makes the following contributions:

- A SciBERT-based classifier trained on 35,000 balanced arXiv abstracts spanning seven computer science subject categories.
- A consensus fusion method that combines SHAP and Integrated Gradients attributions into a single, more faithful set of keywords, converted into templatebased natural-language explanations.
- A quantitative fidelity assessment of the consensus method against five baselines (random, LIME, SHAP, IG and rank-average fusion)
- A within-subjects user study (N = 300) quantifying the impact of explanations on users' trust, reliability perception, and understanding

## II. RELATED WORK

Automated classification of scientific documents has moved from handcrafted features to transformer-based language models. BERT [16] is a deep bidirectional model pretrained for language understanding, and SciBERT [1] extends it to the scientific domain, achieving better performance than general-domain models on scientific NLP tasks.

Explanations of text classifiers are very common and are called post hoc explanation methods. LIME [13] and SHAP

[14] measure feature importance by making perturbations or local approximations to a model and Integrated Gradients [15] attribute predictions by calculating a model's internal gradients. They have been used in emotion classification [4], misinformation and fake-news identification [17, 18] and in general text classification [5]–[7] and usually only one explanation method is employed in each.

Attribution methods often disagree on which features matter for the same prediction. Roy et al. found frequent disagreements in feature rankings between LIME and SHAP for software defect predictors and proposed aggregating the two explanations so that the parts on which they agree are emphasised [21]. Laberge et al. seek consensus across models instead, extracting feature-importance statements that hold for all well-performing models [22]. Neely et al. compared several explanation methods on language tasks and found that, in most cases, the methods did not agree [23]. To our knowledge, few studies apply agreement-based aggregation of SHAP and IG to transformer-based text classification and test the result with both fidelity measures and a user study. We take a deliberately simple approach: tokens selected by both SHAP and Integrated Gradients are kept first, and the result is compared with plain rank-average (Borda) fusion and other baselines.Benchmarks such as ERASER [8] and conceptual frameworks [9] propose quantitative measures such as sufficiency and comprehensiveness, and work on interpretability stresses the need for human-centered evaluation [10], [19], [20], but few applied studies report both kinds of evidence together. The use of validated questionnaires like the Trust in Automation scale [11] and the Technology Acceptance Model [12] to measure the impact of explanations on user perception is also relevant, as is recent research that has emphasized the importance of user-centred design of XAI in clinical [2] and language-tool [3] contexts.

## III. METHODOLOGY

### *A. Research Design*

The study followed a six-stage design combining quantitative model development with a within-subjects user evaluation (Fig. 1): (1) a pilot survey to confirm the need for explainable classification; (2) construction of a labelled dataset of computer science abstracts; (3) fine-tuning a transformer classifier; (4) generation of explanations using SHAP, Integrated Gradients, and a proposed consensus fusion; (5) fidelity evaluation of the six explanation methods; and (6) a user study measuring the effect of explanations on trust and understanding.

### *B. Dataset and Model development*

A balanced dataset of 35,000 arXiv abstracts was constructed across seven computer science subject categories: Machine Learning (cs.LG), Artificial Intelligence (cs.AI), Computer Vision (cs.CV), Natural Language Processing (cs.CL), Robotics (cs.RO), Cryptography and Security (cs.CR), and Computers and Society (cs.CY), with 5,000 papers per category. To keep a single-label format, papers that matched two or more target categories were excluded. The data were split into training, validation and test sets (70/15/15) using a stratified split with a fixed random seed (42).

The SciBERT model (scibert_scivocab_uncased) was fine-tuned with a linear classification head on the concatenated title and abstract, tokenised with WordPiece to a maximum sequence length of 512. Training used the Hugging Face Trainer API with the AdamW optimiser for 5 epochs, a learning rate of $2\times10^{-5}$, a batch size of 16, fp16 mixed precision and a label smoothing factor of 0.1. The checkpoint with the best validation accuracy was kept. Training ran on a single NVIDIA T4 GPU with the same random seed (42). Table I summarises the configuration.

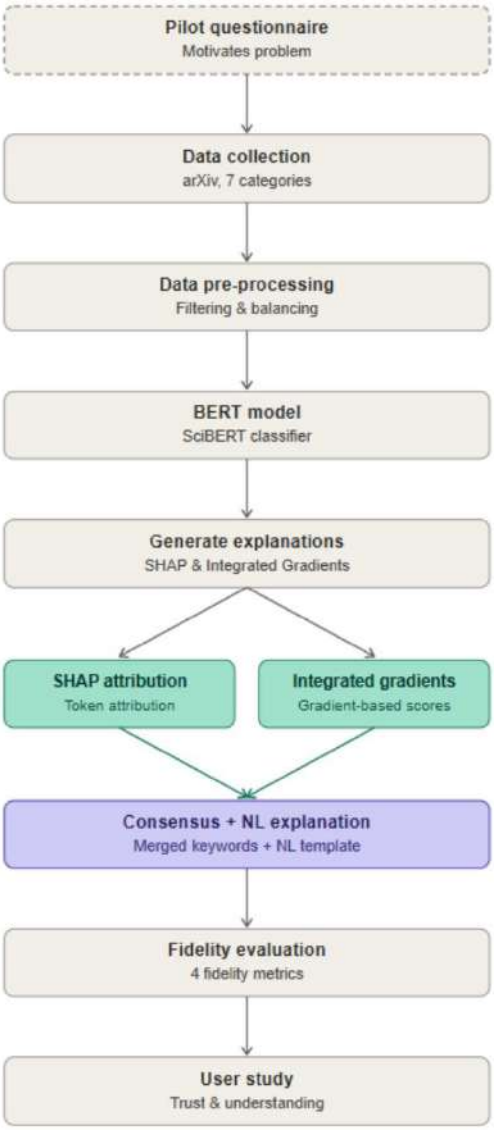



Fig. 1. Methodology Diagram

### *C. Explanation Methods*

The trained classifier was explained with two complementary attribution methods. SHAP computed Shapley-value token attributions by masking parts of the input and measuring the change in the predicted class probability. It used a text masker with the model's tokenizer and the Partition explainer with 500 model evaluations per paper. Integrated Gradients (IG) computed attributions from the gradients along a path from a zero-vector baseline to the input embedding, using 50 integration steps. Attributions were summed over embedding dimensions to give one signed score per token. For both methods, attributions were computed for the predicted class, and only tokens with positive attribution were kept. Stop words and sub-word tokens shorter than three characters were removed,

and the 20 highest-ranked tokens were retained as the candidate set.

Because SHAP and IG estimate importance differently, their top tokens often differ. We treat a token selected by both methods as stronger evidence of importance than a token selected by only one. Algorithm 1 combines the two lists into at most k = 7 keywords. Tokens selected by both methods are kept first, ranked by their mean normalized rank (lower means more important). If fewer than k tokens are agreed, the remaining places are filled with the best-ranked tokens from either list, so neither method is favoured.

TABLE I
SCIBERT FINE-TUNING HYPERPARAMETERS

| Hyperparameter | Value |
|---|---|
| Base model | SciBERT (scivocab_uncased) |
| Learning rate | $2\times10^{-5}$ |
| Batch size (train/eval) | 16 / 16 |
| Epochs | 5 |
| Max sequence length | 512 tokens |
| Label smoothing | 0.1 |
| Mixed precision | fp16 |
| Checkpoint selection | Best validation accuracy |

Algorithm 1: SHAP–IG Consensus

Input: top-20 positive tokens S (SHAP) and G (IG), in rank order; cap k = 7

Output: consensus keyword set C

1. For each token t in S ∪ G and each method M ∈ {S, G}, compute r_M(t) = rank_M(t) / 20. If t is not in M, set r_M(t) = 21/20.

2. Compute the mean rank r(t) = (r_S(t) + r_G(t)) / 2.

3. Compute the agreed set A = S ∩ G.

4. If |A| ≥ k, return the k tokens of A with the lowest r(t).

5. Otherwise, set C = A, then add tokens from (S ∪ G) \ A in ascending order of r(t) until |C| = k.

6. Return C.

The keywords are not shown in raw form. A deterministic template converts the keywords, the predicted category and a confidence descriptor (very high / high / moderate / low) into a sentence. A fixed template was chosen instead of a generative model so that every word in the explanation can be traced to the attribution scores. An example is shown in Section IV-D.

*D. Fidelity Evaluation*

Six explanation methods were evaluated on 500 papers randomly sampled from the held-out test set: random selection of seven tokens, LIME, SHAP, IG, rank-average (Borda) fusion, and the proposed consensus method. All methods used the same keyword budget (k = 7) and the same papers.

LIME used its default text explainer and kept the seven highest-weighted tokens. Rank-average (Borda) fusion gave each token points from its position in the SHAP and IG top- 20 lists (20 points for rank 1 down to 1 point for rank 20, and 0 if absent) and kept the seven tokens with the highest total.

Four metrics were computed in the same way for every method. Sufficiency is the percentage of papers for which the keywords alone reproduce the original prediction. Comprehensiveness is the mean decrease in prediction confidence when the keywords are removed. Flip rate is the percentage of predictions that change when the keywords are removed. Keywords were removed by replacing them with the [MASK] token, and confidence was the softmax probability of the originally predicted class.

Stability measures how consistent the keywords are when the input wording changes slightly. For each paper, 10% of the words that were not selected as keywords by any method were replaced by WordNet synonyms. The perturbed abstract was kept only if the classifier predicted the same class as for the original (462 of 500 papers kept). Stability was the mean top-k overlap, $|C \cap C'| / k$, between the keyword set C of the original abstract and the keyword set C′ of the perturbed abstract, with k = 7. All methods were evaluated on the same perturbed texts, and the random baseline gives the overlap expected by chance. Each metric is reported as a mean with a 95% bootstrap confidence interval (2,000 resamples of the evaluation papers). Consensus was compared with each baseline using paired significance tests: McNemar's test for sufficiency and flip rate, which are binary outcomes per paper, and the Wilcoxon signed-rank test for comprehensiveness and stability, which are continuous. P-values were adjusted with the Bonferroni correction for multiple comparisons.

*E. User Study Design*

The study used a within-subjects design with 300 participants recruited through university mailing lists and social media groups, who gave informed consent. Each participant rated three classification examples taken from published papers (Computer Vision, NLP, and Cryptography and Security), all correctly classified by the model, under two conditions: the classification result alone, and the result with the generated natural-language explanation. The order of the articles was randomised, but the order of the two conditions was fixed. Seven constructs were rated on 5-point Likert scales: confidence, reliability and trust (adapted from the trustin-automation scale [11]), suspicion, efficiency and adoption intent (adapted from the technology acceptance model [12]), and understanding. Differences between conditions were tested with paired-samples t-tests and Wilcoxon signed-rank tests, with Bonferroni correction.

## IV. RESULTS

### A. Pilot Study Findings

The responses from 100 respondents were: 83% had a hard time determining the right category for a computer science article; 86% said an explanation-providing classification system would be helpful (10% unsure, 4% no). Participants had a medium trust in the AI-based classification (M = 3.16/5, SD = 0.92), and they expected explanations to boost their trust in the classification results (M = 3.91/5, SD = 0.93), but they valued the importance of understanding why a system classifies a paper at 4.13/5, SD = 0.87. The results of these investigations directly inspired the classification/explanation framework detailed in Section III.

### B. Classification Performance

A fine-tuned SciBERT model had an accuracy of 90% and weighted F1-scores of 0.90 on a test set of 5250 papers (Table II). F1-scores for each category varied from 0.83 (Machine Learning and Natural Language Processing) to 0.95 (Cryptography and Security, Robotics). The confusion matrix (Fig. 2) illustrates that misclassifications were mostly made between categories that are closely related to each other, such as between Machine Learning and NLP, with 58 and 45 papers, respectively, for instance, or between Computers and Society and NLP, with 26 and 48 papers, respectively. Cryptography and Security and Robotics were not confused with other categories very often.

TABLE II
CLASSIFICATION PERFORMANCE ON THE TEST SET

| Category | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| ML | 0.89 | 0.77 | 0.83 | 750 |
| AI | 0.93 | 0.95 | 0.94 | 750 |
| CR | 0.93 | 0.98 | 0.95 | 750 |
| CY | 0.91 | 0.90 | 0.91 | 750 |
| CV | 0.90 | 0.92 | 0.91 | 750 |
| NLP | 0.82 | 0.85 | 0.83 | 750 |
| RO | 0.94 | 0.95 | 0.95 | 750 |
| Accuracy | — | — | 0.90 | 5250 |
| Macro avg | 0.90 | 0.90 | 0.90 | 5250 |
| Weighted avg | 0.90 | 0.90 | 0.90 | 5250 |

### C. Explanation Fidelity

Table III compares the six explanation methods on the 500-paper evaluation set, using the same keyword budget (k = 7) for all methods

The intervals of consensus and Borda fusion overlap for sufficiency, which agrees with the paired test (p = 0.12) reported below.

The consensus method achieved the best value on all four metrics. Its largest gain was in flip rate (60.8%, compared with 42.8% for SHAP, 48.6% for IG and 54.2% for Borda fusion), which suggests that consensus keywords are more decisive for the model's prediction. Sufficiency (86.2% vs. 80.6–84.0%) and comprehensiveness (0.319 vs. 0.241–0.292) improved in a similar way. Stability was also highest for consensus (93.5%), while LIME and the random baseline were clearly lower. The differences between consensus and SHAP, IG and LIME were statistically significant on all metrics ($p < 0.001$, Bonferroni-corrected). The gap to Borda fusion was smaller: it was significant for flip rate ($p = 0.004$) and comprehensiveness ($p = 0.011$), but not for sufficiency ($p = 0.12$). Random selection performed far worse than every method, which confirms that the metrics measure real explanation quality.

### D. Explanation Example

As an example of the pipeline's output, a representative paper on low-light video enhancement is shown below. The consensus method produced seven keywords related to the central topic of the paper, and the explanation read: "This paper is classified as Computer Vision (cs.CV) with high confidence (92.5%). The prediction is primarily influenced by domain-specific terms such as 'videos', 'enhancement', 'robust', 'vision', 'illumination', 'contrast' and 'noise'. These terms strongly support the model's classification decision." This shows that the explanation uses domainspecific terms rather than generic wording, and it can be read without knowing the underlying attribution methods.

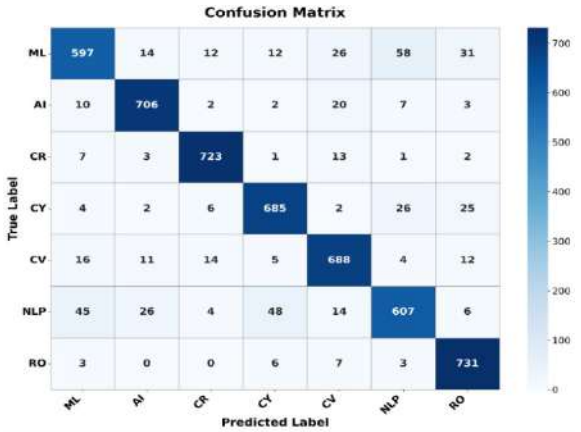



Fig. 2. Confusion matrix on the test set

TABLE III
EXPLANATION FIDELITY COMPARISON (500 PAPERS, K = 7; STABILITY ON 462 PAPERS). VALUES ARE MEANS WITH 95% BOOTSTRAP CIs IN BRACKETS.

| Method | Suff. (%) | Compr. | Flip (%) | Stab. (%) |
|---|---|---|---|---|
| Random 7 tokens | 38.4 [34.1, 42.7] | 0.061 [0.052, 0.070] | 14.8 [11.7, 17.9] | 12.5 [10.9, 14.1] |
| LIME (top-7) | 74.2 [70.4, 78.0] | 0.228 [0.211, 0.245] | 41.0 [36.7, 45.3] | 78.4 [76.2, 80.6] |
| SHAP (top-7) | 80.6 [77.1, 84.1] | 0.241 [0.223, 0.259] | 42.8 [38.5, 47.1] | 86.1 [84.2, 88.0] |
| IG (top-7) | 82.8 [79.5, 86.1] | 0.267 [0.248, 0.286] | 48.6 [44.2, 53.0] | 88.3 [86.5, 90.1] |
| Rank average (Borda) | 84.0 [80.8, 87.2] | 0.292 [0.272, 0.312] | 54.2 [49.8, 58.6] | 91.0 [89.4, 92.6] |
| **Consensus (SHAP+IG)** | **86.2 [83.2, 89.2]** | **0.319 [0.298, 0.340]** | **60.8 [56.5, 65.1]** | **93.5 [92.1, 94.9]** |

*E. User Study Results*

All constructs showed an increase from the withoutexplanation to the with-explanation condition, with Understanding showing the most marked increase (3.17 to 3.96) and Suspicion the most marked decrease (2.88 to 2.21). Of 21 per-article paired samples t-tests (7 constructs × 3 articles, Bonferroni-corrected α = .00238), only 20 were statistically significant, with the exception of Confidence for article #1 (t(299) = 1.426, p = .155) which became significant when averaged across all three articles. All seven of the constructs peaked at the corrected significance level (Table IV), and had medium (Efficiency, d = 0.451) to large (Understanding, d = 0.824; Suspicion, d = −0.766) effect sizes. Wilcoxon signed-ranked tests confirmed each of the t-tests both in direction and statistical significance. Fig. 3 summarizes the overall mean ratings.

TABLE IV
USER STUDY RESULTS: OVERALL MEAN RATINGS AND PAIRED T-TEST RESULTS (N = 300, DF = 299)

| Construct | M (Without) | M (With) | t | Cohen's d |
|---|---|---|---|---|
| Confidence | 3.26 | 3.68 | 8.256 | 0.477 |
| Reliability | 3.27 | 3.83 | 11.759 | 0.679 |
| Trust | 3.06 | 3.66 | 11.671 | 0.674 |
| Suspicion | 2.88 | 2.21 | −13.266 | −0.766 |
| Efficiency | 3.49 | 3.90 | 7.809 | 0.451 |
| Adoption Intent | 3.27 | 3.78 | 9.318 | 0.538 |
| Understanding | 3.17 | 3.96 | 14.271 | 0.824 |

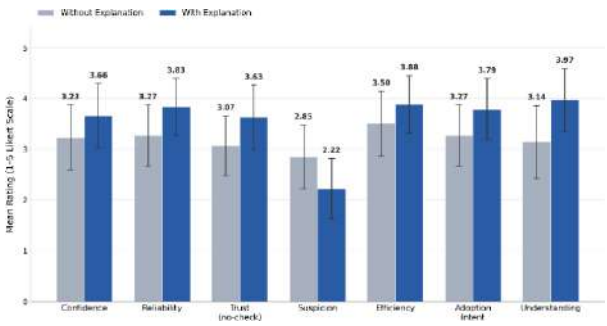



Fig. 3. Overall mean construct ratings, without vs. with explanation (N = 300; error bars denote ± 1 SD).

## V. DISCUSSION

Combined these results form a connection between a user need and a functional technical solution. The pilot study demonstrated that the classification of research category of an article is a genuine challenge and the fine-tuned SciBERT achieved a weighted F1-score of 0.90 on a seven way classification problem with highly overlapping categories. The fidelity evaluation supports the idea that method specific noise is removed by the agreement of complementary attribution methods. The SHAP–IG consensus had the highest performance in all four fidelity metrics on 500 held-out papers with a greater improvement in the flip rate (60.8% vs. 42.8– 54.2% in the baselines). This suggests that combining methods can improve explanation faithfulness, instead of assuming that faithfulness belongs to a single attribution technique. The gap between consensus and Borda fusion was smaller than the gap between consensus and the single methods, so a plain rank average already provides much of the benefit, and the value of giving priority to agreed tokens needs further testing.

The only per article result that was not significant was confidence, which was significant when averaged across the three articles. This implies that the level of confidence expressed in the statement could be influenced by the article/content and not the explanation itself and a wider and more diverse collection of articles is required to assess this. The explanations are template based and based on the attribution scores. This ensures that they are still followable but less fluent, which could be explored in future work to see if there is a positive effect on readability whilst maintaining faithfulness.

Limitations. There are several limitations to this study. The SciBERT classifier was fine-tuned once with a single random seed, so results may differ across training runs. The data come from one repository (arXiv) and use seven singlelabel categories, so the findings may not apply to other disciplines or to multi-label settings. We did not test the sensitivity of the consensus method to the top-20 candidate lists, the cap of seven keywords or the equal weighting of SHAP and IG. Explanations come from a fixed template, which keeps them traceable but less fluent.

The fidelity evaluation used 500 papers and one masking strategy ([MASK] replacement). Stability was tested with one perturbation type (synonym replacement) on 462 papers whose predicted class did not change, so it does not cover paraphrasing in general. Fidelity metrics also show how much the model relies on the keywords, not whether the keywords are correct to a human reader.

The user study could not show that consensus explanations are better than SHAP-only or IG-only explanations, because these were not compared. The condition order was fixed, so part of the increase may be due to order or demand effects. Only three articles and self-reported ratings in one session were used, so we cannot say whether the gains last over time or reflect well-calibrated trust. We did not ablate the components of the consensus mechanism (agreed-token priority, the fill step and the keyword cap).” The Conclusion mentions this, and the report relies on it

## VI. CONCLUSION AND FUTURE WORK

This paper presented an explainable framework for academic article classification that combines a fine-tuned

SciBERT classifier, a SHAP–Integrated Gradients consensus method and template-based natural-language explanations. The classifier achieved 90% accuracy across seven computer science categories. On 500 held-out papers, the consensus method achieved the best value on all four fidelity measures compared with the five baselines, with its largest improvement in flip rate (60.8% vs. 42.8–54.2% for the baselines). In a within-subjects study with 300 participants, adding an explanation to the classification result significantly increased trust, reliability, efficiency, adoption intent and understanding, and significantly decreased suspicion.

These results suggest that combining complementary attribution methods can improve the fidelity of explanations and users' trust in them. They should be read with the limitations above: the classifier was trained only once, stability was tested with a single perturbation type, the components of the consensus mechanism were not ablated, and the user study did not compare consensus explanations with SHAP-only or IG-only explanations. Future work will ablate the consensus mechanism (agreed-token priority, the fill step and the keyword cap), repeat training with several random seeds, add SHAP-only and IG-only conditions to the user study, test a larger and more varied set of articles, and deploy the system in a real digital library.

## REFERENCES


[1] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in Proc. EMNLP-IJCNLP, 2019.

[2] C. Panigutti et al., "Co-design of Human-centered, Explainable AI for Clinical Decision Support," ACM Trans. Interactive Intelligent Systems, vol. 13, 2023.

[3] Y. Kim, T.-L. V. Le, D. Kim, M. Lee, and S.-J. Lee, "Design Opportunities for Explainable AI Paraphrasing Tools: A User Study with Non-native English Speakers," in Proc. ACM DIS, 2025.

[4] M. H. F. Siddiqui, D. Inkpen, and A. Gelbukh, "Towards Interpretable Emotion Classification: Evaluating LIME, SHAP, and Generative AI for Decision Explanations," in Proc. 28th Int. Conf. Information Visualisation (IV), IEEE, 2024.

[5] F. M. Gafarov, V. R. Gafarova, and P. N. Ustin, "Explainable Artificial Intelligence Methods in Text Classification Machine Learning Models for Investigating Human Behavior in Digital Environments," in IEEE Int. Conf. Problems of Informatics, Electronics and Radio Engineering (PIERE), IEEE, 2024.

[6] S. Mustafa and M. H. Saeed, "Empowering Text Classification with NLP and Explainable AI for Enhanced Interpretability," J. Electrical Systems and Information Technology, vol. 12, 2025.

[7] M. A. Mersha, M. G. Yigezu, and J. Kalita, "Evaluating the Effectiveness of XAI Techniques for Encoder-Based Language Models," Knowledge-Based Systems, vol. 310, 2025.

[8] J. DeYoung et al., "ERASER: A Benchmark to Evaluate Rationalized NLP Models," in Proc. ACL, 2020.

[9] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: On the Concept of Explanations," in Proc. ACL, 2020.

[10] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.

[11] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," Int. J. Cognitive Ergonomics, vol. 4, no. 1, 2000.

[12] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, vol. 13, no. 3, 1989.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD, 2016.

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, 2017.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. ICML, 2017.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.

[17] M. J. Sujitha et al., "An Explainable AI Framework for Detecting Misinformation Using Semantic NLP Features and Attention-based Models," in Proc. ICSCDS, IEEE, 2025.

[18] K. Yoshida and J. Zhang, "Faithful Fake News Detection Based on Natural Language Explanation Generation," in Proc. NLPIR, ACM, 2025.

[19] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, 2018.

[20] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," Information Fusion, vol. 58, 2020.

[21] S. Roy, G. Laberge, B. Roy, F. Khomh, A. Nikanjam, and S. Mondal, "Why Don't XAI Techniques Agree? Characterizing the Disagreements Between Post-hoc Explanations of Defect Predictions," in Proc. IEEE Int. Conf. Software Maintenance and Evolution (ICSME), 2022, pp. 444–448.

[22] G. Laberge, Y. Pequignot, A. Mathieu, F. Khomh, and M. Marchand, "Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set," Journal of Machine Learning Research, vol. 24, no. 364, pp. 1–50, 2023.

[23] M. Neely, S. F. Schouten, M. J. R. Bleeker, and A. Lucic, "Order in the Court: Explainable AI Methods Prone to Disagreement," in ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI, 2021, arXiv:2105.03287.

# Adversarial Threats and Robust Defenses in Satellite-Based Deep Learning for Multi-Hazard Early Warning Systems: A Survey of Flood and Landslide Prediction

Satheesha Fernando
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
fernandorawini@gmail.com

Madusanka Mithrananda
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
madusanka.m@nsbm.ac.lk

***Abstract*—Floods and landslides are among the most destructive natural hazards, and events such as Cyclone Ditwah (Sri Lanka, 2025) show the human and economic cost of missed or late warnings. Satellite Earth-observation (EO) data combined with deep learning (DL) is increasingly used to map and assess these hazards within multi-hazard early warning systems (MHEWS), yet the security of such models is rarely examined. This survey states its search protocol and compares itself with existing surveys. It reviews CNN, LSTM and transformer models for flood and landslide analysis; presents a threat model and a structured taxonomy of seven threats, classified by attacker capability, attack surface, detectability and operational consequence; and reviews remote-sensing-specific adversarial evidence. It then assesses seven defense families, including adversarial training, certified robustness, detection, ensembles, explainable AI, federated learning and data provenance, together with their documented limits. Within our search, robustness evidence for hazard-specific models is limited to two early studies, and we found no evaluation of deep flood or landslide segmenters under input-perturbation attacks and no evaluated provenance framework. We close with research gaps and the limitations of this survey.**

***Keywords—Multi-hazard early warning systems, Earth observation, deep learning, adversarial machine learning, cybersecurity***

## I. INTRODUCTION

### *A. Motivation: Cyclone Ditwah*

Floods and landslides are becoming more frequent and severe with climate change, urbanization and land degradation [1]. Cyclone Ditwah made landfall on Sri Lanka's east coast on 28 November 2025 and triggered floods and landslides in all 25 districts. By 17 December, 643 casualties, 183 missing persons and more than 107,000 damaged houses had been recorded [2], and the World Bank estimated direct physical damage at US$4.1 billion, about 4% of GDP [3]. Satellite-derived flood extents were used to map affected agricultural land, about 58,340 ha of paddy in Trincomalee and Batticaloa [2]. The sources we reviewed do not document whether DL models were part of the operational warning chain during Ditwah; we use the event to motivate why the reliability and security of such models matter, not as evidence of their performance.

A multi-hazard early warning system (MHEWS) combines risk knowledge, monitoring and forecasting, warning dissemination and response capability [1], [4]. EO missions such as Sentinel-1 and Sentinel-2 [5], [6] feed DL models for flood mapping and landslide susceptibility

[7]–[9]. The cost of error is asymmetric: a missed warning risks lives, whereas a false alarm erodes trust and wastes response capacity, so a model that is accurate on clean test data but fragile under manipulation is not fit for this role.

### *B. Gap and Contributions*

Surveys of EO-based flood and landslide prediction concentrate on predictive accuracy [10]–[12]. Adversarial machine learning (AML) research in remote sensing concentrates on scene classification, object detection and SAR target recognition [13]–[15]. Their intersection, the security of EO-based DL inside an MHEWS, is rarely examined. This paper contributes: (1) a stated survey protocol and a comparison with existing surveys (Section II, Table I); (2) a threat model and a seven-threat framework mapped onto the MHEWS pipeline (Section IV, Fig. 1, Table III); (3) a review of remote-sensing-specific adversarial evidence and a critical assessment of defenses, including their documented limits (Sections IV–V, Table IV); and (4) research gaps and the limitations of this survey (Section VI).

## II. SURVEY METHODOLOGY AND RELATED SURVEYS

This is a structured narrative survey, not a PRISMA-style systematic review. We searched Google Scholar, IEEE Xplore, Scopus and publisher platforms (Elsevier, MDPI, Springer, Nature) in three passes: S1, EO-based DL for flood and landslide mapping or prediction; S2, adversarial, poisoning, backdoor and data-integrity attacks on remote-sensing or EO systems; S3, secure or trustworthy AI for early warning. Representative strings were S1: ("flood" OR "landslide") AND ("deep learning" OR "transformer") AND ("Sentinel" OR "SAR" OR "satellite"); S2: ("adversarial" OR "poisoning" OR "backdoor") AND ("remote sensing" OR "satellite" OR "SAR"); S3: "early warning" AND ("robust*" OR "secure" OR "trustworthy") AND ("deep learning" OR "AI"). Coverage runs from 2014, when adversarial examples were first reported [16],

to October 2026. We included peer-reviewed articles, archival preprints and official agency reports in English that fit a pass, and excluded works without EO data, purely hydrological models and non-archival material. Candidate studies were read and classified by threat, pipeline stage and defense type. The final corpus is 42 cited works. We did not record per-database hit counts or apply a formal quality score, so the search is not reproducible in the PRISMA sense (Section VI).

Table I compares this survey with the closest existing surveys along five dimensions. No single survey covers hazard-specific EO DL, security, remote-sensing-specific attacks, the MHEWS pipeline and defenses together.

TABLE I
COMPARISON WITH EXISTING SURVEYS

| Survey | Focus | EO DL | Sec. | RS atk | MHEWS | Def. |
|---|---|---|---|---|---|---|
| Mosavi et al. [10] | ML flood prediction | P | N | N | N | N |
| Reichenbach et al. [11] | Landslide susceptibility models | P | N | N | N | N |
| Zhang et al. [12] | DL flood forecasting and warning | P | N | N | P | N |
| Tiggeloven et al. [4] | AI for early warning; guardrails | P | P | N | Y | P |
| Biggio & Roli [17] | Adversarial ML, general | N | Y | N | N | P |
| Zhou et al. [18] | Adversarial ML, cybersecurity view | N | Y | N | N | Y |
| Xu et al. [14] | Attacks/defenses, RS scene classification | N | Y | Y | N | Y |
| Mei et al. [15] | Robustness survey and benchmark, RS | N | Y | Y | N | P |
| **This survey** | Security of EO DL in flood/landslide MHEWS | P | Y | Y | Y | Y |

Y = main focus; P = partly covered; N = not covered. EO DL = deep learning on EO data for flood/landslide; Sec. = adversarial or cyber threats; RS atk = remote-sensing-specific attacks; MHEWS = pipeline or warning-system view; Def. = defenses. Ratings are our reading of titles, abstracts and scope statements.

## III. SATELLITE-BASED DEEP LEARNING FOR FLOOD AND LANDSLIDE ANALYSIS

### *A. Satellite EO Data Sources and Deep Learning Models*

Sentinel-1 C-band SAR is the workhorse for all-weather flood mapping because open water is a specular reflector that returns little backscatter, and SAR works through cloud and at night [5]. Sentinel-2 multispectral imagery supports land-cover and water-index mapping but is limited by cloud and cloud shadow, which is acute during cyclones [6]. For landslides, InSAR time series monitor slow ground deformation [19], while terrain, rainfall and soil-moisture layers condition susceptibility [11]. CNN and U-Net segmenters dominate pixel-level mapping, LSTM models handle rainfall and hydrological series, and transformers add long-range context [8], [12]. Foundation-style models aimed at cross-region transfer are emerging [20].

### *B. What the Models Do, and How They Are Evaluated*

Table II lists representative studies; it is illustrative, not exhaustive. Three points matter for security. First, EO models mostly map an event in progress or zone susceptibility from static factors; pre-event forecasting additionally needs rainfall forecasts and hydrological models [4]. Second, reported accuracy, for example up to 99.1% for landslide susceptibility with RF-GA [9], is measured on clean, region-specific test sets, and cross-region generalization remains a known weakness [20]. Third, to our knowledge none of the studies in Table II reports an adversarial-robustness evaluation.

TABLE II
REPRESENTATIVE EO-BASED FLOOD AND LANDSLIDE STUDIES

| Study | Hazard | EO data | Method | Note |
|---|---|---|---|---|
| Bonafilia et al. [7] | Flood | Sentinel-1; 4,831 chips, 11 events | FCNN baselines | Benchmark; part weakly labelled |
| Sharma & Saharia [8] | Flood | Sentinel-1 | ViT deep ensembles | Uncertainty estimates; weak labels from optical |
| Ullah et al. [21] | Multi-hazard | EO and terrain factors | CNN | Susceptibility maps, E. Hindu Kush |
| Razavi-Termeh et al. [9] | Multi-hazard | Sentinel-1/2 | RF-GA, RF-PSO | Up to 99.1% (landslide); ML, not DL |
| **Study** | **Hazard** | **EO data** | **Method** | **Note** |
| Reichenbach et al. [11] | Landslide | Geospatial factors | Statistical models | Review of susceptibility models |

## IV. THREAT MODEL AND TAXONOMY

### *A. Attacker Model and EO-Specific Constraints*

Fig. 1 shows the MHEWS data pipeline considered here and where each threat acts. We distinguish four attacker types: (i) an external adversary with only query access to a public or partner-facing prediction service (black-box); (ii) a supply-chain adversary who can contribute training data, label inventories or pretrained weights; (iii) an insider or intruder with write access to the archive, processing platform or distribution channel; and (iv) a network or radio-frequency adversary against communications and positioning.

EO inputs impose constraints that natural-image attacks do not face. Imagery is radiometrically calibrated, orthorectified, resampled and, for SAR, speckle-filtered before it reaches a model, so a perturbation crafted on a model input must be injected after that chain, which requires ingest or archive access; physical patches are plausible mainly for very-high-resolution optical data [13]; and SAR statistics differ from natural images [15], [22]. This reasoning is our analysis rather than an empirical finding.

### *B. Evidence from Remote-Sensing Adversarial Research*

Small patch perturbations can fool satellite scene classifiers [13]. Chen et al. studied optical and SAR scene classification [23]; Xu et al. evaluated attacks and defenses for scene classification [14]; Xu and Ghamisi released UAE-RS, a black-box universal-perturbation benchmark whose examples transfer across networks in scene classification and semantic segmentation [22]; Mei et al. benchmarked adversarial and natural robustness for classification and detection [15]; and Chen and Chu reviewed attacks and defenses for aerial detection

across optical, infrared and SAR sensors, noting that most studies use simulated or idealized attacks [24]. Remote-sensing-specific defenses include a task-guided denoiser [25] and detection of adversarial high-resolution images [26]. All of this concerns scene classification, detection or generic segmentation. Hazard-specific evidence is sparse. Kothari et al. simulate annotation errors, framed as label poisoning, on a U-Net for inland-water segmentation from SAR images and find that performance holds up to a point before dropping sharply [27]. A recent study applies the zeroth-order-optimization (ZOO) attack to random-forest and XGBoost landslide-susceptibility models and reports that the more accurate model (RF) was less robust, so accuracy and robustness are distinct properties [28]. Neither study attacks a deep flood or landslide segmenter with input perturbations, which is the gap this survey addresses; both are preliminary, and the first is an unreviewed preprint.

Beyond evasion [16], [29], [30], flood and landslide models are exposed to training-time attacks because they are trained on inventories and weak labels derived from optical imagery [7], [8], which are attractive poisoning targets [17]; backdoored models keep clean accuracy [31], [32]. Prediction APIs can be probed to extract a surrogate model [33], and generative models can fabricate realistic satellite imagery [34].

### *C. Structured Threat Framework*

Table III expands the usual list of threats into a framework with attacker capability, attack surface, target, objective, detectability and operational consequence. Detectability is a qualitative judgment from the cited literature and the pipeline analysis above, not a measured value. Fig. 1 maps the threats and the defenses of Section V onto pipeline stages, and Fig. 2 groups the seven threats into four classes by the access the attacker needs.

### *D. Illustrative Scenario*

Consider a Sentinel-1 flood-mapping service that feeds an inundation layer to a national disaster centre during a cyclone, as in the Ditwah response [2]. This scenario is hypothetical and is not drawn from any reported incident. An intruder with write access to the archive (T4) adds a small, structured perturbation to the post-event SAR scenes of one river basin before inference (T1). The perturbed scenes look normal to analysts, the segmenter returns a smaller flood extent, and the basin is ranked lower for relief. Because the segmenter is not itself altered, model-integrity checks pass; only input provenance, cross-checking against optical imagery or gauges, or a consistency test against the previous acquisition could flag the change. A poisoned label inventory (T2) would produce the same outcome more slowly and persistently, and would be harder to attribute. The scenario shows why Table III lists detectability and consequence separately from the attack itself: the cheapest attack to mount is not always the one that does the most operational harm.

Two properties of EO workflows make this concern practical. First, many agencies consume derived products (flood masks, susceptibility rasters) rather than raw scenes, so tampering after the model is as effective as attacking the model and is not covered by adversarial-training defenses. Second, cyclone events compress decision time: analysts have hours, not days, to cross-check outputs, which favours automated consistency checks built into the pipeline over manual review.

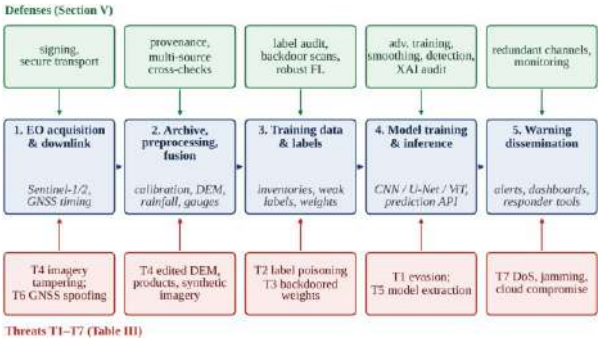



Fig. 1. MHEWS data pipeline with the threats (T1–T7, Table III) and defense families (Section V) that act at each stage. Original diagram prepared by the authors.

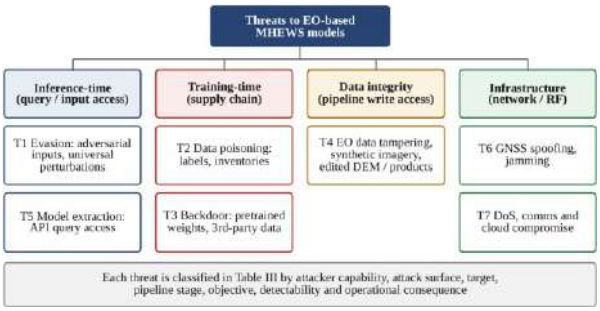



Fig. 2. Original taxonomy of the seven threats by the access an attacker needs; each threat is characterized further in Table III. Prepared by the authors from the analysis in Section IV

TABLE III
STRUCTURED THREAT FRAMEWORK FOR EO-BASED MHEWS

| Threat | Attacker capability | Surface / stage | Target | Objective | Detectability | Operational consequence |
|---|---|---|---|---|---|---|
| T1 Evasion (adversarial examples) | Black-/grey-box; perturbed inputs injected after calibration (ingest or archive access), or physical patch on VIIR optical | Inference input (stage 4) | Deployed model; input scenes | Suppress or fabricate a flood/landslide mask | Low by eye; input-consistency checks help | Missed warning or false alarm |
| T2 Data poisoning | Contributes or edits labels, inventories, crowd or weak labels | Training data (stage 3) | Inventories, label sets | Bias susceptibility maps; degrade accuracy | Low–medium (label audit, outlier tests) | Systematically missed or inflated risk zones |
| T3 Backdoor | Supply chain; pretrained weights, third-party datasets | Training and fine-tuning (3–4) | Model weights | Trigger-conditioned misclassification | Low; clean accuracy preserved | Targeted suppression of warnings |
| T4 EO data manipulation, synthetic imagery | Write access to archive or distribution; generative fabrication | Ingest, archive, fusion (1–2) | Imagery, DEM, derived products | Mislead situational awareness | Low–medium (provenance, forensics) | Misallocated response; false confidence |
| T5 Model extraction | Query access to prediction API | Inference service (4) | Model, training data | Steal model; craft transferable attacks | Medium (query monitoring) | Enables T1; may leak sensitive data |
| T6 GNSS spoofing / jamming | RF capability near receivers or ground stations | Positioning, timing (1) | Geolocation, timestamps | Misleadize hazard; corrupt time series | Medium (multi-constellation checks) | Alerts for the wrong area |
| T7 DoS, comms and cloud compromise | Network access or insider | Dissemination (5); cloud | Services, channels | Delay or block warnings | High (visible outage) | Late or absent alerts |

## V. DEFENSES AND THEIR LIMITATIONS

Table IV summarises seven defense families, the threats they counter, the evidence we found for EO or hazard models, and their documented limits. Adversarial training [35] is the strongest empirical defense against T1 but protects only against the threat model used in training and costs accuracy and compute. Certified randomized smoothing [36] gives provable guarantees, but only for small $\ell 2$-norm radii whose relation to physically realisable perturbations of calibrated EO products is unclear. Detection and denoising have been demonstrated for scene classification [25], [26], yet defenses that resist non-

adaptive attacks are often broken by adaptive ones [30]. Ensembles improve accuracy [9], but adversarial examples transfer between models, and ensembles of weak defenses are not strong [37], [38]. Explainable AI supports audit and human review [39], but attributions can themselves be manipulated [40], so XAI does not prevent attacks. Federated learning

[41] supports data sovereignty across agencies but opens a poisoning and backdoor channel through malicious clients [42]. For provenance and secure fusion we found no evaluation on an operational EO pipeline: signing proves integrity relative to the signer, not correctness of what was signed, and cross-checking independent sources (for example SAR against optical imagery and gauges) is a sound design heuristic rather than an evaluated defense. Defenses should therefore be layered, with provenance upstream, robust training and detection at the model, and XAI-assisted human review before dissemination.

## VI. RESEARCH GAPS AND LIMITATIONS

Gaps. (G1) Adversarial-robustness evidence for hazard-specific EO models is limited to label-noise [27] and tabular ZOO [28] studies, with no benchmark for deep flood or landslide segmenters under input-perturbation attacks; a UAE-RS-style benchmark [22] built on flood data [7] would be a direct next step. (G2) Flood and landslide are modelled separately although cyclone rainfall drives both; multi-hazard work so far relies on susceptibility zoning with CNN or random-forest models [9], [21]. (G3) Cross-region generalization is weak [20], and we found no EO-based DL study with a robustness evaluation specific to Sri Lanka or similar data-scarce tropical settings. (G4) No integrated cyber-resilience framework or data-provenance practice for EO-based MHEWS has been evaluated.

TABLE IV
DEFENSE FAMILIES AND DOCUMENTED LIMITS

| Defense | Threat | Evidence for EO / hazard models | Documented limits |
|---|---|---|---|
| Adversarial training [35] | T1 | RS scene classification [14], [15]; none for hazard models (robustness measured only without a defense [27], [28]) | Covers only the trained threat model; accuracy and compute cost |
| Certified smoothing [36] | T1 | None found for EO hazard models | Small certified radii; fit to realisable EO perturbations unclear |
| Detection / denoising [25], [26] | T1 | RS scene classification | Adaptive attacks can evade [30] |
| Ensembles [37], [38] | T1 | Used for accuracy [9]; robustness untested | Attacks transfer; weak-defense ensembles not strong |
| XAI audit [39], [40] | T2, T3 | Not evaluated for EO hazards | Attributions can be manipulated; no prevention |
| Federated learning [41], [42] | Privacy | Not evaluated for EO hazards | Malicious clients can poison or backdoor the global model |
| Provenance, secure fusion | T4, T6 | No evaluation on an operational EO pipeline found | Signing proves integrity, not correctness; needs infrastructure |

Limitations of this survey. It is a narrative survey: we did not record screening counts or score study quality, so it is not reproducible in the PRISMA sense. Only English-language sources were searched, and some cited works are preprints. Table II is illustrative, not exhaustive. The threat framework is analytical: attacker capabilities, detectability ratings and the EO-specific constraints in Section IV-A are our inference from non-hazard remote-sensing work and general AML, and T5–T7 draw on general cybersecurity rather than EO-specific evidence. The ratings in Table I reflect our reading of scope statements. Finally, we ran no experiments, so claims about defense effectiveness rest on the cited studies. Evaluation Protocol for Future Robustness Studies

A floodor landslide-specific robustness study should state its protocol explicitly so results are comparable. We suggest the following minimum design, which follows evaluation practice in the cited adversarial literature [22], [30], [35] rather than any existing EO-hazard benchmark. (i) Models: at least one CNN or U-Net and one transformer segmenter trained on a public flood dataset [7]. (ii) Attacks: white-box PGD and a black-box transfer attack, with perturbation budgets stated in the units of the calibrated product (for example dB for SAR backscatter), plus a label-poisoning rate sweep. (iii) Metrics: clean and robust IoU, the change in mapped flood area, and the false-negative rate for the most vulnerable administrative units, since a small area error can move a district across a response threshold. (iv) Defenses: adversarial training, randomized smoothing [36] and one detector, each evaluated against an adaptive attacker. (v) Generalization: a held-out region and a held-out event, because robustness measured in-region can overstate protection [20]. Recommended Practices and Future Work

From the analysis above we draw five practical recommendations for agencies and researchers; they are derived from the cited literature and our threat framework, not from field trials. (R1) Record provenance for every EO scene and derived product, including processing version and checksum, so tampering after acquisition is detectable (T4). (R2) Require a cross-source consistency check before dissemination, for example SAR against optical imagery, gauges and the previous acquisition (T1, T4). (R3) Audit label inventories and pretrained weights before reuse, and prefer sources with documented lineage (T2, T3). (R4) Rate-limit and log prediction APIs and avoid exposing raw confidence maps externally (T5). (R5) Report robustness alongside accuracy, using at least one adaptive attack and a held-out region. Research should prioritize a flood-specific robustness benchmark, evaluation of certified and detection defenses on SAR segmentation, and field studies of how analysts act on contested model outputs.

## VII. CONCLUSION

Satellite-based DL is now part of how floods and landslides are mapped and assessed. Cyclone Ditwah illustrates the human and economic cost of such events, but the sources we reviewed do not show that DL models were used in its operational warning chain, so we do not claim a specific model failure or success there. We presented a threat model, a structured seven-threat framework and an assessment of seven defense families together with their limits. The evidence base is thin where it matters most: robustness evidence for hazard-specific EO models is limited to two early studies, with no evaluation of deep flood or landslide segmenters under input-perturbation attacks, and no evaluated provenance framework for EO-based MHEWS. Priorities are robustness benchmarks for hazard models, joint flood–landslide modelling, cross-region

validation, and layered, provenance-aware pipelines evaluated end to end.

## REFERENCES


[1] UNDRR, Global Assessment Report on Disaster Risk Reduction 2023. Geneva, Switzerland: United Nations Office for Disaster Risk Reduction, 2023.

[2] UN OCHA, "Sri Lanka: Cyclone Ditwah – Situation Report 3," Dec. 2025. [Online]. Available: https://www.unocha.org/publications/report/sri-lanka/sri-lanka-cyclone -ditwah-situation-report-3-23-december-2025

[3] World Bank, "Damage from Cyclone Ditwah in Sri Lanka estimated at $4.1 billion," press release, Dec. 22, 2025. [Online]. Available: https://www.worldbank.org/en/news/press-release/2025/12/22/damage -from-cyclone-ditwah-in-sri-lanka-estimated-at-4-1-billion

[4] T. Tiggeloven et al., "The role of artificial intelligence for early warning systems: Status, applicability, guardrails, and ways forward," iScience, vol. 28, no. 11, Art. no. 113689, 2025, doi: 10.1016/j.isci.2025.113689.

[5] R. Torres et al., "GMES Sentinel-1 mission," Remote Sens. Environ., vol. 120, pp. 9–24, 2012.

[6] M. Drusch et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," Remote Sens. Environ., vol. 120, pp. 25–36, 2012.

[7] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1," in Proc. IEEE/CVF CVPR Workshops, 2020, pp. 835–845.

[8] N. K. Sharma and M. Saharia, "DeepSARFlood: Rapid and automated SAR-based flood inundation mapping using vision transformer-based deep ensembles with uncertainty estimates," Sci. Remote Sens., vol. 11, Art. no. 100203, 2025.

[9] S. V. Razavi-Termeh et al., "Optimizing ensemble learning for satellite-based multi-hazard monitoring and susceptibility assessment of landslides, land subsidence, floods, and wildfires," Sci. Rep., vol. 15, Art. no. 30968, 2025, doi: 10.1038/s41598-025-15381-2.

[10] A. Mosavi, P. Ozturk, and K.-W. Chau, "Flood prediction using machine learning models: Literature review," Water, vol. 10, no. 11, Art. no. 1536, 2018.

[11] P. Reichenbach, M. Rossi, B. D. Malamud, M. Mihir, and F. Guzzetti, "A review of statistically based landslide susceptibility models," Earth-Sci. Rev., vol. 180, pp. 60–91, 2018.

[12] Y. Zhang, D. Pan, J. Van Griensven, S. X. Yang, and B. Gharabaghi, "Intelligent flood forecasting and warning: A survey," Intell. Robot., vol. 3, no. 2, pp. 190–212, 2023, doi: 10.20517/ir.2023.12.

[13] W. Czaja, N. Fendley, M. Pekala, C. Ratto, and I.-J. Wang, "Adversarial examples in remote sensing," in Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst., 2018, arXiv:1805.10997.

[14] Y. Xu, B. Du, and L. Zhang, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 2, pp. 1604–1617, Feb. 2021, doi: 10.1109/TGRS.2020.2999962.

[15] S. Mei et al., "A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking," J. Remote Sens., vol. 4, Art. no. 0219, 2024, doi: 10.34133/remotesensing.0219.

[16] C. Szegedy et al., "Intriguing properties of neural networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2014.

[17] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognit., vol. 84, pp. 317–331, 2018.

[18] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity," ACM Comput. Surv., vol. 55, no. 8, Art. no. 163, 2022, doi: 10.1145/3547330.

[19] A. Ferretti, C. Prati, and F. Rocca, "Permanent scatterers in SAR interferometry," IEEE Trans. Geosci. Remote Sens., vol. 39, no. 1, pp. 8–20, Jan. 2001.

[20] Z. Gong et al., "CrossEarth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation," arXiv:2410.22629, 2024.

[21] K. Ullah, Y. Wang, Z. Fang, L. Wang, and M. Rahman, "Multi-hazard susceptibility mapping based on convolutional neural networks," Geosci. Front., vol. 13, no. 5, Art. no. 101425, 2022.

[22] Y. Xu and P. Ghamisi, "Universal adversarial examples in remote sensing: Methodology and benchmark," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5619815, doi: 10.1109/TGRS.2022.3156392.

[23] L. Chen, Z. Xu, Q. Li, J. Peng, S. Wang, and H. Li, "An empirical study of adversarial examples on remote sensing image scene classification," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 9, pp. 7419–7433, Sep. 2021.

[24] Y. Chen and S. Chu, "On the adversarial robustness of aerial detection," Front. Comput. Sci., vol. 6, Art. no. 1349206, 2024, doi: 10.3389/fcomp.2024.1349206.

[25] Y. Xu, W. Yu, and P. Ghamisi, "Task-guided denoising network for adversarial defense of remote sensing scene classification," in Proc. 2nd Workshop Complex Data Challenges Earth Observ. (CDCEO), CEUR Workshop Proc., vol. 3207, 2022.

[26] W. Li et al., "Spear and shield: Attack and detection for CNN-based high spatial resolution remote sensing images identification," IEEE Access, vol. 7, 2019.

[27] S. Kothari, S. Murali, S. Kothari, U. Verma, and J. Sreevalsan-Nair, "Adversarial robustness of deep learning models for inland water body segmentation from SAR images," arXiv:2505.01884, 2025.

[28] [INITIAL] Sun et al. [VERIFY FULL AUTHOR LIST], "Evaluating robustness of landslide susceptibility mapping models using ZOO attack and machine learning," Trans. GIS, vol. 30, no. 2, 2026, doi: 10.1111/tgis.70247.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. ICLR, 2015.

[30] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Security Privacy (SP), 2017, pp. 39–57.

[31] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv:1708.06733, 2017.

[32] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 1, pp. 5–22, Jan. 2024.

[33] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in Proc. 25th USENIX Security Symp., 2016, pp. 601–618.

[34] B. Zhao, S. Zhang, C. Xu, Y. Sun, and C. Deng, "Deep fake geography? When geospatial data encounter artificial intelligence," Cartogr. Geogr. Inf. Sci., vol. 48, no. 4, pp. 338–352, 2021.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018.

[36] J. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019.

[37] F. Tramèr et al., "Ensemble adversarial training: Attacks and defenses," in Proc. ICLR, 2018.

[38] W. He, J. Wei, X. Chen, N. Carlini, and D. Song, "Adversarial example defenses: Ensembles of weak defenses are not strong," in Proc. 11th USENIX Workshop Offensive Technol. (WOOT), 2017.

[39] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020.

[40] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 3681–3688.

[41] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), 2017, pp. 1273–1282.

[42] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in Proc. AISTATS, 2020, pp. 2938–2948.

# PACT: Post-hoc Adaptive Conditional Transfer for Risk-Controlled LLM Abstention in Low-Resource Languages

Senes Fitzroy
*Faculty of Computing, NSBM Green University*
*Colombo, Sri Lanka*
dsfmendis@students.nsbm.ac.lk

Isidara Gallage
*Faculty of Computing, NSBM Green University*
*Colombo, Sri*
Lankacharuniisidara@gmail.com

Sanuji Walgampaya
*Faculty of Computing, NSBM Green University*
*Colombo, Sri*
Lankasanujiwalgampaya2003@gmail.com

***Abstract*—Large language models are increasingly gated by anabstentionmechanism that refuses to answer when an uncertainty score exceeds a threshold. Choosing that threshold so the deployed system provably respects an error budget requires labelled calibration data, which is abundant inEnglish and scarce in low-resource languages such as Sinhala and Tamil. Practitioners therefore calibrate on English and deploy elsewhere. We show this silently voids the guarantee.We give an exact decomposition of the cross-lingual risk gap into acovariateterm, caused by the uncertainty score distribution moving, and aconditionalterm, caused by the score becoming less predictive of correctness. The decomposition has a sharp consequence: importance weighting and weighted conformal prediction—the standard tools for distribution shift—cancel only the covariate term, so they cannot repair cross-lingual calibration whenever the conditional term dominates. In a controlled study whose parameters are matched to effect sizes reported in the multilingual evaluation literature, an English-calibrated conformal gate overruns a10%risk budget by2.2×and violates it in100%of simulated deployments, and we show analytically that98.6%of that gap is conditional, capping what any reweighting method can recover. We propose PACT, which corrects the conditional term with a two-parameter calibration-shift model fitted to a handful of target-language labels andpropagatesitssamplinguncertaintyintoabootstrap upper confidence bound on the risk curve. With 50 labelled target items PACT cuts the violation rate to9%whileretaining76%of oracle answering utility, and an intercept-only variant is shown to be insufficient: modelling the decay of score informativeness, not merely the accuracy drop, is what restores control. The decomposition is model-independent; all performance figures come from the controlled simulation, and validation on real Sinhala and Tamil LLM outputs remains future work.**

***Keywords—large language models, uncertainty quantification, conformal risk control, low-resource languages, selective prediction***

## I. INTRODUCTION

Instruction-tuned large language models (LLMs) are now embedded in tools that answer questions for non-expertusers [1], [2]. Their competence is distributed as unevenlyas their training data: languages spoken by tens of millions remain under-served [3], [4], with double-digit accuracy gaps to English [5]–[7]. In Sri Lanka, where Sinhala and Tamil are the languages of administration, health and education, this gap decides whether an assistant is safe todeploy.

Because LLMs fabricate fluent but false content [8], [9], deployed systems increasingly wrap the model in an *abstention gate*: an uncertaintyscore$s$is computed for each candidate answer, and the systemanswers only when$s \leq \tau$, where$\tau$is the*abstention threshold*, otherwise deferring to a human or a fallback. Semantic entropy and its relativesmake$s$a genuinely useful signal [10]–[12], and conformal risk control (CRC) turns the choice of$\tau$into a statistical guarantee: given labelled calibration data, CRC returns a threshold whose expected loss provably respects a user-specified budget$\alpha$[13], [14].

That guarantee is stated for data exchangeable with the calibration set. Labelled calibration data—prompts paired with verified answer correctness—is expensive, and in Sinhala or Tamil it barely exists. The overwhelmingly practical response is to calibrate the gate once on English and reuse the threshold everywhere. This paper askswhether the guarantee survives that reuse, and finds that it does not.*Research gap:*no existing method tells a practitionerwhetheranEnglish-calibratedthresholdremains validinalow-resourcelanguage,orhowtorepairit withthe handful of target labels that can realistically be collected.We study this question analytically and in a controlled simulation; we do not run real Sinhala or Tamil LLM experiments in this paper.

The failure is not merely large; it is of a kind that the standard remedy cannot address. We give an exact decomposition(SectionIV)ofthecross-lingualriskgapinto a*covariate*term, arising because the score distribution$p(s)$moves between languages, and a*conditional*term, arising because$\eta(s) = \Pr(\text{correct} \mid s)$itself degrades—the same score value simply means something worse in the target language. Importance weighting, and with it weighted conformal prediction under covariate shift [15]–[17], is constructed preciselyto cancelthecovariatetermand leaves the conditional term untouched. Recent evidence that hallucination detectors transfer poorly across languages[18], [19] and that LLMs remain confident while becoming less accurate outside English [20] suggests the conditional term is dominant, which would make the standardmachinery ineffective here.

We confirm this in a controlled study whose generative parameters are fitted to effect sizes reported in the multilingual evaluation literature, and we propose a correction. We contribute (i) an exact risk-gap decomposition separating covariate from conditional shiftfor selective prediction, with the corollary that importance-weighting methods cancel only the

former (Section IV); (ii) aquantifiedfailuremode,inwhichanEnglish-calibrated

CRC gate overruns a10%risk budget by2.2×, violates itin every simulated deployment, and is98.6%beyond the reach of reweighting (SectionVII); (iii) PACT, a post-hoc correctorthatfitsatwo-parametercalibration-shiftmodelon a few target-language labels, transports it onto abundant*unlabelled*target data, and propagatesestimationuncertainty into a bootstrap upper confidence bound on the risk curve (SectionV); and (iv) an ablation with a design consequence—an intercept-only correction, modelling the accuracy drop alone, fails at50%violation, while adding a slope term for the decay of score informativeness succeedsat9%. Our evidence is a controlled simulation, not an LLM measurement campaign; Section VII-F gives the instantiation protocol for LLM-scale validation.

## II. BACKGROUND AND RELATED WORK

### *A. UncertaintyquantificationforLLMs*

Black-box uncertainty is typically estimated by sampling several continuations and measuring disagreement.Semantic entropy clusters samples by meaning before computing entropy, so paraphrases are not counted as disagreement [10], [11], and probes approximate it from a single forward pass [12]; related signals include self-consistency [21], [22] and self-evaluation [23], surveyed in [24], [25]. Recent work warns that entropy alone is an insufficient basis for safe selective prediction [26]. All of these produce a score; none fixes a threshold, and all inherit the classical observation that model confidence is miscalibrated [27].

detection transfer degrades, or supplies risk control that presumes exchangeability. To our knowledge no prior work treats*language*as the shift variable in a conformal risk-control problem, decomposes the resulting risk gap, orshows that the covariate-shift remedy is structurally inadequate for it.

## III. PROBLEM FORMULATION

Let a prompt $X$ drawn from a language distribution produce a generation from a fixed LLM, and let $Y \in \{0,1\}$ indicate whether that generation is correct. A scoring function assigns an uncertainty score $s = s(X) \in [0,1]$, larger meaning less certain. A *selective predictor* with threshold $\tau$ answers when $s \leq \tau$ and abstains otherwise.

We control the joint error rate

$$R(\tau) = \Pr(s \leq \tau \wedge Y = 0), \quad (1)$$

the probability of answering *and* being wrong. $R$ is non-decreasing in $\tau$ and bounded in $[0,1]$, so it is exactly the monotone bounded loss that CRC handles. Utility is the acceptance rate $A(\tau) = \Pr(s \leq \tau)$; the goal is the *largest* $\tau$ with $R(\tau) \leq \alpha$.

Write $p(s)$ for the density of the score and $\eta(s) = \Pr(Y = 1 \mid s)$ for the conditional accuracy, so that $R(\tau) = \int^{\tau} p(s)\,(1-\eta(s))\,ds$. We have a source language $\mathcal{S}$. (English) with $n_S$ labelled calibration items, and a target language $\mathcal{T}$ (e.g. Sinhala) with $n_U$ *unlabelled* items, which are cheap, and only $n_T \ll n_S$ labelled items.

Given calibration data $\{(s_i, y_i)\}_1^n$ , CRC returns

$$\hat{\tau} = \sup\{\tau: nR\hat{}_n(\tau) + 1/n + 1 \leq \alpha\}, \quad (2)$$

where $R\hat{}(\tau) = \frac{1}{n}\sum \mathbf{1}\{s \leq \tau, y\}$ $\mathbb{E}[R(\tau)] \leq \alpha$ when the deployment data are exchangeable

with the calibration data [13]. The question of this paper is what happens to $R_{\mathcal{T}}(\hat{\tau})$ when $\hat{\tau}$ is computed from $\mathcal{S}$.

### *B. MultilingualLLMsandthelow-resourcegap*

English-centric LLMs route non-English inputs through English-like internal representations [36], [37], so non-English behaviour is largely a projection of English competence. Accuracy falls sharply outside high-resource languages [5] and evaluation itself becomes unreliable there [38].Forabstention,cross-lingualconsensusandcausal

$\mathbb{E}[R(\hat{\tau})] \leq \alpha$whenthedeploymentdataareexchangeable

with the calibration data [13]. The question of this paper is what happens to$R_{\mathcal{T}}(\hat{\tau})$when$\hat{\tau}$is computed from$\mathcal{S}$.

## IV. DECOMPOSINGCROSS-LINGUALCALIBRATIONFAILURE

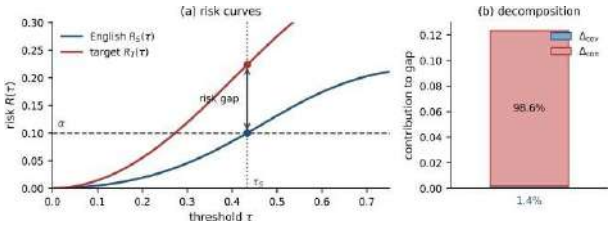



Fig. 1.*Cross-lingual risk gap. The English-calibrated threshold*$\hat{\tau}_S$*satisfies the budget in English but incurs*2.2×*the budget in the targetlanguage. Shading separates the covariate contribution (score distributionmoves) from the conditional contribution (the score means less), computedexactly from (3).*

**Theorem1(Risk-gapdecomposition).***Foreverythreshold*$\tau$,

methodshavebeenproposed[39],[40],andhallucination

$$R_{\mathcal{T}}(\tau) - R_{\mathcal{S}}(\tau) = \underbrace{\int_0^{\tau} (p_{\mathcal{T}} - p_{\mathcal{S}})(1-\eta_{\mathcal{S}})}_{\Delta_{\text{cov}}(\tau)} + \underbrace{\int_0^{\tau} p_{\mathcal{T}} (\eta_{\mathcal{T}} - \eta_{\mathcal{S}})}_{\Delta_{\text{con}}(\tau)}. \quad (3)$$

*Proof* .Add and subtract$\int_0 p_{\mathcal{T}}(s)(1-\eta_{\mathcal{S}}(s))ds$ in

$_{\tau}R(\tau) - R(\tau) = \int[p(1-\eta) - p(1-\eta)]$and group terms.

The decomposition is exact and requires no assumptions. In plain terms, the covariateter masks whether the model is

*uncertain about different questions*in the target language (t0he scores move), while the*conditional*term asks whether*the same uncertainty value now means a higher chance of being*

*wrong*(a score of0.3was right85%of the time in English but only70%in Sinhala). Its value is what itimplies about remedies.

**Corollary1(Reweightingisblindto$\Delta_{\text{con}}$).***Let$w(s)=p_{\mathcal{T}}(s)/p_{\mathcal{S}}(s)$and let$R^{-}(\tau)=\mathbb{E}_{\mathcal{S}}[w(s)\mathbf{1}\{s\le\tau\}(1-Y)]$be the importance-weighted source risk that weighted conformalpredictionestimates.Then$R^{-}(\tau)=R_{\mathcal{S}}(\tau)+\Delta_{\text{cov}}(\tau)$, so$R_{\mathcal{T}}(\tau)-R^{-}(\tau)=\Delta_{\text{con}}(\tau)$for all$\tau$.*

Thatis,weightedconformalpredictionremovesexactly

$\Delta_{\text{cov}}$and exactly nothing else. Whenever$|\Delta_{\text{con}}|\gg|\Delta_{\text{cov}}|$—which is what the multilingual literature leads us to expect, sincemodelsremainconfidentwhilelosingaccuracy

replicate refits only$(\theta_0,\theta_1)$with the English log-odds held fixed as an offset; and$s^{-}=1/2$centres the slope term. The importance-weighting baseline is implemented by transporting$\hat{\eta}_{\mathcal{S}}$onto the unlabelled target scores, which equals weighting each English item by$p_{\mathcal{T}}/p_{\mathcal{S}}$(Corollary1) withoutestimatingadensityratio.Fittingcosts$O(n_{\mathcal{S}}+$

$Bn_{\mathcal{T}})$plus$O(Bn_{\mathcal{U}})$;at$B$=200,$n_{\mathcal{T}}$=50,$n_{\mathcal{U}}$=2000this

ismillisecondsonaCPU,onceperdeployment,andadds

nothingatinferencetime.

### A. Whatisandisnotguaranteed

CRC'sguaranteeisfinite-sampleanddistribution-free; PACT'sboundisneither,beingasymptoticallyvalidunder

(4)andinheritingthat model's misspecificationerror, which SectionVIIprobesdirectly.Whereahardguaranteeis

required,thehybrid$\tau$^H=min($\tau$^PACT,$\tau$^$\mathcal{T}$)—thesecond

outsideEnglish[18],[20]—thestandardcovariate-shift repairisclosetoano-op.SectionVIIconfirmsthis

termbeing(2)onthe

CRC

$n_{\mathcal{T}}$targetlabels—satisfies

empirically.

Estimating$\Delta_{\text{con}}$

ispreciselywhatrequirestargetlabels:

$\mathbb{E}[R_{\mathcal{T}}(\tau$^H)]≤$\alpha$exactlybymonotonicityof$R_{\mathcal{T}}$.Wereportit asPACT-H.

$\Delta_{\text{cov}}$depends only on$p_{\mathcal{T}}$, which unlabelled text determines, whereas$\Delta_{\text{con}}$involves$\eta_{\mathcal{T}}$. The design question is how to spend a very small labelling budget on$\eta_{\mathcal{T}}$.

## V. PACT

PACT answers that question by refusing to estimate$\eta_{\mathcal{T}}$nonparametrically. Instead it models the target conditionalas a low-dimensional*calibration shift*of the source conditional, on the log-odds scale:

$$\text{logit}\,\eta_{\mathcal{T}}(s)=\text{logit}\,\eta_{\mathcal{S}}(s)+\theta_0+\theta_1(s-s^{-}).\quad(4)$$

Here$\theta_0$absorbstheuniformaccuracydropand$\theta_1$

absorbs the*change in how informative the score is*:$\theta_1>0$flattens the conditional, which is exactly a fall in the detector's AUROC. Two parameters suffice to express both documented failure modes, and two parameters are cheap toestimate.

I. Algorithm1:PACT

1. **Input:**sourcelabelled$\{(s_i,y_i)\}^{n_{\mathcal{S}}}$;targetunlabelled $\{s_j\}^{n_{\mathcal{U}}}$; target labelled $\{(s_k,y_k)\}^{n_{\mathcal{T}}}$; budget $\alpha$;confidence$\delta$; replicates$B$
2. Fit$\eta$^$\mathcal{S}$bylogisticregressiononthesourcelabelledset
3. $o_k\leftarrow\text{logit}\,\eta$^$\mathcal{S}(s_k)$foreachtargetlabelleditem//offset
4. $b$=1to$B$
5. Resamplethe$n_{\mathcal{T}}$targetlabelleditemswithreplacement
6. $\theta^{(b)}\leftarrow$logisticfitof(4)withfixedoffset$o$
7. $\eta$^$(b)(s)\leftarrow\sigma(\text{logit}\,\eta$^$(s)+\theta$^$(b)+\theta$^$(b)(s-s^{-}))$

## VI. CONTROLLEDSIMULATIONSTUDY

### A. Whysimulation

Measuring*violation rates*requires repeating an entire calibrate-and-deploy cycle hundreds of times under known ground truth, which no single pass over an LLM benchmark provides. We therefore fit the simulator to published effect sizes and claim no measured LLM numbers.

### B. Data

No external dataset is used and no LLM is queried: every item is synthetic, drawn from the generative model below,so the study isolates the calibration procedure from any one model, scorer or benchmark. Each item is a pair$(s,y)$: an uncertaintyscore$s\in[0,1]$andacorrectnesslabel$y\in\{0,1\}$.

### C. Generativemodel

Scores follow Beta marginals and the conditionals are logistic,logit$\eta_{\mathcal{S}}(s)=\beta_0+\beta_1$sandlogit$\eta_{\mathcal{T}}(s)=\beta_0+\beta_1 s+D_0+D_1(s-1/2)$.Wesolvedfor$(\beta_0,\beta_1,D_0,D_1)$so thetwolanguagesreproducereportedeffectsizes:English

error rate0.220with detector AUROC0.820and mean score0.286; target error rate0.420with AUROC0.700and mean score0.346. The12-point AUROC drop matches the magnitudeofcross-lingual detectordegradationreported in [18], [19], and the mild+0.06shift in mean scoreencodes the overconfidence documented in [20]: the model becomesmuchlessaccuratewhilebarelybecomingless

8. $R^{(b)}(\tau)\leftarrow\frac{1}{n_{\mathcal{U}}}\sum_j\mathbf{1}\{s_j\le\tau\}(1-\eta_{\mathcal{T}}^{(b)}(s_j))$

confident.Thesolvedvaluesare$s\sim\text{Beta}(1.602,3.998)$ (English)and$\text{Beta}(1.938,3.662)$(target),$\beta$=3.863,

9. $R^{-}(\tau)\leftarrow$pointwise$(1-\delta)$quantileof$\{R^{(b)}(\tau)$

0

return

$\tau \triangleq \sup\{\tau : R^{-}(\tau) \le \alpha\}$

$b=1$

$\beta_1 = -7.667, \quad D_0 = -0.313, \quad D_1 = 3.527.$ These areanchoredtothe*directionandmagnitude*ofpublished

Algorithm1 makes the division of labour explicit: the shape of the risk curve over the target population is computed on$n_U$unlabelled items, so itischeap andprecise, while the$n_T$labels are spent only on the two shift parameters and the bootstrap propagates their sampling error into apointwise upper confidence bound on the wholecurve [46]. Monotonicity makes thresholding that bound conservative. Concretely, both logistic fits use Newton's method witha$10^{-3}$ ridgepenalty (which keeps the fitst able when the $n_T$ label sarenearly separable); each bootstrap effects,not fitted to any one model's raw outputs; Section VII varies them.

### *D. Protocol*

Each trial draws$n_S$= 2000labelled source items, $n_U$= 2000unlabelled target items and$n_T$labelled target items, runs every method, and evaluates the resulting threshold against the*exact*target risk and acceptance computed by quadrature Weuse$\alpha$ = 0.10, $\delta$ = 0.05, $B$ = 200and1000trials (500 for sweeps), seed20260814. We report 95% intervals: normal intervals formean risk and acceptance, Wilson intervals for violation rates. The simulator and every baseline are a single Python script (NumPy/SciPy), available from the corresponding author.

### *E. Methods*

SOURCE-ONLYCR Capplies (2) on English and deploys unchanged (current practice).TARGET-ONLYCRCapplies (2) to the$n_T$ target labels alone. RE WEIGHT-ONLYtransports $\hat{\eta}_S$onto the unlabelled target scores, correcting$\Delta_{cov}$and nothing else. PACT-1 restricts (4) to$\theta_0$; PACT-2 is the full method. TARGET-PLATTfitsafreshtwo-parameterlogistic recalibration on the$n_T$target labels alone and applies the same bootstrap bound, isolating the value of PACT'sEnglish prior. ORACLEknows$\eta_T$exactly and marks the utility ceiling.

TABLE I
CROSS-LINGUAL SELECTIVE PREDICTION,$\alpha = 0.10$,$n_T = 50$LABELLED TARGET ITEMS, 1000 TRIALS. RISK IS THE REALISED JOINT ERROR RATE;VIOLATIONISTHEFRACTIONOFDEPLOYMENTSEXCEEDING$\alpha$(95%WILSONI NTERVAL).

| Method | Risk | Violation[95%CI] | Accept | Sel.err. |
|---|---|---|---|---|
| Source-onlyCRC | 0.222 | 1.000 [0.996, 1.000] | 0.687 | 0.323 |
| Reweight-only | 0.220 | 1.000 [0.996, 1.000] | 0.684 | 0.322 |
| Target-onlyCRC | 0.098 | 0.457 [0.426, 0.488] | 0.379 | 0.250 |
| Target-Platt | 0.065 | 0.081 [0.066, 0.100] | 0.281 | 0.228 |
| PACT-1 (intercept) | 0.103 | 0.501 [0.470, 0.532] | 0.400 | 0.255 |
| **PACT-2 (ours)** | **0.071** | **0.090** [0.074, 0.109] | 0.299 | 0.232 |
| PACT-H (hybrid) | 0.068 | 0.089 [0.073, 0.108] | 0.289 | 0.230 |
| Oracle(utilityceiling) | 0.099 | 0.388 [0.358, 0.419] | 0.392 | 0.253 |

*Riskandacceptanceintervalsareallwithin*±0.008.

### *F. Metrics*

Mean realised target risk; *violation rate*, the fraction of trials with$R_T(\tau) > \alpha$; and acceptance rate. A method is useful only if it controls risk*and*answers often. Oracle targets meanriskexactly$\alpha$, so its violationrate is near50%by construction: it is a utility reference, not a validity one.

## VII. RESULTS

### *A. Thegapisalmostentirelyconditional*

Before comparing methods we measure thedecomposition itself, which Theorem1 lets us compute exactly. At the English-calibrated threshold$\tau_S$= 0.433, the risk is0.100in English by construction and0.223in the target language, agapof0.123. Ofthatgap,$\Delta_{cov}$= +0.0017and$\Delta_{con}$= +0.1217: the covariate term accounts for**1.4**%ofthefailureandtheconditionalterm for**98.6**%

(Fig.1). Under the effect sizes reported in the multilingual literature, cross-lingual calibration failure is a conditional-shift phenomenon almost to the exclusion of anything else. By Corollary1 this caps what any reweighting method can recover at1.4%of the excess risk, before it is evenimplemented.

This is not an artefact of the particular shift we chose. Holding the target conditional fixed and sweeping the covariate shift from zero to four times our nominal value,$\Delta_{con}$stays adverse and dominant throughout (0.055to0.138), while$\Delta_{cov}$ranges only over$[-0.037,+0.002]$;$\tilde{R}$stays below the true target risk at every setting and below$\alpha$at five of seven, so reweighting can argue for an*even more permissive*threshold. No setting in this family lets$\Delta_{cov}$-correction alone suffice.

### *B. English-calibratedgatesfail,andreweightingdoesnothelp*

TableI is the central result. The English-calibrated gate realisesatargetriskof0.222againstabudgetof0.100—a

2.2 ×overrun—and exceeds the budget in100%of deployments. It answers a great deal (68.7%), which is precisely why the failure is dangerous: the system looks confident and productive while silently emitting more than twice the contracted error rate.

The second row is Corollary1 made concrete: correcting the covariate term moves realised risk from0.222to0.220, recovering1.6%of the excess, within sampling error of the1.4%ceiling predicted analytically. Weighted conformal prediction is provably capped here, and the cap is tight.

### *C. Correctingtheconditionaltermrestorescontrol*

PACT-2 attains a violationrate of0.090(95% CI0.074–0.109),nearitsnominal$\delta = 0.05$andanorderof magnitude better than the gate's1.000, while stillanswering29.9%of queries—76%of the oracle's39.2%. PACT-H costs little extra:0.089violation at28.9%acceptance,withanexactexpectationguarantee.Target-only CRC is well calibrated*in expectation*($0.098 \leq \alpha$) yet violates in46%of individual deployments, because 50 labels give a very noisy empirical risk curve; an expectation guarantee averaged over hypothetical redeployments issmall comfort to the one deployment a ministry actuallyruns.

The PACT-1 row isolates the sharpest design lesson. An intercept-onlycorrection—thenaturalfirstguess,andwhata simple accuracy-gap adjustment implements—violates at0.501,nobetterthanignoringEnglishentirely.Onlywhen

$\theta_1$is admitted, letting the model express that the score has become*less discriminative*, does control return. It is not enough to know the model is less accurate in Sinhala; one must knowits confidence signal is less trustworthy, and that is what the slope term encodes.

The Target-Platt row tempers the role of the Englishprior. A two-parameter recalibration on the target labels alone, with the same bootstrap bound, reaches similarcontrol (0.081) at slightly lower utility (28.1%). In a 500-triallabelsweeptheprior'sadvantageisinutilityatthe

smallest budgets: at$n_T = 10$and25PACT-2 answers29.2%and28.9%of queries against25.2%and25.4%for Target-Platt, at comparable violation (0.206/0.120vs0.178/0.108); by$n_T \geq 100$the two coincide. What restores control is therefore the slope-bearing conditional model transported onto unlabelled target data with a bootstrap bound; the English prior buys about15%more answering when labels are scarcest.

### *D. Labelefficiency*

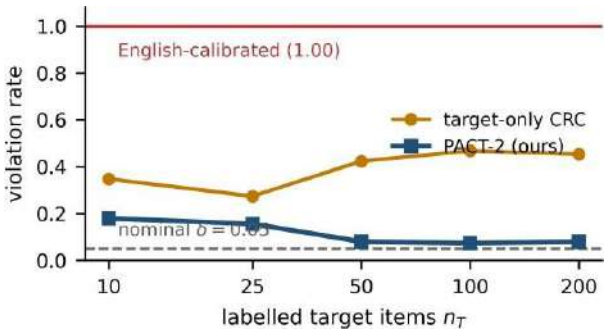



Fig. 2.*Label efficiency. PACT-2 is below*0.2*violation from ten labelledtarget items and near nominal from fifty, while target-only CRC neverapproaches control because its noise does not shrink fast enough at thesesample sizes. Acceptance rises from*0.286*to*0.328*across the samerange, so the control is not bought by abstaining from everything.*

Figure2 shows PACT reaching useful control almost immediately:its violationrate is0.180at$n_T = 10$,0.156at25and0.080at50, and stays at0.07–0.08up to$n_T = 200$while acceptance climbs from0.286to0.328. Target-only CRC sits between0.27and0.47throughout and does not improve monotonically—more labels raise its threshold toward the optimum, increasing utility and the chance of overshooting together. Twenty-five to fifty verified items is a realistic annotation task; two thousand is not.

### *E. Robustnesstomodelmisspecification*

PACT assumes the conditional shift is affine inson the log-odds scale. To probe the cost of that assumption we add aquadraticterm$D_2(s - 1/2)^2$to thetruetargetconditional, whichPACT-2cannotrepresent.Sweeping$D_2 \in$

$\{-6,-3,0,+3,+6\}$at$n_T$=50(300trialseach)gives

violationrates0.427,0.107,0.093,0.057,0.100with

acceptance rising monotonically from0.255to0.376: the exposureisstrongly*one-sided*.Positivecurvature makesthe target better than the affine fit predicts where items are accepted,andcontrolstaysina0.06–0.10bandwhileutility improves. Negative curvature is the dangerous direction, mildatfirst(0.107at$D_2 = -3$against0.093well-specified) but unbounded: at$D_2 = -6$violation reaches0.427, comparable to target-only CRC (0.41–0.46). Severe unmodelled curvature thus erases PACT's advantage over the label-only baseline, though both stay far below thegate's1.000. Where it is suspected and the guarantee is contractual, PACT-H is the right choice: its expectation bound survives misspecification by construction. The affine familyalso cannot expressshifts that differ byquestiontype or domain; with a larger label budget the same machinery accepts a richer shift (e.g. a spline inson per-domain intercepts),atthepriceofmorelabelsperparameter.A

practical check is a calibration-curve test of the fitted$\eta \hat{} \mathcal{T}$on held-out target labels; a failed test should trigger PACT-H.

### *F. InstantiationforanLLM-scalestudy*

The abstract scoresinstantiates as follows. Per prompt, sample$m$generations at non-zero temperature and compute semanticentropy[10],[11].Itsclusteringstepnormallyuses a natural-language-inference model for bidirectional entailment, the weak link for Sinhala and Tamil where no reliable NLI model exists; we instead cluster in a language-agnostic embedding space [47], [48], consistent with evidence that such

models carry language-agnosticsemantics in deeper layers [19], [36]. Target labels come from human verification: items sampled across the score range, each answer marked correct or incorrect by two native-speakerannotatorsagainstareference,disagreements adjudicated by a third, and agreement (Cohen's$\kappa$) reported; with SinhalaMMLU [6], SinhaLegal [45] and NSINA-derived tasks [43] as prompt pools and MMLU [49] on the English side. The calibration procedure is unchanged: the harness takes a two-column table of$(s,y)$per language, so any scorer substitutes directly, and because calibration is a one-off per-deployment cost the pipeline suits the small, edge-deployed models that low-resource settings actually run. Multilingual abstention methods [39], [40] produce a score$s$rather than a calibrated threshold, so they are complementstoPACT,notcompetitors:PACTcancalibrate any of them.

## VIII. DISCUSSIONANDLIMITATIONS

### *A. Scopeofevidence*

Two kinds of result appear here and they should not be confused.Theorem1andCorollary1are mathematical facts that hold for any model, language and scorer. Every number in SectionVII is a simulation result: it validates a*calibration procedure*under a generative model fitted to published effectsizes, doesnotmeasureanyparticular LLM inSinhalaorTamil,anditsmagnitude will varywithmodel, task and scorer. Real-LLM validation with human-verified Sinhala and Tamil labels (Section VII-F) is the necessary next step and was not feasible within this revision. The decomposition, not the simulated magnitudes, is what grounds our recommendation against covariate-onlyremedies.

### *B. Weakerguarantee,minimalmodel*

PACT trades CRC's distribution-free finite-sample validity for a parametric asymptotic bound; where a hard guarantee is mandated, PACT-H recovers the exact expectation bound at a small utility cost. The affine shift family is deliberately minimal, and the sweep above prices that choice. Extending finite-sample validity to conditional shift, and selecting the shift family from the label budget,are both open.

### *C. Beyondlanguage,andequity*

Nothing in the derivation is specific to language; domain, dialect and time plausibly induce conditional shift too. A gate that abstains more often in Sinhala than in English delivers unequal service; PACT makes that inequalityvisible and budgeted, not absent, and it should be disclosed to users [50], [51].

## IX. CONCLUSION

Calibrating an LLM abstention gate in English and deploying it in a low-resource language voids the risk guarantee, overrunning a10%budget by2.2×and failingin every simulated deployment. We showed why the standard covariate-shift repair cannot fix this—it cancels a term that is98.6%beside the point—and that a two-parameter conditional-shift correction fitted to as few as twenty-fiveverifiedtarget-languageitemsrestorescontrolin simulation while keeping most achievable utility. The necessary ingredient is the slope term: recalibrating for lost accuracy is not enough, one must recalibrate for lost*informativeness*. Real-LLM validation in Sinhala and Tamil is the next step.

## ACKNOWLEDGMENT

The authors thank the Faculty of Computing, NSBM Green University.

## REFERENCES

[1] T. B. Brown et al., "Language Models are Few-Shot Learners,"arXiv:2005.14165,2020.

[2] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned ChatModels," arXiv:2307.09288, 2023.

[3] P. Joshi et al., "The State and Fate of Linguistic Diversity andInclusion in the NLP World," arXiv:2004.09095, 2020.

[4] K. Huang et al., "A Survey on Large Language Models withMultilingualism: Recent Advances and New Frontiers,"arXiv:2405.10936,2024.

[5] W. Xuan et al., "MMLU-ProX: A Multilingual Benchmark forAdvanced Large Language Model Evaluation," arXiv:2503.10497,2025.

[6] A. Pramodya et al., "SinhalaMMLU: A Comprehensive Benchmarkfor Evaluating Multitask Language Understanding in Sinhala,"arXiv:2509.03162,2025.

[7] S. Kishanthan et al., "Large Language Models for Math Education inLow-Resource Languages: A Study in Sinhala and Tamil,"arXiv:2602.14517,2026.

[8] Z. Ji et al., "Survey of Hallucination in Natural LanguageGeneration," arXiv:2202.03629, 2022.

[9] A. Alansari and H. Luqman, "Large Language Models Hallucination:A Comprehensive Survey," arXiv:2510.06265, 2025.

[10] S. Farquhar et al., "Detecting hallucinations in large language modelsusing semantic entropy," Nature, vol. 630, no. 8017, pp. 625–630,2024.

[11] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic Uncertainty: LinguisticInvariances for Uncertainty Estimation in Natural LanguageGeneration," arXiv:2302.09664, 2023.

[12] J. Kossen et al., "Semantic Entropy Probes: Robust and CheapHallucination Detection in LLMs," arXiv:2406.15927, 2024.

[13] A. N. Angelopoulos et al., "Conformal Risk Control,"arXiv:2208.02814,2022.

[14] ——,"Learn then Test:CalibratingPredictiveAlgorithmstoAchieveRisk Control," arXiv:2110.01052, 2021.

[15] R. J. Tibshirani et al., "Conformal Prediction Under Covariate Shift,"arXiv:1904.06019,2019.

[16] A. Bhattacharyya and R. F. Barber, "Group-Weighted ConformalPrediction," arXiv:2401.17452, 2024.

[17] D. Prinster et al.,"Conformal Validity Guarantees Existfor AnyDataDistribution (and How to Find Them)," arXiv:2405.06627, 2024.

[18] D. Datta et al., "Do LLM hallucination detectors suffer from low-resource effect?" arXiv:2601.16766, 2026.

[19] A. Alansari, M. Alkhorasani, and H. Luqman, "CrossHallu: DoHallucination Signals Generalize Across Languages and Domains inLarge Language Model's Internals?" arXiv:2607.04029, 2026.

[20] E. A. Chimoto, M. Elhoushi, and B. A. Bassett, "Calibrating BeyondEnglish: Language Diversity for Better Quantized MultilingualLLM," arXiv:2601.18306, 2026.

[21] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative LargeLanguage Models," arXiv:2303.08896, 2023.

[22] X. Wang et al., "Self-Consistency Improves Chain of ThoughtReasoning in Language Models," arXiv:2203.11171, 2022.

[23] S. Kadavath et al., "Language Models (Mostly) Know What TheyKnow," arXiv:2207.05221, 2022.

[24] J. Genget al., "ASurveyofConfidenceEstimationandCalibration inLarge

Language Models," arXiv:2311.08298, 2023.

[25] S. Kang et al., "Uncertainty Quantification for HallucinationDetection in Large Language Models: Foundations, Methodology,and Future Directions," arXiv:2510.12040, 2025.

[26] E. Phillips et al., "Entropy Alone is Insufficient for Safe SelectivePrediction in LLMs," arXiv:2603.21172, 2026.

[27] C. Guo et al., "On Calibration of Modern Neural Networks,"arXiv:1706.04599,2017.

[28] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in aRandom World. New York, NY, USA: Springer, 2005.

[29] A. N. Angelopoulos and S. Bates, "A Gentle Introduction toConformal Prediction and Distribution-Free UncertaintyQuantification," arXiv:2107.07511, 2021.

[30] V.Quachetal.,"ConformalLanguageModeling,"arXiv:2306.10193,2023 .

[31] Y. A. Yadkori et al., "Mitigating LLM Hallucinations via ConformalAbstention," arXiv:2405.01563, 2024.

[32] Y. Xu, W. Guo, and Z. Wei, "Selective Conformal Risk Control,"arXiv:2512.12844,2025.

[33] M. Salem et al., "Hierarchical Group-Conditional Conformal RiskControl for Selective Prediction in Language Models,"arXiv:2607.24562,2026.

[34] J. Opoku and D. Banahene, "ToolChain-CRC: Conformal RiskControl for Agentic AI Under Retrieval and Tool-Use Drift,"arXiv:2606.18467,2026.

[35] R. El-Yaniv and Y. Wiener, "On the foundations of noise-freeselective classification," Journal of Machine Learning Research, vol.11, pp. 1605–1641, 2010.

[36] C. Wendler et al., "Do Llamas Work in English? On the LatentLanguage of Multilingual Transformers," arXiv:2402.10588, 2024.

[37] M. Wang et al., "Lost in Multilinguality: Dissecting Cross-lingualFactual Inconsistency in Transformer Language Models,"arXiv:2504.04264,2025.

[38] E. Zhou et al., "Lower-Resource, Higher Scores: Language Bias inLLM Evaluators," arXiv:2607.14480, 2026.

[39] S. Duwal, "MKA: Leveraging Cross-Lingual Consensus for ModelAbstention," arXiv:2503.23687,S 2025.

[40] Y. Sun et al., "CausalAbstain: Enhancing Multilingual LLMs withCausal Reasoning for Trustworthy Abstention," arXiv:2506.00519,2025.

[41] W. Zheng et al., "CCL-XCoT: An Efficient Cross-LingualKnowledge Transfer Method for Mitigating HallucinationGeneration," arXiv:2507.14239, 2025.

[42] J. Huang et al., "Investigating the Multilingual Calibration Effects ofLanguage Model Instruction-Tuning," arXiv:2601.01362, 2026.

[43] H. Hettiarachchi et al., "NSINA: A News Corpus for Sinhala,"arXiv:2403.16571,2024.

[44] H. W. K. Aravinda et al., "SinLlama – A Large Language Model forSinhala," arXiv:2508.09115, 2025.

[45] M.LasandiandN.Jayatilleke,"SinhaLegal:ABenchmarkCorpusforInformation Extraction and Analysis in Sinhala Legislative Texts,"arXiv:2603.04854,2026.

[46] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. NewYork, NY, USA: Chapman & Hall/CRC, 1993.

[47] F. Feng et al., "Language-agnostic BERT Sentence Embedding,"arXiv:2007.01852,2020.

[48] A. Conneau et al., "Unsupervised Cross-lingual RepresentationLearning at Scale," arXiv:1911.02116, 2019.

[49] D. Hendrycks et al., "Measuring Massive Multitask LanguageUnderstanding," arXiv:2009.03300, 2020.

[50] A. Biswas and W. Talukdar, "Guardrails for trust, safety, and ethicaldevelopment and deployment of Large Language Models (LLM),"arXiv:2601.14298,2026.

[51] B. Torkian and J. Zhou, "Designing Safety-Constrained LLMSystems for Public Health Information Access," arXiv:2607.13038,2026.

# AI-Powered Smart Passenger Travel Assistant: A Full Design, Implementation and Evaluation Study of Machine Learning Flight Delay Prediction, Route Guidance, and Sleep-Onset Boarding Alerts

H. A. Javindi Nethnika Heshani
*Department of Software Engineering, Faculty of Computing*
*NSBM Green University*
*Colombo, Sri Lanka*
ajnheshani@students.nsbm.ac.lk

Hirushi Dilpriya Thilakarathna
*Department of Computer and Data Science, Faculty of Computing*
*NSBM Green University*
*Colombo, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Passengers manage flight status, route planning, weather and their own alertness in separate applications, and delay forecasts rarely reach them as actionable guidance. This paper presents the design, implementation and evaluation of an AI-powered smart passenger travel assistant, built under Design Science Research, that integrates flight-delay prediction, traffic-aware departure guidance, weather alerts, an LLM travel assistant and wearable sleep-onset boarding alerts in one Flutter application backed by one Flask service. A custom ESP32 wearable (MAX30102 heart rate, MPU6050 motion) streams data over Bluetooth Low Energy to the phone, where a rule-based classifier triggers a boarding alert and a vibration wake-up. A Random Forest delay classifier, trained on 2024 U.S. BTS and METAR data for 30 airports and scored on a chronological test set of 744,968 flights, reached macro-F1 0.582, AUC-ROC 0.652, delayed-class precision 0.309 and recall 0.329, against a majority-class baseline of macro-F1 0.451 and AUC-ROC 0.500; accuracy (0.749) is below the 0.821 baseline and is not a headline metric. All ten functional tests passed, a defect found in testing was fixed (push delivery 1/15 to 15/15), and eleven participants gave a mean SUS of 63.4 with 87.9% task completion. The model does not serve Colombo departures, the sleep classifier is not yet validated and route guidance used simulated traffic; the results support feasibility of the integration, not deployment readiness.**

***Keywords—flight delay prediction, class imbalance, Design Science Research, wearable sleep detection, ESP32, Bluetooth Low Energy, mobile application***

## I. Introduction

### A. Background and Problem

Air travel is time-sensitive, yet passengers coordinate flight status, airport travel, weather and their own readiness with separate tools. Delays are costly: an FAA-commissioned study put U.S. air-transport delay costs at USD 32.9 billion for 2007 [1], and Airlines for America reports about USD 33 billion for 2019 [2]. In our test partition (Section III-D), 17.88% of BTS flights [3] were over 15 minutes late. Airline applications report a delay only after it is confirmed, and no widely available system combines delay risk, traffic-aware departure guidance and passenger alertness.

Falling asleep before boarding is a further risk. In a formative online survey of 62 Sri Lankan respondents (February–March 2026, open link, not a random sample), 41 (66.1%) had missed a bus or train stop, 11 of them (26.8%) mainly because they fell asleep, and 54 (87.1%) fall asleep on journeys at least sometimes. The survey concerned buses and trains, so it supports the plausibility of the need, not its prevalence among air passengers.

### B. Problem Statement

Four gaps motivate this work. *Gap 1:* delay information is reactive, and most prediction studies never reach a passenger-facing tool. *Gap 2:* no system combines a flight deadline with live traffic to compute the latest time to leave home. *Gap 3:* wearable sleep sensing has not been applied to boarding alerts. *Gap 4:* no travel assistant combines a mobile app, a prediction backend, live flight, weather and maps data, and physiological sensing in one architecture.

### C. Research Questions

The primary question is how an integrated travel assistant combining delay prediction, traffic-aware route guidance, wearable sleep sensing and personalised alerts can be designed, implemented and evaluated to reduce passenger inconvenience. Table I decomposes it and states how far each sub-question is answered.

TABLE I. Research Sub-Questions and Their Status

| # | Sub-question | Status |
|---|---|---|
| RQ1 | Which of Logistic Regression, Random Forest and XGBoost gives the best minority-class-aware performance (macro-F1) for delay classification under a chronological split? | Partly answered (§V-A) |
| RQ2 | How can heart-rate and motion data from a custom ESP32 wearable over BLE be processed to classify sleep with enough sensitivity to trigger alerts? | Open (§V-B) |
| RQ3 | How should the app and backend be architected to integrate flight, weather and maps data with low latency and battery cost? | Partly answered (§V-D) |
| RQ4 | Which metrics best characterise integrated-system performance? | Partly answered (§V, VI) |

### D. Contributions

No new learning algorithm is proposed. The contributions are (i) a three-tier Flutter/Flask architecture with an ESP32 wearable peripheral, nine modules, and thirteen functional and seven non-functional requirements; (ii) a chronologically evaluated delay classifier (744,968 test flights) benchmarked against a majority-class baseline, with macro-F1 and AUC-ROC as headline metrics; (iii) an implemented wearable sleep-onset alert pipeline with a documented validation protocol, for which no sensitivity or specificity is yet claimed; (iv) functional, non-functional and usability results, including a push-notification defect found and fixed; and (v) documented design iterations in sensing hardware and backend structure.

### E. Scope and Claim Boundaries

Only the delay module is a trained model; the other decision modules are deterministic calculations, and sleep detection is a rule-based threshold classifier. We do not claim that the system reduces missed flights, that its delay predictions justify changing travel plans, or that the sleep classifier is accurate (Section VI).

## II. Related Work

### A. Flight Delay Prediction

Machine-learning delay prediction has been studied extensively. Kalyani et al. [4] trained XGBoost and linear regression on U.S. flight and weather data; Gui et al. [5] compared models on ADS-B, weather, schedule and airport data (90.2% binary accuracy, Random Forest); Hu et al. [6] used a support vector machine on flight-plan information; Anees and Huang [7] trained a Random Forest on U.S. data; Dou [8] reported 80.44% accuracy with CatBoost; and Demirkaya et al. [9] reported $R^2$ of 91.9% (XGBoost) and found that weather observations improved accuracy, supporting our weather features. Most stop at model evaluation; two [7], [9] add a web front end but no departure guidance or wearable alerting. These figures are not comparable with ours (different task, data, class balance and split). Our features use only information known at scheduled departure and our test set is chronological (Section III-D); we claim no superiority and benchmark against a trivial baseline (Section V-A).

### B. Travel Assistants, Wearables and Passenger Flow

Travel Assist [10] resolves Indian travel queries with an intent classifier and real-time APIs, and Suresh et al. [11] combine weather, generative-AI itineraries, crowd density and a chatbot; neither predicts delays or senses the passenger. Commercial tools such as TripIt Pro notify delays only after confirmation. On the sensing side, the HOT Watch [12] streams multi-sensor health data to a phone over Bluetooth, Anjum et al. [13] survey wearable sensing over BLE, and Wu et al. [14] validate a BLE sensor patch against a commercial reference. Zouein et al. [15] optimise boarding at airport-operator level; we take the individual passenger's view. Our wearable reports only heart rate and motion and classifies sleep state rather than assessing health.

### C. Positioning

Table II shows that, among the works reviewed, none combines all five capabilities.

TABLE II. Capability Comparison with Related Work

| System | Delay | Route | Wearable | Sleep alert | Unified app |
|---|---|---|---|---|---|
| Kalyani et al. [4] | Yes | No | No | No | No |
| Gui et al. [5] | Yes | No | No | No | No |
| Hu et al. [6] | Yes | No | No | No | No |
| Anees and Huang [7] | Yes | No | No | No | Part. |
| Dou [8] | Yes | No | No | No | No |
| Demirkaya et al. [9] | Yes | No | No | No | Part. |
| Dheemonth et al. [10] | No | Part. | No | No | Part. |
| Suresh et al. [11] | No | Part. | No | No | Part. |
| Madavarapu et al. [12] | No | No | Yes | No | No |
| Zouein et al. [15] | No | No | No | No | No |
| **Proposed system** | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** |

Part. = partial ([7], [9]: web front end for the model only; [10], [11]: travel information without deadline-based guidance).

The contribution is thus a combination, not a method: each ingredient is established, but none of the reviewed works feeds a delay forecast, a deadline-based leave-home calculation and a wearable sleep state into a single passenger alert. We claim no algorithmic novelty.

## III. System Design and Methodology

### A. Design Science Research

The work follows the Design Science Research Methodology (DSRM) of Peffers et al. [16]: problem identification (the four gaps), objective definition, design and development (nine modules, two iterations), demonstration (a simulated passenger journey), evaluation and communication. Evidence is matched to module type: held-out testing for the trained model, labelled-episode validation for the sleep classifier (planned, Section III-E), scenario tests and measurements for deterministic modules, and a usability study.

### B. Final Architecture

The deployed system is a three-tier client–server design (Fig. 1). Tier 1 is the Flutter application (Riverpod, flutter_blue_plus for BLE, Firebase for authentication and push messaging); Tier 2 is a single Flask service hosting the decision-logic modules and the trained delay model; Tier 3 comprises external flight-status (AviationStack), weather (OpenWeatherMap), maps (Google) and LLM services. The ESP32 wearable is a peripheral of Tier 1: it has no network path and stores no API key.

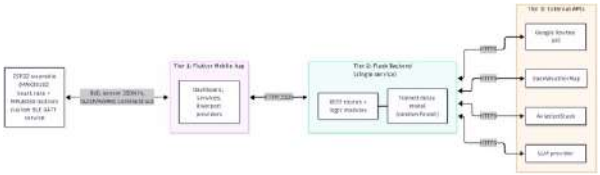


Fig. 1. Final three-tier architecture with the ESP32 wearable as a peripheral of the mobile app.

### C. Design Iterations

Two decisions changed during development and are treated as DSRM iterations. *Sensing hardware.* Motion first came from the phone's accelerometer and heart rate from a commercial BLE band, because commercial wearables expose no raw motion over standard BLE services. This was reversed once an ESP32-WROOM rig (MAX30102, MPU6050, vibration motor) gave both signals over one link; the thresholds carried over because the movement scale was kept identical. *Backend.* Separate Flask services and an orphaned FastAPI sleep service were consolidated into one Flask application.

### D. Delay-Prediction Data, Training and Metrics

*Data.* We use 2024 U.S. Bureau of Transportation Statistics (BTS) on-time data [3] for the 30 busiest airports, joined to Iowa Environmental Mesonet METAR observations [17] by origin airport and scheduled UTC departure time (nearest observation within one hour). The target is binary: more than 15 minutes late. Features use only information known at scheduled departure: UTC hour, day of week, month, one-hot carrier, three historical delay rates (origin, route, route over the previous 30 days) and five weather fields (temperature, dew point, wind speed, visibility, precipitation). Post-departure fields are excluded to prevent leakage; delays above the 99th percentile are clipped and missing weather values are filled with training medians.

*Why U.S. data.* To our knowledge BTS [3] is the only open per-flight record at this scale joinable to open METAR data [17]; no comparable open dataset exists for Colombo departures. The U.S. model therefore demonstrates the pipeline (leakage-free features, chronological evaluation, serving endpoint, baseline comparison), not delay behaviour at Colombo; it can be retrained on local records, and until then the endpoint returns unsupported_airport for non-U.S. origins (TC2).

*Split and models.* The split is chronological: training before 1 September 2024, validation September–October, test from 1 November 2024, used once, for the selected model only. Logistic Regression, Random Forest [18] and XGBoost [19] were tuned with scikit-learn [20]; Logistic Regression and Random Forest used balanced class weights. Within each family, hyperparameters were chosen by five-fold grid search scored by macro-F1 on a stratified sample of 1,000,000 training rows (seed 42); as these folds are not chronological, they only ranked settings within a family. The families were then compared on the chronological validation partition (Section V-A). The grids held three, four and four combinations respectively, the XGBoost grid being limited by memory and compute. A parallel Random Forest regressor predicts delay minutes.

*Why macro-F1.* About 82% of flights are on time, so accuracy rewards a classifier that never predicts a delay. Selection and headline reporting therefore use macro-F1 (the unweighted mean of per-class F1, which cannot be raised by ignoring the minority class), supported by AUC-ROC (threshold-independent ranking quality) and delayed-class precision and recall. Balanced accuracy and the Matthews correlation coefficient (Table III) cannot be inflated by a majority-class predictor. Accuracy is shown only beside the 0.821 majority-class baseline, never as evidence of skill.

### E. Wearable Sleep Detection

*Wearable.* The ESP32-WROOM-32 reads heart rate (MAX30102) and acceleration (MPU6050) on separate I²C buses and drives a vibration motor (Fig. 2). A custom BLE GATT server notifies a JSON payload {bpm, ax, ay, az, mpu_ok} about every 500 ms and accepts SLEEP, AWAKE, WAKE and NORMAL commands. The firmware performs no classification and transmits only while a finger is on the MAX30102; despite its firmware name (ESP32-HealthSensor), it reports only heart rate and motion.

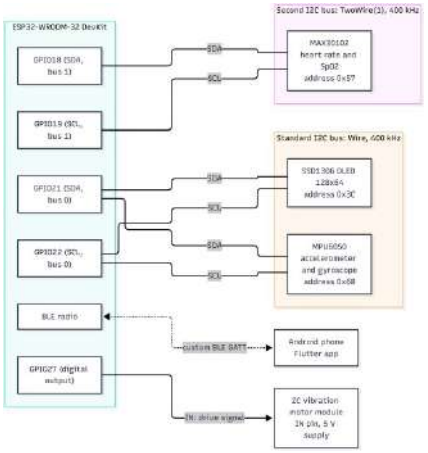



Fig. 2. ESP32 wearable wiring. The MAX30102 can also measure $SpO_2$, but the firmware reports heart rate only; the OLED is a debug display.

*Classifier.* The phone derives a movement value as the peak-to-peak spread of acceleration magnitude over the last 20 samples, multiplied by 6 and clamped to 0–20. Movement above 10.0 means awake; otherwise heart rate at or below 58 bpm gives deep sleep, at or below 60 bpm REM-cycle, and anything higher light sleep. These phase names are heuristic labels, not polysomnographic stages; any non-awake phase counts as asleep. On a change between awake and asleep, the app writes SLEEP or AWAKE to the wearable, which pulses the motor until an awake command arrives, and the boarding-risk logic combines the phase with the time remaining to decide whether to alert (Fig. 3).

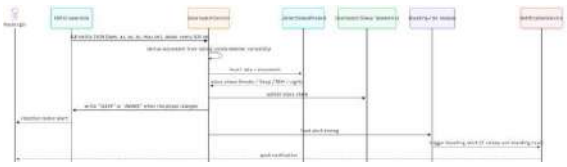


Fig. 3. Sleep-onset alert sequence: wearable telemetry, on-phone classification, wearable vibration command and boarding-alert push notification.

*Threshold provenance.* The constants (10.0 movement; 58 and 60 bpm) were set empirically by the developer during bench testing in the phone-accelerometer iteration and carried over unchanged. No labelled dataset, participants or formal tuning

were involved, so they are untuned defaults, not validated operating points. A movement value of 10.0 equals a peak-to-peak spread of about 1.7 m/s² (≈0.17 g) over roughly 10 s (20 samples at 2 Hz); the heart-rate cut-offs are population-level constants, not personal baselines. Because quiet sitting produces little movement and no minimum duration is enforced, we expect high sensitivity but reduced specificity: a hypothesis, not a finding.

*Literature rationale.* Heart rate falls from wakefulness to stable sleep [21] and movement drops at sleep onset, which is why wrist actigraphy scores sleep from motion. Fixed rules on wrist motion are established: the Cole–Kripke algorithm thresholds a weighted sum of activity counts and was validated against polysomnography [22], and van Hees et al. [23] detect sustained inactivity as arm-angle change below 5° for at least 5 min. These support the direction of our rules, not our constants, which use a different movement measure and remain untuned; unlike [23], our rule has no minimum-duration condition.

*Why rule-based.* No trained classifier was used because (i) no labelled airport sleep dataset exists and building one was outside scope; (ii) a threshold rule runs on the phone in real time with no server round trip; and (iii) it can be inspected and tuned without reflashing hardware. Wrist-accelerometry and photoplethysmographic classifiers reach about 90% sleep–wake accuracy against polysomnography [24] but need labelled training data. The rules are therefore a transparent on-device baseline, not a claim that fixed rules outperform learned classifiers; a trained classifier should be compared against them once labelled data exist.

*Known limitations and validation protocol.* No minimum stillness duration is applied, so quiet sitting can be labelled as sleep. In the validation protocol, the researcher alternates between simulated asleep postures (still, lying down) and awake activity, logging ground-truth labels beside the values the classifier saw; 30–60 labelled episodes yield sensitivity, specificity and a confusion matrix. Because one researcher simulating sleep cannot establish performance on real sleepers, the protocol is to be extended to several volunteers with separate tuning and test episodes, a threshold sweep, quiet-wake episodes and confidence intervals (with 30 episodes, an observed 90% sensitivity has a 95% Wilson interval of about 74–97%).

### F. Decision-Support Layer

Layer 1, the only trained model, is the Random Forest delay classifier and regressor. Layer 2 (arrival risk, leave-home time, boarding risk) is time-margin arithmetic over live data. Layer 3 (sleep detection, route ranking, recommendations, notification scheduling, risk score, chatbot) uses thresholds, deterministic rules and an LLM. Arrival margin is the gate-closing time minus (now + travel ETA + walking + security), bucketed at 30, 10 and 0 minutes (On Time, Slight Risk, High Risk, Miss Flight). Leave-home time is boarding close minus (travel + security + a weather buffer of 0, 10 or 20 min). Boarding risk is min(99, 100 × (walking + queue estimate) / time remaining), and risk-score confidence is 50 + 45 × (inputs supplied / 4).

### G. Requirements

Thirteen functional requirements were derived from the objectives: delay prediction for supported U.S. origins (FR1); sleep-phase classification, a demo-mode indicator and wearable commands (FR2–FR4); arrival-risk, leave-home, boarding-risk, route-ranking and recommendation logic (FR5–FR9); notification timing (FR10); an anonymising LLM assistant with local fallback (FR11); a travel-risk score (FR12); and authentication (FR13). Seven non-functional requirements cover operation with labelled demo data when APIs fail, latency, push reliability, API quota caching, visible labelling of dummy data, chatbot PII scrubbing and an acceptable SUS score; their status is reported in Section V.

## IV. Implementation

### A. Mobile Application and Backend

The Flutter client reaches its screens and services through Riverpod providers; one service is the only HTTP client for the backend and another handles BLE. With no wearable connected, the app produces a simulated sensor stream and shows a "Simulated/Demo Mode" badge. The Flask backend exposes 15 REST endpoints (authentication, flight status, weather, delay prediction, route ranking, prediction modules, alerts and the chatbot). The delay endpoint loads one joblib bundle (classifier, regressor, medians, lookup tables) per process, reports whether weather was live or a fallback median, and returns an explicit unsupported_airport error for origins outside the 30 trained airports, including Colombo. A local SQLite database stores sensor logs; its temperature and oxygen columns hold app-generated placeholders, as the wearable reports neither.

### B. External Services and Evaluation Data

Maps, flight-status and weather clients share a common interface; when a key is missing or a call fails they return deterministic dummy data labelled by a data_source field. Google's Directions API returned REQUEST_DENIED (legacy API not enabled), so evaluation used deterministic simulated traffic, which also keeps the decision-layer test cases reproducible. The client was later migrated to the Routes API, but live-traffic results were not collected, so no claim is made about route quality under real traffic.

### C. Chatbot and Privacy

The chatbot uses a provider-agnostic LLM client (DeepSeek deepseek-chat is deployed, after OpenAI gpt-4o-mini credit ran out). The backend injects current module outputs as context and, before dispatch, replaces email-like strings and runs of seven or more digits with placeholders; on failure it falls back to a local template engine.

## V. Results and Evaluation

### A. Flight Delay Prediction

Random Forest had the best validation macro-F1 (0.5756; four grid combinations, balanced weights), ahead of Logistic Regression (0.5604; three, balanced) and XGBoost (0.5124; four, no class weights). Because the XGBoost search was limited by memory and compute, the ranking reflects these searches and is not a verdict on the algorithms; it bounds the answer to RQ1.

Table III compares the selected model with a majority-class baseline that always predicts "on-time", computed from the

test-set class counts (611,800 on-time; 133,168 delayed); test-set results are reported for the selected model only.

TABLE III. TEST-SET RESULTS, PER-CLASS RESULTS AND CONFUSION MATRIX

| Metric | Random Forest | Majority class |
|---|---|---|
| Accuracy | 0.7486 | 0.8212 |
| Macro-F1 | 0.5824 | 0.4509 |
| AUC-ROC | 0.6521 | 0.5000 |
| Delayed-class precision | 0.3092 | undefined |
| Delayed-class recall | 0.3292 | 0.0000 |
| Balanced accuracy | 0.5846 | 0.5000 |
| Matthews corr. coeff. | 0.165 | 0 |
| Regressor MAE (min) | 23.10 | n/a |
| On-time P / R / F1 (n = 611,800) | 0.852 / 0.840 / 0.846 | n/a |
| Delayed P / R / F1 (n = 133,168) | 0.309 / 0.329 / 0.319 | n/a |
| Confusion matrix (test set, 744,968 flights): true on-time predicted on-time 513,884; true on-time predicted delayed 97,916; true delayed predicted on-time 89,333; true delayed predicted delayed 43,835. | | |

Baseline values, balanced accuracy and MCC were derived by us from the test-set class counts and the confusion matrix above.

Accuracy (0.7486) is below the 0.8212 of always predicting on-time, so it is not evidence of skill here. The model gains 0.131 in macro-F1 and 0.152 in AUC-ROC over the baseline: real but modest ranking power. It flags 19.0% of flights; 30.9% of those are delayed, 1.73 times the 17.88% base rate, while two thirds of actual delays are missed. The regressor's MAE of 23.10 minutes exceeds the 15-minute label threshold, and its residuals were not analysed.

Despite balanced class weights and macro-F1 selection, delayed-class recall is only 0.329. Further imbalance handling (threshold tuning on the validation precision–recall curve, cost-sensitive objectives, resampling such as SMOTE, a fuller XGBoost search) was not performed and is the first modelling item of future work; because the classifier outputs probabilities, its operating point can be moved along the precision–recall curve. The delay output is therefore a directional risk indicator, not a basis for changing plans. Higher accuracies in the literature (e.g., 90.2% in [5]) should not be read against our 0.749: accuracy depends on class balance, prediction target, features and split, and we did not reproduce those studies.

The main validity constraint is scope: the model was trained and tested only on 30 U.S. airports and never evaluated on Colombo departures (TC2), so no claim is made about the population the application is meant to serve.

### B. Sleep Detection

No validation data have been collected yet, so no sensitivity, specificity or accuracy is reported or implied. The BLE link was verified functionally on a Samsung SM-A226B (Android 13): payloads arrived about every 500 ms and the demo-mode badge cleared on connection. We expect quiet wakefulness to be mislabelled as sleep (Section III-E); results will be reported once the protocol is completed.

### C. Functional Testing

Ten scenarios were run against a local backend with real weather credentials and the trained model; all passed on first execution (Table IV). They cover boundary values and failure paths, not advice quality on real journeys; cases such as unavailable location, exhausted quota or no paired wearable were not executed.

TABLE IV. FUNCTIONAL TEST CASES (10/10 PASS)

| # | Module | Scenario | Result |
|---|---|---|---|
| TC1 | Delay | Valid U.S. flight (JFK to LAX) | Pass |
| TC2 | Delay | Non-U.S. origin (CMB): unsupported_airport | Pass |
| TC3 | Delay | Weather API unreachable: fallback medians | Pass |
| TC4 | Arrival | Margin at the 30-minute boundary | Pass |
| TC5 | Leave-home | High weather severity (+20 min) | Pass |
| TC6 | Boarding | Time remaining zero or negative | Pass |
| TC7 | Route | No deadline supplied | Pass |
| TC8 | Risk score | No inputs: missing_fields error | Pass |
| TC9 | Risk score | One of four inputs: lower confidence | Pass |
| TC10 | Chatbot | Email and phone removed before dispatch | Pass |

### D. Non-Functional Testing

*Route latency.* Over 30 timed requests the route endpoint gave a minimum of 287 ms, a median of 381.8 ms, a mean of 489.6 ms and a 95th percentile of 820.5 ms. Every request first made a rejected Directions API call and then fell back to simulated data, so the figures describe that configuration only, not live-traffic latency.

*Push notifications.* The first run on an Android emulator (Pixel 6, API 35) delivered 1 of 15 test pushes and crashed the app four times: three AndroidManifest.xml entries left from an abandoned native design had no compiled classes, so each message raised a ClassNotFoundException. After removing them, 15 of 15 pushes arrived without crashes, each within about 1–1.2 s. The test covered delivery through the Firebase Admin SDK, not end-to-end alert dispatch.

*Network drop.* With connectivity off, fetched cards kept rendering with their DEMO DATA badges, and the card needing a fresh call showed an error rather than an invented number; however, the error was a raw exception message and the card did not recover until restart. BLE latency and battery cost were not measured.

### E. Usability

A task-based study was run on 3–6 September 2026 as an anonymous online form with eleven adult volunteers. Each attempted three tasks (find delay risk, ask the assistant when to leave home, check sleep status), rated feature ease and usefulness on five-point scales, and completed the SUS [25]. Table V gives the results.

TABLE V. USABILITY RESULTS (N = 11)

| Measure | Result |
|---|---|
| SUS mean (SD) | 63.4 (10.4) |
| SUS median; range | 70.0; 47.5–75.0 |
| 95% CI of mean (t, our calculation) | 56.4–70.4 |
| Delay-risk task completed | 9/11 (81.8%) |
| Leave-home task completed | 9/11 (81.8%) |
| Sleep-status task completed | 11/11 (100%) |
| All tasks completed | 29/33 (87.9%) |

| Measure | Result |
|---|---|
| Feature ratings (10 items, 1–5) means | 4.09–4.64 |

The mean SUS lies below the common benchmark of 68 [26], between the "OK" and "Good" adjective anchors [27]; the median is above 68 and the confidence interval spans it. Seven participants scored 67.5–75.0 and four 47.5–52.5; the two who timed out on the delay-risk and leave-home tasks are among the lowest scorers. The weakest ratings, understanding the information shown (4.09) and usefulness of the leave-home recommendation (4.10), concern the assistant and route outputs. As no SUS target was fixed in advance, the result is reported as marginal. The sample is small, unmoderated and self-reported, participants were not screened for flying frequency, and one duplicated SUS item was scored from its first copy (the second changes no score).

## VI. Discussion and Limitations

The evidence supports one conclusion: the components can be integrated in one architecture and a small group of volunteers can use it. It does not show that the delay predictions or sleep alerts can be relied on, or that the system reduces missed flights; that needs a prospective study. RQ1 is answered only for the searches run, RQ2 is open, RQ3 is answered for latency of the simulated route path but not battery cost, and RQ4 by the metric set but not alert timeliness.

*Threats to validity.* Delay data are U.S.-only, and the survey and usability samples are small, self-selected and not airport-specific. Route guidance was evaluated on simulated traffic and sleep behaviour was never measured. A single chronological split was used, the XGBoost search was limited, within-family tuning used non-chronological folds, and test-set results are reported for one model only. Data sources, features, split dates, class weights, grid sizes, the random seed and the classifier thresholds are stated; exact grid values will be released with the code.

*Deployment claims.* We make no claim of deployment readiness: the delay model has not been validated on Colombo-origin flights, the sleep classifier not against labelled episodes, and route guidance not on live traffic. Until then, the delay output is a directional indicator, the sleep alert a prototype behaviour, and the route output a decision-logic demonstration.

## VII. Conclusion and Future Work

We presented the design, implementation and evaluation of an AI-powered smart passenger travel assistant. A Random Forest delay classifier reached macro-F1 0.582 and AUC-ROC 0.652 on 744,968 held-out U.S. flights, above a majority-class baseline (0.451, 0.500) but weak on delayed flights (recall 0.329); ten of ten functional tests passed, a push-delivery defect was fixed (1/15 to 15/15), and eleven participants gave a mean SUS of 63.4 with 87.9% task completion. The system is a working, testable platform, not yet a service passengers can rely on.

Future work, in priority order: (1) retrain and evaluate the delay model on Colombo-origin flights; (2) improve delayed-class recall with threshold tuning, cost-sensitive learning, resampling (e.g., SMOTE) and a fuller XGBoost search, and analyse the regressor's residuals; (3) validate the sleep classifier with several participants following the Section III-E protocol, add a minimum-stillness duration and personal heart-rate baselines, and compare against a trained classifier; (4) evaluate route guidance with live Routes API traffic; (5) measure BLE latency and battery cost; (6) run an observed usability study; and (7) add retry and plain-language errors to cards that fail offline.

## References


[1] M. Ball et al., "Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States," NEXTOR, Univ. California, Berkeley, CA, USA, Oct. 2010. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/6234

[2] Airlines for America, "U.S. passenger carrier delay costs," *A4A Data & Statistics*. [Online]. Available: https://airlines.org/dataset/u-s-passenger-carrier-delay-costs/

[3] Bureau of Transportation Statistics, U.S. Department of Transportation, "Airline on-time performance data." [Online]. Available: https://www.transtats.bts.gov/

[4] N. L. Kalyani et al., "Machine learning model-based prediction of flight delay," in *Proc. 2020 4th Int. Conf. I-SMAC*, 2020, pp. 577–581, doi: 10.1109/I-SMAC49090.2020.9243339.

[5] G. Gui et al., "Flight delay prediction based on aviation big data and machine learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 140–150, 2020, doi: 10.1109/TVT.2019.2954094.

[6] P. Hu et al., "Research on flight arrival delay prediction based on support vector machine," in *Proc. 2022 IEEE 4th Int. Conf. Civil Aviation Safety and Information Technology (ICCASIT)*, 2022, pp. 593–598, doi: 10.1109/ICCASIT55263.2022.9986792.

[7] Anees and W. Huang, "Flight delay prediction: Data analysis and model development," in *Proc. 2021 26th Int. Conf. Automation and Computing (ICAC)*, 2021, pp. 1–6, doi: 10.23919/ICAC50006.2021.9594260.

[8] X. Dou, "Flight arrival delay prediction and analysis using ensemble learning," in *Proc. 2020 IEEE 4th ITNEC*, 2020, pp. 836–840, doi: 10.1109/ITNEC48623.2020.9084929.

[9] Y. Demirkaya et al., "Forecasting flight delays from operational and weather data: A regression-based machine learning application," in *Proc. 2025 9th Int. Symp. Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2025, pp. 1–8, doi: 10.1109/ISMSIT67332.2025.11267898.

[10] K. Dheemonth et al., "Travel Assist: Resolution of travel queries using advanced natural language processing for India," in *Proc. 2024 1st Int. Conf. Communications and Computer Science (InCCCS)*, 2024, pp. 1–5, doi: 10.1109/InCCCS60947.2024.10593579.

[11] Y. Suresh et al., "Smart travel assistance using machine learning and generative artificial intelligence," in *Proc. 2025 Int. Conf. Intelligent Computing and Control Systems (ICICCS)*, 2025, pp. 757–762, doi: 10.1109/ICICCS65191.2025.10985115.

[12] J. B. Madavarapu et al., "HOT Watch: IoT-based wearable health monitoring system," *IEEE Sensors J.*, vol. 24, no. 20, pp. 33252–33259, 2024, doi: 10.1109/JSEN.2024.3424348.

[13] Anjum et al., "Wearable biomedical sensors for continuous monitoring of chronic disease progression," in *Proc. 2025 7th Int. Conf. Inventive Material Science and Applications (ICIMA)*, 2025, pp. 222–227, doi: 10.1109/ICIMA64861.2025.11073848.

[14] T. Wu et al., "A rigid-flex wearable health monitoring sensor patch for IoT-connected healthcare applications," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6932–6945, 2020, doi: 10.1109/JIOT.2020.2977164.

[15] P. Zouein et al., "AI-driven optimization of passenger flow: Integrating computer vision, machine learning, and simulation for enhanced efficiency and revenue generation," in *Proc. 2025 IEEE Int. Conf. Engineering, Technology, and Innovation (ICE/ITMC)*, 2025, pp. 1–8, doi: 10.1109/ICE/ITMC65658.2025.11106669.

[16] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A

design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[17] Iowa State University, "Iowa Environmental Mesonet: ASOS/AWOS/METAR data." [Online]. Available: https://mesonet.agron.iastate.edu/

[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[21] H. J. Burgess, J. Kleiman, and J. Trinder, "Cardiac activity during sleep onset," *Psychophysiology*, vol. 36, no. 3, pp. 298–306, May 1999.

[22] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–469, 1992, doi: 10.1093/sleep/15.5.461.

[23] V. T. van Hees et al., "A novel, open access method to assess sleep duration using a wrist-worn accelerometer," *PLoS ONE*, vol. 10, no. 11, Art. no. e0142533, 2015, doi: 10.1371/journal.pone.0142533.

[24] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, vol. 42, no. 12, Art. no. zsz180, 2019, doi: 10.1093/sleep/zsz180.

[25] J. Brooke, "SUS: A 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

[26] J. Sauro, *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. Denver, CO, USA: Measuring Usability LLC, 2011.

[27] Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the System Usability Scale," *Int. J. Human-Comput. Interact.*, vol. 24, no. 6, pp. 574–594, 2008.

# Coherence, Not Volume: Sub-Threshold Dilution Attacks on Provenance-Based Agent Memory and a Spectral Defence

Senesh Fitzroy
*Faculty of Computing, NSBM Green University*
*Colombo, Sri Lanka*
dsfmendis@students.nsbm.ac.lk

Isidara Gallage
*Faculty of Computing, NSBM Green University*
*Colombo, Sri Lanka*
charuniisidara@gmail.com

Sanuji Walgampaya
*Faculty of Computing, NSBM Green University*
*Colombo, Sri Lanka*
sanujiwalgampaya2003@gmail.com

Thyagi Luvishewa
*Faculty of Computing, NSBM Green University*
*Colombo, Sri Lanka*
charuniisidara@gmail.com

***Abstract*—Large language model (LLM) agents persist knowledge in long-term memory and periodically consolidate retrieved items into derived beliefs. To contain memory poisoning, current defences label items with provenance and propagate a scalar taint along derivation edges, quarantining beliefs whose taint exceeds a threshold. We show that this defence family shares a structural blind spot: propagated taint is an additive functional of provenance labels and retrieval weights, and is therefore content-independent. It measures the volume of untrusted ancestry, never its coherence. We formalise this as a volume-invariance result and exploit it with DILUTE, an attack that adds no untrusted writes but makes those an agent already accepts collude along a shared claim direction under a per-item stealth cap. We propose CIRCA, which scores coalitional coherence via the leading eigenvalue of the hollow Gram matrix of the untrusted premise set, whitened against benign covariance; a Baik–Ben Arous–Péché argument predicts detectability when . We evaluate both in simulation and on an agent-memory corpus of 320 untrusted, 121 trusted and 132 adversarial records using 384-dimensional sentence embeddings, with a live LLM performing consolidation. Taint-based detection sits at chance throughout (AUROC 0.489–0.505), while attack efficacy rises; across five consolidation hops residual taint decays 256× as efficacy grows. Against a stealth-enforced adversary on real text, content screening stays near chance at $k = 4$ (0.478); CIRCA whitened against a held-out benign reference reaches 0.713 at $k = 4$ and 0.887 at $k = 6$, and whitening is essential. We also report where CIRCA fails: it flags naturally coherent single-topic benign clusters and loses power on topics its reference has not seen. Live-LLM consolidation is steered from 0.00 to 0.46 endorsement of the attacker's target claim (one-sided Fisher $p < 10^{-3}$), at a cost of 122 µs per consolidation.**

***Keywords—LLM agents, memory poisoning, data provenance, spectral detection, AI security***

## I. INTRODUCTION

Autonomous agents built on large language models (LLMs) no longer treat context as ephemeral. Systems such as ReAct [1], Reflexion [2], Generative Agents [3] and MemGPT [4] persist observations, distil them into reusable lessons, and retrieve them across sessions. This persistence enables long-horizon autonomy, and it also converts a transient prompt-injection incident [5] into durable state corruption.

The resulting threat, *memory poisoning*, is well documented. Attacks plant malicious trajectories through the ordinary agent pipeline [6]–[9], lie dormant until triggered [10], [11], propagate across agents [12], [13], and survive naive cleanup [14]; surveys now organise the whole memory lifecycle [15], [16].

Defences have converged on one architecture: treat memory as a derivation graph, label items with the provenance of their origin, and propagate a taint or trust value along derivation edges. MAP-Graph [17] checks supporting memories for untrusted ancestors before high-risk actions; MemLineage [18] anchors per-entry signatures in a Merkle log; SENTINEL [19] applies provenance labelling and taint-threshold filtering; belief-based schemes cap trust by source provenance [20]. All reduce a derived belief's risk to a scalar compared against a threshold.

**We argue the scalar measures the wrong thing.** Propagated taint is an additive functional of provenance labels weighted by retrieval weights; it never inspects *what the untrusted items say*. Two premise sets with identical provenance composition receive identical taint, even if one contains items discussing unrelated topics and the other contains items secretly aligned on a single adversarial claim. Taint measures the *volume* of untrusted ancestry; what determines whether the agent's conclusion is hijacked is its *coherence*.

We exploit this with **DILUTE**. The adversary accepts a per-item stealth cap $\epsilon$: every injected write must individually look ordinary. Crucially it does *not* inject more items — it repurposes the untrusted writes the agent already ingests (retrieved pages, tool outputs, peer messages) so that $k$ of them share a common latent direction. Aggregate influence grows with $k$ because aligned components add coherently while benign residue adds incoherently; taint does not move at all.

We then propose **CIRCA** (Coalitional Influence for Robust Contamination Accounting), which scores exactly what taint discards. A colluding coalition plants a rank-one spike in the second-moment structure of the untrusted premise set; CIRCA computes the leading eigenvalue of the *hollow* Gram matrix, whitened against benign covariance, and compares it to the edge expected under independent provenance, giving a threshold of Baik–Ben Arous–Péché (BBP) type [21], [22].

Contributions:

- A volume-invariance result (Prop. 1): any additive taint scheme has detection power exactly equal to chance against an adversary that changes coherence without changing untrusted volume.
- DILUTE, a sub-threshold dilution attack instantiating that bound, and evidence that multi-hop consolidation *launders* it — taint decays geometrically with depth while efficacy grows.
- CIRCA, a content-aware spectral detector with a closed-form threshold , $O(n_u^2\, d)$ cost, and a validated phase transition in claim-space dimension.
- A real-text evaluation on a 384-dimensional embedded agent-memory corpus with a *live LLM* performing consolidation, with held-out whitening, bootstrap intervals, a clean-data-only threshold, and an explicit account of CIRCA's failure modes (topic-coherent benign clusters, unseen topics).

**What is new.** Existing defences cannot see a property that exists only jointly. Taint schemes [17]–[19] do not read content at all (Prop. 1); per-item screens, including activation and gradient detectors [29], [30], score items one at a time and pass every item a stealth cap admits; post-hoc auditing [31] acts after the belief is written. CIRCA is a set-level, pre-write test with a closed-form threshold that needs only the embeddings of the untrusted premises of the consolidation being checked.

## II. BACKGROUND AND RELATED WORK

**Agent memory.** Retrieval-augmented generation [23] supplies non-parametric knowledge; agent memory adds *writes*. MemGPT [4] pages between context tiers, HippoRAG [24] builds associative indices, Zep [25] maintains a temporal knowledge graph, and reflective agents [2], [3] summarise episodes into lessons. Consolidation — aggregating several retrieved items into one derived belief that is written back — is the common primitive.

**Memory poisoning.** AgentPoison [6] and PoisonedRAG [7] corrupt retrieval corpora; indirect prompt injection [5] supplies the write channel. Systematic studies map the write surface [8], [26]; dormant variants [10], [11] echo classical backdoors [27] and data poisoning [28]. Detection has used activation signatures [29], [30], causal attribution [31] and trajectory signals, whose provenance observability is itself bounded [32].

**Provenance and information flow.** Tracking derivation is old: database provenance [33], [34], information flow [35], [36] and dynamic taint analysis [37]. Agent-memory defences import this machinery [17], [18], [38], [39]. A recurring difficulty is over-tainting: strict propagation eventually marks everything, so systems adopt thresholds — precisely what we target. Belief-revision and temporal-validity work [40], [41] addresses *stale* beliefs, an orthogonal failure mode — our attacker's memories are current.

**Sub-threshold evasion.** Distributing an attack below a per-step alarm is known in trajectory monitoring [42], and coalition-aware attribution has been studied via Shapley values [43], [44]. To our knowledge the interaction between per-item stealth caps and *provenance* accounting in agent memory has not been analysed, and no agent-memory defence scores coalitional coherence. Our detector draws on spiked random matrix theory [21], [22], [45] and planted structure recovery [46], [47].

## III. THREAT MODEL AND FORMALISATION

**Memory model.** Each item $i$ carries a unit *claim vector* $x_i \in \mathbb{S}^{d-1}$ and an origin label $\tau_i \in \{0,1\}$, with $\tau_i = 1$ for untrusted channels. A consolidation event retrieves a premise set $S$, $|S| = s$, with retrieval weights $w \in \Delta^{s-1}$, and writes

Let $U = \{i \in S : \tau_i = 1\}$, $n_u = |U|$, $W_U = \sum_{i \in U} w_i$.

**Adversary.** It controls $A \subseteq U$, $|A| = k$, of the untrusted writes the agent *already* accepts, and targets a claim $v^\star \in \mathbb{S}^{d-1}$. It is bound by a **per-item stealth cap** $\epsilon$: each controlled item satisfies $\langle x_j, v^\star \rangle \le \epsilon$ , with the remainder spent on benign-looking camouflage. This is the binding constraint in practice — a single strongly on-message write is caught by content screening — and is more realistic than a global payload budget. Efficacy is $\eta = \langle c(S), v^\star \rangle$. Crucially the adversary does **not** change $n_u$ or $w$ ; it changes only the *geometry* of untrusted claims.

**Defences. D1** binary taint [36] quarantines any belief with an untrusted ancestor. **D2** scalar taint propagation, the dominant design [17]–[19], computes $t(c) = \sum_{i \in S} w_i \tau_i = W_U$ and flags $t > \theta$. **D3** content anomaly screening flags the premise set by its most anomalous untrusted item, standing in for embedding-outlier and activation-probe screening. **D4** is CIRCA (Section V). D4 needs the claim embeddings of the untrusted items in the premise set being consolidated, not the whole memory store, model internals, or extra LLM calls.

## IV. THE DILUTE ATTACK

**Proposition 1 (Volume-invariance of additive taint).** *Let $t(c) = \sum_{i \in S} w_i \tau_i$ be any additive propagated-taint functional over origin labels. Then $t$ is a function of $(\tau, w)$ alone and is constant on the premise configurations $\{x_i\}$. Hence for any pair of episodes with identical provenance labels and retrieval weights — one benign, one containing an arbitrary colluding coalition — $t$ takes the same value, and any threshold rule on $t$ achieves detection AUROC exactly .*

*Proof.* Immediate: $t$ does not read $\{x_i\}$. The adversary alters only $\{x_i\}$, so the distributions of $t$ under $H_0$ and $H_1$ coincide and every threshold gives equal true- and false-positive rates.

Proposition 1 is elementary; its force is that it applies to *every* defence in the taint family regardless of how carefully the graph is signed or the threshold tuned. Cryptographic provenance such as MemLineage [18] authenticates *who* wrote an item, which is orthogonal to *what* the items jointly assert.

**Proposition 2 (Coalitional efficacy).** *Let* $W_A = \sum_{j \in A} w_j$ *. With benign claims uniform on* $\mathbb{S}^{d-1}$*, strictly increasing in* $W_A$*. Since* $W_A$ *can be raised to* $W_U$ *with no change to* $t(c) = W_U$ *, the adversary buys efficacy at exactly zero taint cost.*

*Proof Sketch.* Decompose $\sum_i w_i x_i$ into its $v^\star$ component, $\epsilon W_A$ plus an benign term, and its orthogonal component, whose squared norm concentrates at . Normalising by (1) gives the expression.

**Laundering.** The attack is amplified by the mechanism defences rely on. Once a derived belief is written it is *agent-authored*, hence trusted-origin; at the next hop its taint is multiplied by its retrieval weight among trusted siblings and decays geometrically. But if several sibling beliefs were each consolidated from premise sets carrying the planted direction, the payload accumulates coherently across the merge. Provenance dilutes; influence compounds.

## V. CIRCA

**Detector.** Let $X_U \in \mathbb{R}^{nu \times d}$ stack the retrieval-weighted untrusted claims, . Define the *hollow* Gram matrix $G = X_U X_U^\top$, $G_{ii} := 0$. Hollowing is essential: the diagonal is identically one for unit claims, carries no coalition information, and would otherwise dominate. CIRCA reports

flagging when $\Psi > \theta_\Psi$ . The leading eigenvector recovers an estimate $\hat{v}$ of the planted direction, from which the coalition's influence on the written belief, $|\langle c(S), \hat{v} \rangle| \, W_U$, follows.

**Proposition 3 (Detectability threshold).** *Under* $H_0$ *the off-diagonals of* $G$ *are approximately i.i.d. , so* $G$ *is Wigner and . Under* $H_1$ *a coalition of size* $k$ *with pairwise alignment* $\epsilon^2$ *contributes a rank-one block of spectral mass* $\simeq k\epsilon^2$ *. The regimes separate exactly when*

**Corollary 1 (Dimension helps the defender).** *Detectability is monotone in* $d$ *: benign claims become more nearly orthogonal as the claim space grows, lowering the edge as while the planted mass* $k\epsilon^2$ *is dimension-free. High-dimensional claim embeddings are a security asset, not merely a retrieval-quality one.*

**Proposition 4 (Coherence–efficacy frontier).** *If the adversary splits* $A$ *across* $r$ *mutually orthogonal directions to suppress the spike, the leading planted mass falls to , but only items project onto* $v^\star$*, so by Prop. 2 efficacy falls by the same factor. The adversary cannot reduce* $\Psi$ *without proportionally reducing* $\eta$*.*

**Whitening.** Eq. (3) assumes benign claims are isotropic. Real sentence embeddings are strongly anisotropic: on our corpus the nominal $d = 384$ carries an effective dimension of only $\approx 70$ (participation ratio of the benign spectrum), so unmodelled topical structure alone produces spikes. CIRCA therefore estimates the benign mean and covariance from a reference pool of trusted and historical items using Ledoit–Wolf shrinkage, and scores the whitened residual. Section VII shows this is not optional: on real embeddings unwhitened CIRCA is near chance where whitened CIRCA is decisive.

**Cost.** One $n_u \times d$ product and one symmetric eigendecomposition of an $n_u \times n_u$ matrix: $O(n_u^2 d + n_u^3)$, with $n_u$ the untrusted premises in a single retrieval (typically $< 16$). No model internals, gradients, or extra LLM calls; it composes with, rather than replaces, existing provenance tracking.

## VI. SIMULATION EVALUATION

We first validate the propositions in a controlled testbed implementing Section III exactly: $d = 256$, $s = 12$, $n_u = 8$, $\epsilon = 0.30$ unless stated, softmax retrieval weights, 800 Monte-Carlo episodes per configuration. Detection is scored as AUROC against clean episodes with *identical* untrusted volume — the discrimination a deployed filter actually faces.

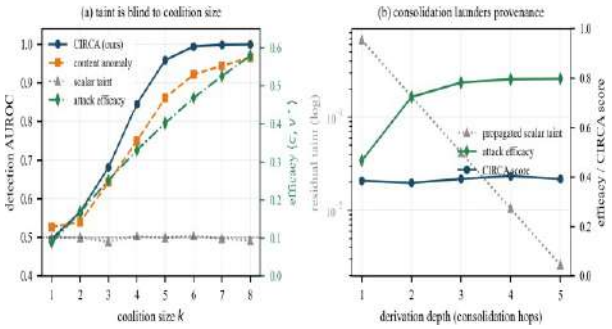



Fig. 1. *Simulation. (a) Attack efficacy rises* 6.6 × *with coalition size while scalar taint stays pinned at chance, confirming Prop. 1. (b) Across consolidation hops, residual taint falls* 256 × *(log scale) while efficacy grows.*

**Taint is blind.** Fig. 1(a) sweeps $k$ from 1 to 8. Efficacy grows from 0.088 to 0.579; scalar taint AUROC stays within [0.489,0.505], indistinguishable from 0.5, exactly as Prop. 1 requires. Binary taint (D1) attains AUROC 0.5 for the complementary reason: every episode has $n_u = 8$ untrusted premises, so D1 fires on every attacked *and* every benign consolidation. Its 100% recall in Table I is therefore not detection; it is 100% benign quarantine.

**Consolidation launders provenance.** Fig. 1(b) is the most damaging result for the taint family. Over five hops mean residual taint falls from 0.667 to 0.0026 — a factor of 256 — so any fixed threshold releases the belief within two or three consolidations. Over the same hops efficacy *rises* from 0.466 to 0.798. Provenance and influence move in opposite directions, which no monotone function of propagated taint can reconcile.

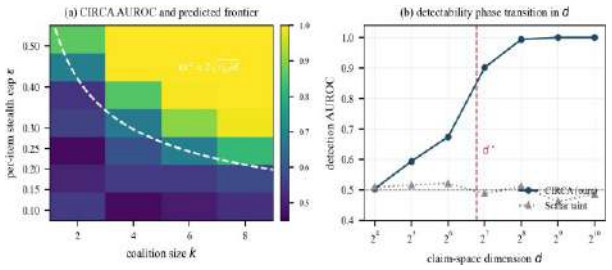



Fig. 2. *(a) CIRCA AUROC over the stealth-cap/coalition-size grid; the dashed curve is the parameter-free prediction of Prop. 3. (b) Detectability phase transition in claim-space dimension, with* $d^\star$ *the predicted crossing.*

**The predicted frontier is observed.** Fig. 2(a) sweeps $\epsilon$ against $k$; the transition from chance to near-perfect detection tracks (4) closely, and at $\epsilon = 0.10$ CIRCA never leaves chance for any $k \leq 8$ — the regime where no coherence test can succeed. Fig. 2(b) tests Corollary 1: AUROC climbs $0.504 \rightarrow 1.000$ as $d$ goes $16 \rightarrow 1024$, crossing where the predicted margin changes sign.

**Operating points.** At a matched 5% false-positive budget (Table I) CIRCA detects 0.978 of DILUTE consolidations against 0.490 for content screening and 0.090 for scalar taint. Binary taint achieves perfect recall only by quarantining everything — an availability failure, precisely the over-tainting pressure that motivated thresholds. Mean CIRCA latency is 122 µs per consolidation on one CPU core.

**Adaptive attacker.** A CIRCA-aware adversary splitting across $r$ orthogonal decoy directions traces the frontier of Prop. 4: at $r = 1$ efficacy is 0.466 with AUROC 0.995; pushing detection to chance (0.488 at $r = 6$) costs 81% of efficacy, leaving 0.090. Evasion is possible but not free.

*TABLE I*
*OPERATING POINTS AT A MATCHED 5% FALSE-POSITIVE BUDGET (SIMULATION, $k = 6, \epsilon = 0.30, n_u = 8, d = 256$).*

| Defence | TPR@5%FPR | AUROC | Benign quarantine |
|---|---|---|---|
| D1 binary taint † | 1.000 | 0.500 | 100% |
| D2 scalar taint | 0.090 | 0.523 | 5% |
| D3 content anomaly | 0.490 | 0.911 | 5% |
| **D4 CIRCA (ours)** | **0.978** | **0.996** | 5% |

†D1 fires on any untrusted ancestor, hence on every episode: recall and benign quarantine are both 100%.

## VII. REAL-TEXT EVALUATION

**Corpus.** We instantiate a long-lived SRE assistant agent whose memory holds 121 trusted records (verified runbooks, internal monitoring, CI output) and 320 untrusted records generated across 16 infrastructure topics. Six attack scenarios each target an operationally harmful claim (e.g. "disabling mTLS verification on the payments gateway is low-risk"), with 132 adversarial records in total. Every record is embedded with a 384-dimensional sentence encoder (all-MiniLM-L6-v2); claim vectors are therefore learned, not stipulated.

**Enforcing the stealth cap.** A rational adversary with embedding access keeps only fragments whose alignment with $v^\star$ stays inside the benign band. We enforce this by selecting fragments at or below the benign 95th percentile, leaving 9–12 usable fragments per scenario (Fig. 3(a)). This is attacker optimisation *within* the threat model, and it matters: without it the most on-message fragments are individually conspicuous and content screening alone is competitive.

**Detection.** Fig. 3(b) and Table II give the stealth-enforced sweep (1500 attacked and 1500 clean episodes per $k$; 95% bootstrap intervals). Scalar taint is at chance for every $k$ (0.493–0.512), as Prop. 1 demands. Content screening stays near chance through $k = 4$ (0.478 [0.458, 0.500]). The whitening reference matters, so we report it two ways. If the reference contains every benign record (*in-sample*), CIRCA reaches 0.939 [0.931, 0.946] at $k = 4$; this is an upper bound, since clean episodes are drawn from the records the covariance was fitted on. With a *held-out* reference (half of each topic's benign records plus the trusted records; episodes use the other half; five random splits) CIRCA reaches 0.713 (0.692–0.747) at $k = 4$, 0.887 at $k = 6$ and 0.964 at $k = 8$, above every baseline from $k = 3$. Unwhitened CIRCA is at chance at $k = 4$ (0.450), so whitening is necessary. A non-spectral alternative, the mean pairwise whitened cosine, trails $\lambda_{\max}$ at every $k$: a coalition occupying part of $U$ is diluted in an average but concentrated in the leading eigenvalue. Corollary 1 also holds on real text: truncating the whitened space to $d = 8$ drops AUROC to 0.590, rising to 0.993 at full dimension.

**Choosing $\theta_\Psi$ without attack data.** We set $\theta_\Psi$ to the 95th percentile of $\Psi$ on clean episodes alone. On fresh clean episodes the realised FPR is 4.5% (2.6–6.2% across splits) held-out and 5.7% in-sample, so only benign history is needed; no attack example enters calibration. The resulting TPR is in Table II.

**Naturally coherent benign clusters.** CIRCA scores coherence, not intent. When all eight untrusted premises come from one benign topic, which topical retrieval routinely produces, CIRCA exceeds that threshold in 87% of episodes (50–100% by topic), and a $k = 4$ attack is *less* coherent than such a cluster (AUROC 0.371 [0.351, 0.390]). When the reference has never seen the episodes' topics, held-out AUROC at $k = 4$ falls to 0.442. Shrinking the reference to 25, 50, 100 and 200 records gives 0.69, 0.72, 0.76 and 0.85 at $k = 4$. CIRCA is therefore reliable only when its reference covers the retrieved topics and $\Psi$ is compared with topic-matched benign coherence.

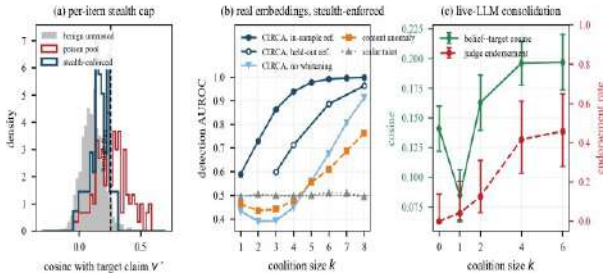



Fig. 3. Real-text evaluation. (a) Stealth-enforced fragments sit inside the benign alignment band. (b) Against that adversary, scalar taint and content screening remain near chance while whitened CIRCA separates; whitening is necessary. (c) A live LLM performing consolidation is steered toward the attacker's claim as $k$ grows (bars: s.e. for cosine, Wilson 95% for endorsement; $n = 24$ per point). (b) shows CIRCA with an in-sample and a held-out whitening reference.

*TABLE II*
*REAL-TEXT DETECTION AUROC, STEALTH-ENFORCED ATTACKER. TPR USES A THRESHOLD SET FROM CLEAN EPISODES ONLY (REALISED FPR 4.5%).*

| | **D2** | **D3** | **Pairwise** | **D4** | **D4** | **D4 TPR** |
|---|---|---|---|---|---|---|
| $k$ | taint | content | coh. (in) | (in) | (held-out) | (held-out) |
| 3 | 0.500 | 0.442 | 0.729 | 0.863 | 0.599 | 0.10 |
| 4 | 0.497 | 0.478 | 0.844 | 0.939 | 0.713 | 0.22 |
| 6 | 0.512 | 0.610 | 0.975 | 0.992 | 0.887 | 0.58 |
| 8 | 0.494 | 0.764 | 0.998 | 0.999 | 0.964 | 0.80 |

**Does it actually steer a real agent?** Fig. 3(c) replaces the modelled aggregation of (1) with a live LLM asked to write the durable lesson it takes from a retrieved premise set. We measure both the cosine of the derived belief with the target claim and an LLM judge's score for whether the belief endorses that claim. With no coalition the agent never endorses the target (0.00, Wilson 95% interval [0.00, 0.14]); endorsement then rises monotonically to 0.46 [0.28, 0.65] at $k = 6$ (one-sided Fisher exact $p = 1.1 \times 10^{-4}$ against $k = 0$ ). The geometric cosine moves far less (0.141 to 0.197) and is not monotone at $k = 1$: because stealth-enforced fragments are deliberately oblique, cosine with $v^{\star}$ understates an effect that the semantic judge registers clearly. We therefore read the endorsement trend, not either endpoint, as the result. The attack transfers from the geometric model to natural-language consolidation using fragments that no individual content check flags.

## VIII. DISCUSSION AND LIMITATIONS

**Scope.** Prop. 1 does not say provenance tracking is useless. Provenance answers questions CIRCA cannot — who wrote an item, when, under what authority — and remains necessary for forensics, revocation and access control [17], [18]. Our claim is narrower: provenance *volume* is not a proxy for adversarial influence, and a deployed system needs both axes.

**What the real-text evaluation is and is not.** The corpus is LLM-generated and the adversarial fragments are authored to a stated stealth constraint, not harvested from observed attacks. The "agent" is a consolidation prompt, not a production stack with tools and multi-session state; the judge is an LLM and inherits its biases; and with $n = 24$ episodes per point the $k$-trend, not any single value, is the claim.

**Topical coherence is the main open problem.** Whitening removes global anisotropy, not topical structure: Section VII shows that single-topic benign clusters look more coherent than a $k = 4$ coalition, and that a reference blind to the retrieved topics costs most of CIRCA's power. Our in-sample real-text figures are upper bounds and the held-out ones are what a deployment should expect. Per-topic reference covariance and per-topic thresholds are the required next step, and CIRCA should be deployed alongside provenance, not in place of it.

**Below the frontier.** When the coalition is information-theoretically invisible to any coherence test, CIRCA included. This is a genuine limit: defence there must come from reducing $n_u$ (admission control) or raising $d$, both of which our analysis quantifies.

**Next steps.** A larger live-LLM panel, and validation on a live agent stack with tool use and persistent multi-session memory, against benchmarks such as STALE [41], MemSecBench [26] and TAMAS [12], is the essential next step.

## IX. CONCLUSION

Provenance-based defences for LLM agent memory share a structural blind spot: propagated taint is an additive functional of origin labels and is therefore invariant to what untrusted memories actually say. DILUTE turns that invariance into a capability, raising attack efficacy $6.6\times$ while taint-based detection sits at chance, and multi-hop consolidation launders it further, decaying residual taint $256\times$ while efficacy grows. CIRCA addresses the gap by scoring coalitional coherence spectrally, with a parameter-free detectability threshold our simulations reproduce; on real embeddings it outperforms taint and content screening with a held-out reference, but it needs topic-matched calibration to avoid flagging benign topical clusters. The broader lesson for secure agent design is that contamination accounting must be *super-additive*: what makes a set of memories dangerous is not how many came from untrusted sources, but how strongly they agree.

## REFERENCES


[1] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in Proc. ICLR, 2023.

[2] N. Shinn et al., "Reflexion: Language agents with verbal reinforcement learning," in Proc. NeurIPS, 2023.

[3] J. S. Park et al., "Generative agents: Interactive simulacra of human behavior," in Proc. ACM UIST, 2023.

[4] C. Packer et al., "MemGPT: Towards LLMs as operating systems," arXiv:2310.08560, 2023.

[5] K. Greshake et al., "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in Proc. ACM AISec, 2023.

[6] Z. Chen et al., "AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases," in Proc. NeurIPS, 2024.

[7] W. Zou et al., "PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models," in Proc. USENIX Security, 2025.

[8] P. Dash et al., "From untrusted input to trusted memory: A systematic study of memory poisoning attacks in LLM agents," arXiv:2606.04329, 2026.

[9] K. Wang et al., "OEP: Poisoning self-evolving LLM agents via locally correct but non-transferable experiences," arXiv:2605.18930, 2026.

[10] S. Pulipaka et al., "Hidden in memory: Sleeper memory poisoning in LLM agents," arXiv:2605.15338, 2026.

[11] Y. Li et al., "Plant, persist, trigger: Sleeper attack on large language model agents," arXiv:2605.28201, 2026.

[12] I. Kavathekar et al., "TAMAS: Benchmarking adversarial risks in multi-agent LLM systems," arXiv:2511.05269, 2025.

[13] H. Liu et al., "Memory poisoning propagation and repair mechanism in multi-agent collaborative environments," in Proc. ICAIDMTSC, ACM, 2026, pp. 225–231, doi:10.1145/3806262.3806294.

[14] C. Lam et al., "Governing evolving memory in LLM agents: Risks, mechanisms, and the stability and safety governed memory (SSGM) framework," arXiv:2603.11768, 2026.

[15] Z. Lin et al., "A survey on long-term memory security in LLM agents: Attacks, defenses, and governance across the memory lifecycle," arXiv:2604.16548, 2026.

[16] Y. Wang et al., "From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents," arXiv:2606.04990, 2026.

[17] Y. Wang et al., "MAP-graph: Provenance-aware shared memory for multi-agent workflows," arXiv:2608.10509, 2026.

[18] C. Ouyang and R. Hou, “MemLineage: Lineage-guided enforcement for LLM agent memory,” arXiv:2605.14421, 2026.

[19] N. Karamchandani et al., “Your agent's memories are not its own: Forged reasoning attacks on LLM agent memory and defenses,” arXiv:2607.05029, 2026.

[20] P. Singh, “When does belief-based agent memory help? Reliability-conditional updating and provenance-capped poisoning defense,” arXiv:2606.22030, 2026.

[21] J. Baik, G. Ben Arous, and S. Péché, “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,” Ann. Probab., vol. 33, no. 5, pp. 1643–1697, 2005.

[22] I. M. Johnstone, “On the distribution of the largest eigenvalue in principal components analysis,” Ann. Statist., vol. 29, no. 2, pp. 295–327, 2001.

[23] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. NeurIPS, 2020.

[24] B. J. Gutiérrez et al., “HippoRAG: Neurobiologically inspired long-term memory for large language models,” in Proc. NeurIPS, 2024.

[25] P. Rasmussen et al., “Zep: A temporal knowledge graph architecture for agent memory,” arXiv:2501.13956, 2025.

[26] X. Chen et al., “MemSecBench: Tracking agent memory poisoning from persistence to consequence and repair,” arXiv:2607.27080, 2026.

[27] E. Hubinger et al., “Sleeper agents: Training deceptive LLMs that persist through safety training,” arXiv:2401.05566, 2024.

[28] N. Carlini et al., “Poisoning web-scale training datasets is practical,” in Proc. IEEE S&P, 2024.

[29] I. Gowda, “MEMSAD: Gradient-coupled anomaly detection for memory poisoning in retrieval-augmented agents,” arXiv:2605.03482, 2026.

[30] X. Tan et al., “RevPRAG: Revealing poisoning attacks in retrieval-augmented generation through LLM activation analysis,” arXiv:2411.18948, 2024.

[31] Z. Tan et al., “MemAudit: Post-hoc auditing of poisoned agent memory via causal attribution and structural anomaly detection,” arXiv:2605.23723, 2026.

[32] J. W. Leong, “Retrieval observability bounds on provenance detection for agent memory poisoning: Measured coverage and a falsified standalone detector,” arXiv:2606.30566, 2026.

[33] P. Buneman, S. Khanna, and W.-C. Tan, “Why and where: A characterization of data provenance,” in Proc. ICDT, 2001.

[34] T. J. Green, G. Karvounarakis, and V. Tannen, “Provenance semirings,” in Proc. ACM PODS, 2007.

[35] D. E. Denning, “A lattice model of secure information flow,” Commun. ACM, vol. 19, no. 5, pp. 236–243, 1976.

[36] K. J. Biba, “Integrity considerations for secure computer systems,” MITRE Corp., Tech. Rep. ESD-TR-76-372, 1977.

[37] J. Newsome and D. Song, “Dynamic taint analysis for automatic detection, analysis, and signature generation of exploits on commodity software,” in Proc. NDSS, 2005.

[38] Z. Fei et al., “Selection integrity for LLM graph memory: An accumulability criterion for information-flow-blind retrieval,” arXiv:2606.12290, 2026.

[39] A. Alvarez et al., “ProvenanceGuard: Source-aware factuality verification for MCP-based LLM agents,” arXiv:2606.18037, 2026.

[40] X. Li et al., “MemTX: Transactional belief commit for stateful agent memory,” arXiv:2607.23929, 2026.

[41] H. Chao et al., “STALE: Can LLM agents know when their memories are no longer valid?” arXiv:2605.06527, 2026.

[42] Q. Zhang and Y. Gao, “The distributed detectability band against marginal-preserving attacks,” arXiv:2606.10456, 2026.

[43] L. S. Shapley, “A value for n-person games,” in Contributions to the Theory of Games II, Princeton Univ. Press, 1953, pp. 307–317.

[44] A. Ghorbani and J. Zou, “Data Shapley: Equitable valuation of data for machine learning,” in Proc. ICML, 2019.

[45] V. A. Marchenko and L. A. Pastur, “Distribution of eigenvalues for some sets of random matrices,” Math. USSR-Sbornik, vol. 1, no. 4, pp. 457–483, 1967.

[46] N. Alon, M. Krivelevich, and B. Sudakov, “Finding a large hidden clique in a random graph,” Random Struct. Algorithms, vol. 13, no. 3–4, pp. 457–466, 1998.

[47] Q. Berthet and P. Rigollet, “Complexity theoretic lower bounds for sparse principal component detection,” in Proc. COLT, 2013.

# Policy As Code Based System for Detection and Risk Prioritization of AWS IAM, S3 and Security Group Misconfigurations

Uvindu Perera
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwaththa, Pitipana, Homagama, Sri Lanka*
uvinduashitha1006@gmail.com

Madusanka Mithrananda
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwaththa, Pitipana, Homagama, Sri Lanka*
madusanka.m@nsbm.ac.lk

***Abstract*—Cloud misconfiguration remains a significant security concern because small configuration errors can expose cloud resources, broaden permissions, or weaken important protection mechanisms. Although existing cloud security tools can identify many of these issues, their findings often focus on compliance status or individual policy violations rather than explaining how the observed configuration evidence relates to practical risk. This study presents Policy as Code based prototype for detecting and prioritizing misconfigurations in AWS Identity and Access Management (IAM), Amazon S3, and EC2 Security Groups. The system collects AWS configuration data through read only APIs, normalizes the collected information and evaluates it using 18 versioned CloudFormation Guard policies. Failed policy evaluations are then resolved to resource specific evidence, while overlapping detections that represent the same underlying condition are consolidated before risk assessment. Likelihood and impact values are derived from observable configuration evidence and combined to produce an explainable risk score and severity level, with MITRE ATT&CK mappings used only as supporting threat context. Policy level verification covered 177 predefined Guard test cases, with all expected outcomes satisfied. The complete workflow was further evaluated using controlled secure and misconfigured scenarios across the three selected AWS services. Across 26 predefined policy-resource evaluation units, the prototype produced 8 true positives and 18 true negatives, with no false positives or false negatives within the controlled dataset. The results demonstrate the feasibility of combining Policy as Code detection with evidence based risk prioritization while maintaining transparency***

***Keywords—AWS security, cloud misconfiguration, policy as code, risk prioritization.***

## I. INTRODUCTION

Cloud computing is a technology that is widely used in the modern world. The reason why cloud computing or cloud services are widely used is because of its convenient and flexible access to computing resources [1]. Global demand for cloud infrastructure services continues to increase, with AWS, Microsoft Azure and Google Cloud remaining the three leading providers. In the first quarter of 2026, AWS maintained the largest worldwide market share at 28%, followed by Microsoft at 21% and Google at 14% [2]. The reason for this widespread adoption is the range of services provided by AWS.

AWS offers cloud services tailored to organizations of different sizes and requirements, including dedicated solutions for small and medium sized businesses (SMBs) and large enterprises [3], [4]. According to the 2026 Flexera State of the Cloud Report, 77% of surveyed SMBs reported running some or significant workloads on AWS [5]. As SMEs and large enterprises increasingly adopt cloud services, AWS emphasizes the importance of protecting data and cloud resources. To define the respective security obligations, AWS introduced the Shared Responsibility Model, under which AWS is responsible for security “of” the cloud, while customers are responsible for security “in” the cloud [6].

Under the AWS Shared Responsibility Model, customers retain responsibility for securing the configurations and resources they deploy in the cloud. In IAM, customers are responsible for configuring identities, credentials, and appropriate permissions [7]. In Amazon S3 they are responsible for managing data access, permissions, and relevant security controls [8] and in Amazon EC2, customers are responsible for controlling network access through VPCs and security groups, managing credentials, and configuring IAM roles associated with instances [9]. Consequently, the security posture of an organization depends substantially on the correct configuration of these services. Cloud misconfiguration, including over-privileged identities and publicly exposed resources, has been identified as a significant source of cloud security risk [10]. The consequences of such configuration weaknesses are evident in several reported AWS related security incidents. The 2019 Capital One breach involved a firewall configuration weakness followed by access to IAM credentials and private S3 resources [11].

Various cloud security solutions are available to detect and mitigate cloud misconfigurations. Large enterprises are generally better positioned to adopt and manage such security solutions because they tend to have greater IT security investment capacity, skilled personnel, documented processes, and formal IT-budget planning, whereas these capabilities are often less developed in SMEs [12]. Policy as Code (PaC) [13] provides a potential approach for addressing these limitations by expressing security, compliance, and organizational requirements as machine readable rules that can be automatically evaluated and enforced. Therefore, this study

proposes a lightweight prototype that applies Policy as Code to detect misconfigurations in AWS IAM, S3, and EC2 Security Groups and combines the resulting configuration evidence with likelihood and impact based risk prioritization [14] to provide explainable security findings.

However, a lightweight and explainable approach for detecting AWS misconfigurations and prioritizing them based on their associated risk is still needed, despite the presence of cloud security mechanisms. Hence in this paper we discuss the combination of Policy as Code based detection and evidence driven risk prioritization for AWS IAM, S3 and EC2 Security Groups. The next section reviews the related literature on cloud misconfiguration, Policy as Code, cloud security tools, and risk prioritization techniques.

## II. Related work

Cloud misconfigurations are frequently the result of human and operational mistakes during the configuration or modification of cloud resources [15].The complexity of the configuration of cloud environments and the fact that even with security controls and recommended practices in place, errors can still occur during administration. Mistakes in the settings, omissions, miscommunication and mistakes during automatic configuration processes could be the reason for these errors. Therefore, some configuration weaknesses may not be immediately apparent when deploying or modifying resources. As we have mentioned in the Introduction, misconfigurations in services like AWS IAM, S3 and EC2 Security Groups can create large security exposure. This underscores the necessity of continuous automated configuration assessment that can identify misconfigured resources, unauthorized changes, and configuration drift before they lead to security incidents [10].

Manual configuration can lead to errors and inconsistent security settings. Therefore policy driven automation is important in cloud and network security. The policy based management can reduce configuration errors and support automated security management [16]. Their later work also showed that security requirements can be converted into enforceable network controls [17] . Also policy validation can identify incorrect security policies before they are applied [18]. These studies show that automation can reduce manual effort and improve configuration consistency. However most of these approaches mainly focus on network and container environments. This creates a need to consider Policy as Code for AWS resource level misconfiguration detection.

Policy as Code means expressing human readable security and compliance requirements as machine readable rules that can be evaluated automatically [13]. These rules can be stored and managed in the same way as software code. This makes security policies easier to test, update, and apply consistently. It also reduces the need to check every cloud configuration manually. Policy as code can therefore be useful for detecting configurations that do not meet defined security requirements [19]. Finding a policy breach, however, does not necessarily reveal the true degree of risk involved. Therefore, in order to comprehend the finding's possible security effect, it should be assessed using the configuration evidence that is now accessible [14].

Organizations currently use different tools to improve cloud security. AWS provides native services such as AWS Config, Security Hub to assess configurations, collect security findings, and identify unintended access [20], [21]. These services may introduce usage based costs, while effective cloud security management also requires suitable technical knowledge [12]. Open source tools are also available. Prowler supports cloud security and compliance assessment [22]. Checkov detects misconfigurations in Infrastructure as Code before deployment [23].

These tools also have some limitations. Checkov mainly focuses on static analysis of Infrastructure as Code rather than live cloud resources [23].Prowler provides a broad set of security checks, but its approach is different from the specific evidence based risk model considered in this study [22].

Cloud security tools provide several benefits. They help identify configuration weaknesses and security issues. They improve visibility across cloud resources. They also support continuous security and compliance assessment [21]. However, the specific combination of Policy as Code, resource level evidence, and likelihood and impact based risk prioritization remains the focus of this study. Therefore, a lightweight prototype is proposed to combine these elements for AWS IAM, S3, and EC2 Security Groups.

AWS IAM, S3, and EC2 Security Groups are important services for controlling access to cloud resources. IAM follows the principle of least privilege, where users and roles should receive only the permissions required for their tasks [24]. Incorrect IAM permissions can therefore create unnecessary access to AWS resources. Amazon S3 also requires careful access configuration. AWS recommends using S3 Block Public Access to prevent unintended public access to buckets and objects [25]. EC2 Security Groups control inbound and outbound network traffic. Allowing unrestricted access to sensitive ports such as SSH or RDP can increase the exposure of cloud resources [26]. These examples show that different AWS services can create different types of security risk when they are not configured correctly. Therefore, detecting these misconfigurations is important for maintaining a secure cloud environment.

Detecting a misconfiguration only shows that a security rule has been violated. It does not show which issue should be fixed first. Risk assessment helps to compare findings by considering their likelihood and potential impact [14]. This allows higher-risk findings to receive more attention. The MITRE ATT&CK framework can also provide additional security context for these findings. ATT&CK includes adversary tactics and techniques related to cloud environments, including Infrastructure as a Service (IaaS) [27]. Mapping selected misconfigurations to relevant ATT&CK techniques can help explain how a configuration weakness may relate to possible adversary behavior [27]. In this study, MITRE ATT&CK is used only as supporting security context and not to calculate the risk score.

The reviewed studies and tools show that cloud misconfigurations can be detected through automated assessment, policy-based controls, and security auditing [10], [17], [18], [19]. Existing approaches also support configuration validation and risk assessment [10], [14], [19]. However, these

functions are often handled separately. Limited attention has been given to a lightweight AWS focused approach that connects Policy as Code findings with resource level configuration evidence and then prioritizes them using likelihood and impact. Existing approaches also provide limited integration of MITRE ATT&CK as supporting context for explaining how a detected weakness may relate to adversary behavior. Therefore, this study addresses this gap by developing a prototype for selected AWS IAM, S3, and EC2 Security Group misconfigurations. The proposed approach combines Policy as Code detection, configuration evidence, likelihood- and impact-based risk prioritization, and MITRE ATT&CK mapping to produce explainable security findings.

## III. METHODOLOGY AND SYSTEM DESIGN

### *A. Research Design and System Architecture*

This study follows a Design Science Research approach. This approach is suitable because the study develops and evaluates a functional security artifact [28] . The artifact is a Policy as Code-based prototype for detecting and prioritizing selected AWS misconfigurations. The development process covers configuration collection, policy evaluation, evidence extraction, risk prioritization, and threat-context mapping.

Fig. 1 presents the proposed system architecture. The process starts by collecting configuration data from the AWS environment. The collected data are converted into normalized JSON structures. These data are then evaluated against Policy as Code rules. Failed rules are linked to the affected resources and their configuration evidence. Related evidence is consolidated before risk calculation. The final findings contain the affected resource, policy violations, supporting evidence, likelihood, impact, risk score, severity, and relevant MITRE ATT&CK context.

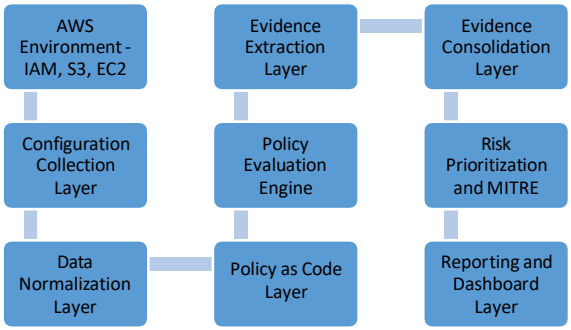



Fig. 1. Proposed architecture of the Policy as Code-based AWS misconfiguration detection and risk prioritization system.

### *B. Policy Selection, Configuration Collection and Policy Evaluation*

The prototype focuses on AWS IAM, Amazon S3, and EC2 Security Groups. The policies were selected using AWS security best practices and AWS Security Hub controls for these services [24], [25], [26], [29], [30]. The selected controls cover important configuration areas such as excessive IAM permissions, missing MFA, wildcard actions, unused credentials, public S3 access, weak S3 protection settings, and unrestricted Security Group access. The controls were then matched with the scope of this study and the configuration evidence that can be collected from AWS.

A total of 18 policies were implemented. Eight policies assess IAM configurations. Five policies assess Amazon S3 configurations. Five policies assess EC2 Security Groups. This number represents the selected scope of the prototype and not a complete list of AWS security controls.

Configuration data are collected using the AWS SDK for Python, Boto3 [31]. The collectors use read-only operations to retrieve the security properties required by the selected policies. Separate collectors are used for IAM, S3, and Security Groups. The collected data are converted into normalized JSON structures. This keeps the policy evaluation process independent from the original AWS API response format.

AWS CloudFormation Guard is used as the Policy as Code engine [19]. Guard provides a domain-specific language for defining policy rules and evaluating structured JSON or YAML data. Each selected misconfiguration is stored as an external .guard policy. The rules are kept separate from the main application logic. This allows the policies to be updated and tested independently.

Each normalized dataset is evaluated against the relevant Guard policies. A failed rule indicates that a security requirement is not satisfied. The failed policy is then passed to the evidence extraction stage, which identifies the configuration values responsible for the violation. This evidence may include public CIDRs, exposed ports, wildcard permissions, MFA status, credential state, public S3 access, and disabled logging or recovery controls.

### *C. Evidence Consolidation and Risk Prioritization*

One configuration weakness may trigger multiple policies and create overlapping findings. The system therefore consolidates related evidence before risk scoring. Findings are grouped by the same resource and underlying security condition, while contributing policy identifiers are retained.

Risk prioritization is based on likelihood and impact. This follows the general risk-assessment principle described in NIST SP 800-30 [14]. Likelihood in this study is not treated as a measured probability of compromise. It is a configuration-derived estimate. Observable evidence is used to determine the likelihood level. Examples include public exposure, missing MFA, active credentials, broad trust relationships, and the accessibility of the affected resource. Likelihood and impact are each assigned a value from 1 to 5. Their multiplication produces a raw risk score ranging from 1 to 25. A scaling factor of 4 is applied because the maximum raw score of 25 multiplied by 4 produces a maximum score of 100. This converts the risk score to a more interpretable 100-point scale while preserving the relative ordering of the findings. The risk score is therefore calculated as:

$$Risk\ Score = Likelihood \times Impact \times 4 \quad (1)$$

The resulting score ranges from 4 to 100. Scores from 80 to 100 are classified as Critical, 60 to 79 as High, 40 to 59 as Medium, 20 to 39 as Low, and below 20 as Informational. The 1–5 scales, scaling factor, and severity bands are defined by this

study. NIST SP 800-30 provides the basis for considering likelihood and impact but does not prescribe this specific scoring formula [14].

Several evidence factors may support the same likelihood or impact level. In such cases, the strongest supported level is retained. Policy scores are not combined, which prevents related rules from artificially increasing the risk score and keeps prioritization linked to observable evidence.

*D. Threat Context and Evaluation Design*

MITRE ATT&CK is applied after risk calculation to provide supporting threat context [27]. Relevant findings are mapped to ATT&CK techniques associated with the observed configuration weakness. The mapping does not indicate detected adversary activity and does not affect the risk score.

The prototype is evaluated in two stages. First, individual Policy as Code rules are tested using the CloudFormation Guard unit-testing capability [32]. Each test contains input data and an expected rule outcome. This verifies whether individual rules return the expected PASS or FAIL result.

The complete system is evaluated using controlled secure and misconfigured AWS scenarios. Expected results are defined before testing and compared with the prototype findings. Detection outcomes are recorded as true positives, true negatives, false positives, and false negatives. Risk prioritization is evaluated separately by comparing the expected and generated likelihood, impact, risk score, and severity. Resource counts and execution time are also recorded to assess prototype performance.

## IV. IMPLEMENTATION AND EVALUATION

*A. Prototype Implementation*

The prototype was implemented in Python using a modular pipeline for configuration collection, policy evaluation, evidence extraction, risk prioritization, MITRE ATT&CK mapping, and reporting. Configuration data from IAM, S3, and EC2 Security Groups are normalized into JSON before policy evaluation.

The Policy as Code layer contains 18 CloudFormation Guard policies: five for Security Groups, eight for IAM, and five for S3. These policies cover selected issues such as unrestricted access, excessive IAM permissions, missing MFA, public S3 access, disabled versioning, logging, and incomplete Block Public Access.

CloudFormation Guard evaluates the normalized configuration data against the relevant policy files. The Guard validate operation determines whether the input data satisfy or violate the defined rules. A PASS result indicates that the evaluated configuration satisfies the policy condition. A FAIL result indicates a policy violation.

Failed evaluations are passed to service-specific evidence extractors. The extractors identify the configuration values that contributed to each violation. Related evidence is then consolidated when several policies describe the same underlying weakness. The resulting findings are processed by the likelihood- and impact-based risk model described in Section III. MITRE ATT&CK mappings are added only as supporting threat context [27].

The final findings are presented through a Streamlit based dashboard. The dashboard shows the affected resource, contributing policies, configuration evidence, likelihood, impact, risk score, severity, and relevant ATT&CK context.

*B. Policy-Level Verification*

Policy level verification was performed before the complete pipeline was evaluated. Test conditions and expected results were defined before execution. CloudFormation Guard also provides a built-in unit-testing framework for testing Guard rules against predefined JSON or YAML test data [32]. The framework allows expected rule outcomes to be specified before the test is executed. All 18 Guard policies were tested with both violating and non-violating configurations. The five Security Group policies were evaluated using 39 test cases. The eight IAM policies were evaluated using 88 test cases. The five S3 policies were evaluated using 50 test cases. A total of 177 predefined test cases were executed. All 177 tests produced the expected outcomes.

TABLE I
POLICY-LEVEL VERIFICATION RESULTS

| Service | Policies | Test Cases | Expected Outcomes Satisfied |
|---|---|---|---|
| Security Groups | 5 | 39 | 39/39 |
| IAM | 8 | 88 | 88/88 |
| S3 | 5 | 50 | 50/50 |
| Total | 18 | 177 | 177/177 |

These results verify that the implemented Guard rules behaved as expected for the predefined test inputs. They are not interpreted as evidence of overall system accuracy. Policy-level unit tests evaluate individual rule behavior and do not measure complete end-to-end detection performance.

*C. Prototype Scan Result*

A live AWS scan produced three prioritized findings: one IAM and two S3 findings. Four overlapping IAM violations were consolidated into one High-risk finding with a likelihood of 3, impact of 5, and risk score of 60. The S3 findings represented disabled versioning and disabled server access logging, with scores of 24 (Low) and 8 (Informational), respectively. No Security Group violation was detected. These results show that final severity is derived from configuration evidence rather than directly from a failed policy.

*D. End-to-End Validation*

Fig. 2 shows three controlled scenarios evaluated for Security Groups, S3, and IAM using misconfigured resources and secure negative controls. The Security Group scenario produced one Medium finding for public SSH access with a score of 48. The S3 scenario produced one Critical finding for public access with a score of 100 and one Low finding for incomplete Block Public Access with a score of 24. The IAM scenario consolidated four overlapping violations into one High finding with a score of 60.

Across 26 policy-resource evaluation units, the prototype produced 8 true positives and 18 true negatives, with no false

positives or false negatives. Precision, recall, and accuracy were 1.00, while the false-positive rate was 0.00. All expected risk classifications were also reproduced. These results apply only to the controlled scenarios and should not be interpreted as perfect accuracy in production AWS environments.

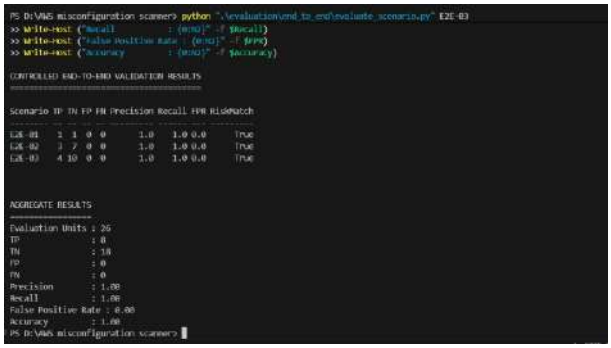



Fig. 2. Controlled end-to-end validation results across the three evaluation scenarios.

## V. DISCUSSION

The evaluation shows that the proposed workflow connects Policy as Code violations with resource-specific evidence for risk prioritization. A failed policy identifies a configuration problem, while the evidence explains its security significance and supports the assigned likelihood and impact values. Evidence consolidation also reduces duplicate findings. In the S3 and IAM scenarios, overlapping policy violations were grouped when they represented the same underlying condition, while separate weaknesses remained independent. This prevents related failures from increasing the final risk score unnecessarily. The results were consistent across the selected IAM, S3, and Security Group scenarios. The prototype evaluates only 18 selected policies and was tested using controlled AWS scenarios. Therefore, the results demonstrate system behavior within the defined research scope and should not be interpreted as production-level detection accuracy. Likelihood values are based on observable configuration evidence rather than measured probabilities of compromise. The current risk model does not include organization-specific factors such as workload criticality or data sensitivity, because these were outside the defined scope. The prototype was not benchmarked against established CSPM tools and should therefore be interpreted within the defined research scope.

## VI. CONCLUSION AND FUTURE WORK

This study developed a Policy as Code-based prototype for detecting and prioritizing AWS IAM, S3, and EC2 Security Group misconfigurations. The system combines read-only AWS configuration collection, CloudFormation Guard evaluation, evidence extraction, evidence consolidation, and likelihood-impact risk prioritization. Overlapping policy violations are consolidated before risk scoring so that findings are based on resource-level evidence rather than individual policy failures.

The prototype contains 18 policies. All 177 predefined Guard test cases produced the expected outcomes. End-to-end evaluation across 26 policy-resource units produced 8 true positives and 18 true negatives, with no false positives or false negatives in the controlled scenarios. The expected risk scores and consolidation behavior were also reproduced.

Only 18 policies were checked, and future work can extend the policy set and evaluate the system in larger AWS environments. Additional context such as workload criticality, data sensitivity, resource relationships, and network reachability could improve risk estimation. Future versions may also explore AI-assisted policy generation, remediation guidance, organization-level scanning, historical analysis, automated remediation, and comparison with established CSPM tools.

## REFERENCES

[1] A. Hassan, S. Bhatti, S. Shujaat and Y. Hwang, "To adopt or not to adopt? The determinants of cloud computing adoption in information technology sector," *Decision Analytics Journal,* p. 100138, 2022.

[2] Synergy Research Group, "Cloud Market Annual Revenue Run Rate Topped Half a Trillion Dollars in Q1 as Growth Surge Continues," 29 April 2026. [Online]. Available: https://www.srgresearch.com/articles/cloud-market-annual-revenue-run-rate-topped-half-a-trillion-dollars-in-q1-as-growth-surge-continues. [Accessed 08 August 2026].

[3] Amazon Web Services, "Resources for Small and Medium Businesses," Amazon Web Services, [Online]. Available: https://aws.amazon.com/smart-business/resources-for-smb/. [Accessed 08 August 2026].

[4] Amazon Web Services, "Cloud Computing for Enterprise Businesses," Amazon Web Services, [Online]. Available: https://aws.amazon.com/enterprise/. [Accessed 08 August 2026].

[5] Flexera, "2026 State of the Cloud Report," Flexera, 2026.

[6] Amazon Web Services, "Shared Responsibility Model," Amazon Web Services, [Online]. Available: https://aws.amazon.com/compliance/shared-responsibility-model/. [Accessed 08 August 2026].

[7] Amazon Web Services, "Security in IAM and AWS STS," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/IAM/latest/UserGuide/security.html. [Accessed 08 August 2026].

[8] Amazon Web Services, "Security in Amazon S3," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/AmazonS3/latest/userguide/security.html. [Accessed 08 August 2026].

[9] Amazon Web Services, "Security in Amazon EC2," [Online]. Available: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-security.html. [Accessed 08 August 2026].

[10] T. Kennedy A, S. Muhammad I. H., F. Cheng and C. Meinel, "Continuous auditing and threat detection in multi-cloud infrastructure," *Computers & Security,* vol. 102, p. 102124, 2021.

[11] Cloud Security Alliance, "A Technical Analysis of the Capital One Cloud Misconfiguration Breach," Cloud Security Alliance, 09 August 2019. [Online]. Available: https://cloudsecurityalliance.org/blog/2019/08/09/a-technical-analysis-of-the-capital-one-cloud-misconfiguration-breach. [Accessed 08 August 2026].

[12] H. Margareta, G. Jin P and B. Peter, "Investigating the Security Divide between SME and Large Companies: How SME Characteristics Influence Organizational IT Security Investments," *Information Systems Frontiers,* vol. 21, no. 6, pp. 1285-1305, 2019.

[13] C. R. China, "What is policy as code (PaC)?," IBM, 18 June 2026. [Online]. Available: https://www.ibm.com/think/topics/policy-as-code. [Accessed 08 August 2026].

[14] National Institute of Standards and Technology, "Guide for Conducting Risk Assessments," National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 2012.

[15] C. Nobles, "Investigating Cloud Computing Misconfiguration Errors using the Human Factors Analysis and Classification System," *Scientific Bulletin,* vol. 27, no. 1, p. 59–66, 2022.

[16] B. Daniele, M. Guido, S. Riccardo and V. Fulvio, "Automation for Network Security Configuration: State of the Art and Research Trends," *ACM Computing Surveys,* vol. 56, no. 3, p. 1–37, 2023.

[17] B. Daniele, M. Guido, S. Riccardo, V. Fulvio and Y. Jalolliddin, "Automated Firewall Configuration in Virtual Networks," *IEEE Transactions on Dependable and Secure Computing,* vol. 20, no. 2, p. 1559–1576, 2023.

[18] K. Bom and L. Seungsoo, "KubeAegis: A Unified Security Policy Management Framework for Containerized Environments," *IEEE Access,* vol. 12, p. 160636–160652, 2024.

[19] Amazon Web Services, "What is AWS CloudFormation Guard?," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/cfn-guard/latest/ug/what-is-guard.html. [Accessed 09 August 2026].

[20] Amazon Web Services, "What Is AWS Config?," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/config/latest/developerguide/WhatIsConfig.html. [Accessed 09 August 2026].

[21] Amazon Web Services, "Introduction to AWS Security Hub CSPM," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/securityhub/latest/userguide/what-is-securityhub.html. [Accessed 09 August 2026].

[22] Prowler, "Introduction," Prowler, [Online]. Available: https://docs.prowler.com/introduction. [Accessed 09 August 2026].

[23] Checkov, "What is Checkov?," Checkov, [Online]. Available: https://www.checkov.io/1.Welcome/What%20is%20Checkov.html. [Accessed 09 August 2026].

[24] Amazon Web Services, "Security best practices in IAM," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/IAM/latest/UserGuide/best-practices.html. [Accessed 09 August 2026].

[25] Amazon Web Services, "Security best practices for Amazon S3," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/AmazonS3/latest/userguide/security-best-practices.html. [Accessed 09 August 2026].

[26] Amazon Web Services, "Security Hub CSPM controls for Amazon EC2," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/securityhub/latest/userguide/ec2-controls.html. [Accessed 09 August 2026].

[27] The MITRE Corporation, "Matrix - Enterprise - Cloud," The MITRE Corporation, [Online]. Available: https://attack.mitre.org/matrices/enterprise/cloud/. [Accessed 09 August 2026].

[28] P. Ken, T. Tuure, R. Marcus A and C. Samir, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems,* vol. 24, no. 3, p. 45–77, 2007.

[29] Amazon Web Services, "Security Hub CSPM controls for AWS Identity and Access Management," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/securityhub/latest/userguide/iam-controls.html. [Accessed 09 August 2026].

[30] Amazon Web Services, "Security Hub CSPM controls for Amazon S3," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/securityhub/latest/userguide/s3-controls.html. [Accessed 09 August 2026].

[31] Amazon Web Services, "Boto3 documentation," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/boto3/latest/. [Accessed 09 August 2026].

[32] Amazon Web Services, "AWS CloudFormation Guard," Amazon Web Services, [Online]. Available: https://docs.aws.amazon.com/cfn-guard/latest/ug/cfn-guard-test.html. [Accessed 09 August 2026].

# A Comparative Evaluation of Regression and Classification Approaches for Student Performance Prediction Using Machine Learning

R.K.A.I.C. Gunarathna
*Department of Computer and Data Science*
*NSBM Green University*
*Mahenwaththa, Pitipana, Homagama, Sri Lanka*
aicgunarathna@students.nsbm.ac.lk

R.W.K.T. Rajapaksha
*Department of Computer and Data Science*
*NSBM Green University*
*Mahenwaththa, Pitipana, Homagama, Sri Lanka*
kavishka.r@nsbm.ac.lk

***Abstract*—Student academic performance is influenced by academic, behavioural, and lifestyle-related factors. This study presents an empirical comparison of regression and direct classification approaches for predicting university student performance using primary survey data. A structured online questionnaire collected approximately 250 responses, of which 242 valid records were retained after data cleaning. The dataset initially contained 24 candidate student-related factors. Data encoding issues involving screen time and multi-select study methods were corrected, and permutation feature importance was used to identify the ten most predictive factors. The prepared dataset was divided into 193 training and 49 testing records. Multiple regression and classification algorithms, together with Voting and Stacking ensembles, were evaluated. The selected Voting Hybrid Regressor achieved a test MAE of 0.2121, RMSE of 0.2798, and $R^2$ of 0.6016. The selected XGBoost classifier achieved 69.39% test accuracy and a Macro F1-score of 0.7059 across the three classes represented in the test outcomes and predictions. However, the dataset contained only one Poor-category record, which was assigned to the training subset; consequently, the test set contained no Poor examples. The common-basis evaluation showed that direct XGBoost classification achieved higher category accuracy (69.39%) than regression-derived category assignment (63.27%) on the same test records. The findings demonstrate the practical potential of both prediction formulations while highlighting the need for a larger and more balanced dataset, particularly to evaluate the underrepresented Poor category.**

***Keywords—student performance prediction, machine learning, XGBoost, hybrid ensemble feature selection, regression versus classification***

## I. Introduction

### A. Background

Student academic performance is an important concern for students, educators, and higher-education institutions. Educational Data Mining and Learning Analytics provide methods for examining student information and identifying patterns associated with academic outcomes. This study investigates the prediction of university student academic performance using academic, behavioural, and lifestyle-related information collected through a structured questionnaire. Two prediction formulations are examined. The regression approach estimates Grade Point Average (GPA) as a continuous numerical value, whereas the direct classification approach predicts a predefined academic performance category. Evaluating both formulations using the same student dataset and selected input factors provides an opportunity to examine their respective predictive results and practical uses. The study also considers the effect of data preparation, feature selection, model selection, and repeated train-test evaluation on the resulting performance.

### B. Problem Statement

Existing studies in Educational Data Mining have applied a range of machine learning techniques to predict student academic outcomes. However, studies may differ in their datasets, selected variables, target definitions, preprocessing procedures, and evaluation measures, making their reported results difficult to compare directly. Furthermore, predicting a continuous GPA and predicting an academic performance category are related but distinct tasks. A model that performs well for GPA estimation does not necessarily provide the best direct category prediction. The present study addresses this issue through an empirical comparison of regression and direct classification formulations using the same primary student dataset and selected predictive factors. It also examines survey data encoding, feature importance, individual algorithms, hybrid ensembles, and performance variation across repeated random train-test splits. The study further identifies limitations arising from the small sample size and the severe under-representation of the Poor performance category. These considerations define the scope of the investigation and the extent to which its findings can be generalised.

### C. Aim and Objectives

The aim of this study is to develop and evaluate a machine-learning-based system for predicting university student academic performance by investigating regression and classification formulations and their suitability for practical deployment.

Identify academic, behavioural, lifestyle, and support-related factors relevant to student performance.

Prepare the collected data, correct encoding issues, and select predictive factors using feature-importance analysis.

Develop and compare individual regression and classification algorithms, together with Voting and Stacking ensembles.

Evaluate both formulations using task-appropriate metrics and examine performance variation across repeated random train-test splits.

Integrate the selected models into a web application that provides GPA estimates, performance categories, and recommendations.

## II. LITERATURE REVIEW

### A. Educational Data Mining and Student Performance

Educational Data Mining applies data mining and machine-learning techniques to educational data to identify patterns that may support educational decision-making [1]. Previous research has investigated factors such as attendance, study habits, and family-related information in predicting academic outcomes [2], [3], [4]. These studies demonstrate the value of student-related data; however, their findings depend on the population, available variables, target definitions, and evaluation procedures used.

### B. Machine Learning Models

Linear Regression provides an interpretable baseline for continuous prediction, while regularised approaches such as Ridge, Lasso, and Elastic Net help control model complexity [5]. Random Forest combines decision trees to model non-linear relationships [6], [7]. Gradient Boosting and XGBoost develop sequential tree-based models to improve predictive performance [8], [9]. Support Vector methods provide alternative approaches for regression and classification. Voting and Stacking ensembles further allow predictions from multiple models to be combined.

### C. Regression Versus Classification Formulations

Student performance can be represented as a continuous GPA prediction or as a discrete performance category. These formulations address related but distinct prediction tasks and require different evaluation metrics. Although previous studies have examined individual prediction algorithms and ensemble approaches [2], [7], [9], comparing results across different datasets and evaluation settings does not establish which formulation is more suitable for a particular student dataset.

### D. Research Gap and Contribution

Previous studies have demonstrated the application of machine learning to educational prediction [2] – [9]. However, differences in student populations, input variables, target definitions, and evaluation procedures make their findings difficult to compare directly. In addition, regression-based GPA estimation and direct performance-category classification provide different types of output. This study addresses these issues through an empirical investigation of both formulations using the same primary dataset and selected predictive factors. It evaluates individual algorithms and hybrid ensembles, examines feature importance, and investigates performance variation across 30 random train-test splits. The contribution is an application-oriented comparison and implementation of the selected models in a web-based student performance prediction system, rather than the introduction of a new algorithm.

## III. METHODOLOGY

### A. Research Design

This study adopts a quantitative, survey-based, and experimental comparative design. The machine-learning workflow was organised with reference to the six CRISP-DM stages [10]. Business understanding defined the need to estimate student GPA and identify performance categories. Data understanding involved examining the questionnaire responses and their distributions. Data preparation included cleaning, encoding correction, and feature selection. Modelling involved training individual regression and classification algorithms and hybrid ensembles. Evaluation used task-specific metrics and repeated random train-test splits. Deployment involved integrating the selected models into the StudentPulse web application. General data-mining principles also informed the preparation and modelling process [11].

### B. Data Collection and Correction

Primary data were collected through a structured Google Forms questionnaire distributed to university students. Approximately 250 responses were received, of which 242 valid records were retained after cleaning. The questionnaire covered academic, behavioural, lifestyle, and support-related factors, including university, academic year, GPA, attendance, study hours, extracurricular activities, sleep, stress, parental support, study-life balance, screen time, and study methods. The dataset contained 24 candidate predictive factors, with GPA used as the regression target and to derive four performance categories. Two encoding issues were corrected: ScreenTime values were converted from time-formatted strings into numerical hours, and multi-select StudyMethod responses were separated into binary indicators. Ordinal responses were encoded according to their recorded order. Data preparation was performed using Python [12].

### C. Feature Selection

Permutation feature importance, computed using a Random Forest model after the encoding corrections described above, was used to select the ten most predictive factors from twenty-four candidates: Attendance, Study-Life Balance, Past Papers, Academic Year, Extracurricular Activity, Screen Time, Early Preparation, Stress Level, Parental Support, and Stress Management. Attendance achieved the highest permutation importance score (0.8193), exceeding the combined scores of the other nine selected factors (0.3510). These scores indicate predictive importance within the fitted model and do not establish causal relationships.

### D. Train-Test Split and Preprocessing

The dataset was divided into 80% training (193 records) and 20% testing (49 records), using a fixed random seed applied consistently across both formulations for direct comparability. Categorical variables were one-hot encoded, and ordinal variables were encoded to preserve their natural ordering. All preprocessing was fit on the training data only.

### E. Models and Evaluation

Two parallel model sets were trained on the same ten features using the scikit-learn library [13]. The regression set comprised Linear Regression, Ridge, Lasso, Elastic Net, Random Forest, Gradient Boosting, XGBoost, and Support Vector Regression, evaluated using MAE, RMSE, and $R^2$. The classification set comprised Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Support Vector Classification, evaluated using accuracy and macro-averaged F1-score. Voting and Stacking hybrid ensembles were evaluated for both sets.

F. Hyperparameter Tuning

Hyperparameter tuning was performed using RandomizedSearchCV [14] on the training data, searching over the number of estimators, learning rate, maximum tree depth, minimum child weight, subsample ratio, and column-sampling ratio for tree-based models.

G. Statistical Stability Analysis

To assess whether the best result in each formulation reflected a reproducible pattern rather than a favorable single split, the train-test split was repeated across 30 random seeds, the best-performing model type was refitted on each resulting training set, and performance was recorded on the corresponding test set each time.

## IV. RESULTS

A. Feature Importance

Table I shows the ten selected factors ranked by permutation importance. Attendance received the highest importance score, indicating its strong predictive contribution within the fitted Random Forest model.

TABLE I
SELECTED FACTORS RANKED BY IMPORTANCE

| Rank | Factor | Importance |
|---|---|---|
| 1 | Attendance | 0.8193 |
| 2 | Study-Life Balance | 0.0623 |
| 3 | Past Papers | 0.0568 |
| 4 | Academic Year | 0.0481 |
| 5 | Extracurricular | 0.0441 |
| 6 | Screen Time | 0.0383 |
| 7 | Early Preparation | 0.0361 |
| 8 | Stress Level | 0.0233 |
| 9 | Parental Support | 0.0214 |
| 10 | Stress Management | 0.0206 |

B. Regression Formulation

The Voting Hybrid Regressor achieved the best reported regression performance, with an MAE of 0.2121, RMSE of 0.2798, and $R^2$ of 0.6016 on the 49-record test set. This indicates that the model explained 60.16% of the observed GPA variance in the test data.

TABLE II
BEST REGRESSION MODEL PERFORMANCE

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Voting Hybrid | 0.2121 | 0.2798 | 0.6016 |

C. Classification Formulation

The baseline XGBoost classifier achieved 69.39% accuracy (34/49 correct predictions) and a Macro F1-score of 0.7059 across the three classes represented in the test outcomes and predictions. Only one record in the full dataset belonged to the Poor category, and it was assigned to the training set. Therefore, the test set contained no Poor examples, preventing evaluation of the model's ability to identify this category.

TABLE III
BEST CLASSIFICATION MODEL PERFORMANCE

| Model | Accuracy | F1-Macro |
|---|---|---|
| XGBoost | 69.39% | 0.7059 |

D. Common-Basis Evaluation: Direct vs. Derived Categorization

On the identical 49-record test set, direct XGBoost classification outperformed the Voting Regressor with GPA-threshold mapping, achieving higher category accuracy (69.39% vs. 63.27%) and represented-class Macro F1-score (0.7059 vs. 0.6431.

TABLE IV
COMMON-BASIS CATEGORICAL EVALUATION

| Prediction formulation | Correct predictions | Category accuracy | Macro F1 |
|---|---|---|---|
| XGBoost classification | 34/49 | 69.39% | 0.7059 |
| Voting regression | 31/49 | 63.27% | 0.6431 |

E. Statistical Stability Analysis

The train-test procedure was repeated across 30 random seeds. For each split, the selected model was refitted using the training subset and evaluated on the corresponding test subset. The mean and standard deviation of the evaluation scores were calculated to examine performance variation across different data partitions. This analysis does not constitute independent external validation.

TABLE V
STABILITY ACROSS 30 REPEATED SPLITS

| Formulation | Mean | SD |
|---|---|---|
| Regression, $R^2$ | 0.6475 | 0.0764 |
| Classification, Accuracy | 70.48% | 4.07 pp |

### F. Class Distribution and Confusion Matrix Analysis

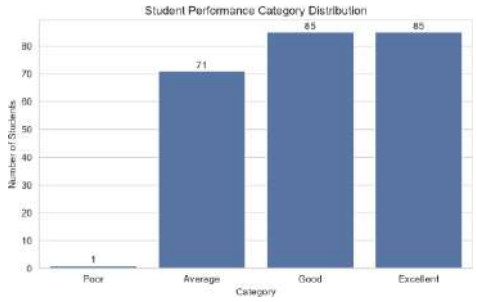



Fig.1. Distribution of performance categories across the 242-record dataset, showing the scarcity of the Poor category.

Fig.1 shows the distribution of performance categories across the 242 records: one Poor, 71 Average, 85 Good, and 85 Excellent. Since the only Poor-category record was assigned to the training set, the test set contained no Poor examples, limiting the evaluation of this category. Fig. 2 presents the confusion matrix for the final XGBoost classifier.

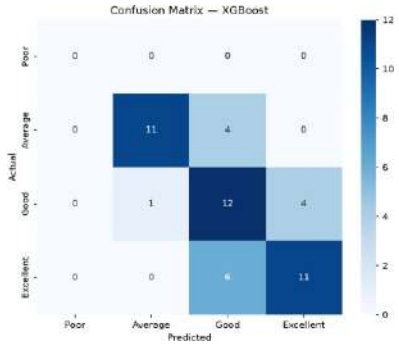



Fig. 2. Confusion matrix for the final XGBoost classification model on the held-out test set.

The severe class imbalance limits reliable evaluation across all four categories. When the Poor class is included with an F1-score of zero, the four-class Macro F1-score is 0.5294. A larger, more balanced dataset is therefore needed for reliable evaluation.

## V. DISCUSSION

The results show that regression and classification provide different but useful outputs: estimated GPA and predicted performance category. The common-basis evaluation showed that direct XGBoost classification performed better than regression-derived category prediction on the same test set. Attendance had the highest permutation importance (0.8193), while the 30-split evaluation demonstrated performance variation across different data partitions. However, the limited representation of the Poor category restricts the interpretation of the classification results. Therefore, the findings apply to the collected dataset and should not be generalised to all university students.

## VI. WEB APPLICATION INTEGRATION

The selected models were integrated into a web application using React for the frontend, Flask for the backend, and MySQL for data storage. Student input is submitted to the Flask prediction endpoint, where shared preprocessing logic prepares the required features. The Voting Hybrid Regressor generates the estimated GPA, while the baseline XGBoost classifier predicts the performance category. Both outputs are returned to the React dashboard. The application also presents personalised guidance using predefined application rules.

## VII. CONCLUSION

This study evaluated regression and classification approaches using 242 valid student records. The Voting Hybrid Regressor achieved an MAE of 0.2121, RMSE of 0.2798, and $R^2$ of 0.6016, while the baseline XGBoost classifier achieved 69.39% test accuracy. Evaluation across 30 random splits provided further insight into model stability. However, the limited dataset and class imbalance restrict generalisation, particularly for the Poor category. Future work will focus on larger datasets, improved class representation, and independent validation.

## VIII. LIMITATIONS AND FUTURE WORK

The study is limited by its relatively small dataset of 242 records and the severe imbalance of the performance categories, particularly the single Poor-category observation. The questionnaire also relies on self-reported information, and repeated random splits do not replace independent external validation. Future work will focus on collecting a larger and more balanced dataset across different university programmes, evaluating the models using independent student data, and improving the prediction system and its recommendations.

## REFERENCES

[1] H. Witten, E. Frank, M. A. Hall and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Cambridge: Morgan Kaufmann, 2016.

[2] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance," in Proceedings of the 5th Future C. Romero and S. Ventura, "Data Mining in Education," R. S. J. d. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," Journal of Educational Data Mining, vol. 1, no. 1, p. 3–17

[3] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, New York: Springer, 2009.

[4] L. Breiman, "Random forests," Machine Learning, vol. 45, p. 5–32,

[5] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, p. 832–844, 1998.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, p.

[8] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," Journal of Data Warehousing, vol. 5, no. 4, p. 13–22, 2000.

[9] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Waltham: Morgan Kaufmann, 2011.

[10] W. McKinney, Python for Data Analysis, 3rd ed., Sebastopol: O'Reilly Media, 2022.

[11] Pedregosa, F. et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, p. 2825–2830, 2011.

[12] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, San Francisco, 1995.

# The Location-Identity Problem in Automated Land Valuation: Satellite Flood Hazard Variables and GN-Division Spatial Validation in Sri Lanka

P. A. K. Dilshan
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
pakdilshan@students.nsbm.ac.lk

Hirushi Dilpriya
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Automated valuation models for Sri Lanka carry no hazard variables, and none of the property valuation studies reviewed here reports spatially-blocked validation. This paper integrates satellite-derived flood hazard and terrain variables into a land price model for Colombo District, Sri Lanka, at Grama Niladhari (GN) division resolution, and measures how much of the model's apparent accuracy depends on having seen a division before. Six hazard and terrain variables, screened under rules fixed in advance, enter a tuned Random Forest on 49 features that predicts log price per perch for 7,034 listings across 194 GN divisions. One model specification and one preprocessing pipeline, with price filtering, per-division trimming and target encoding all fitted inside each training fold, are evaluated under three protocols. On divisions seen in training (random 80/20 split, five seeds) mean $R^2$ is 78.97%; on divisions absent from training (grouped by division, 25 randomised partitions) it is 27.90% (central 95% range 15.93–38.21%), a gap of 51.07 percentage points. The gap persists under the original global preprocessing (87.72% against 30.28%). The target-encoded division identifier carries 75.40% of Gini importance and ranks 1st under permutation importance on a random split, but 34th of 49 on unseen divisions. Removing all six hazard variables changes unseen-division $R^2$ by −0.37 percentage points (bootstrap 95% interval −0.74 to −0.01), which does not meet the study's pre-specified decision rule: any hazard contribution is small and not robust. A neighbour-price baseline that prices a division from its five nearest training divisions reaches 45.57% on unseen divisions, and replacing the identity feature with that neighbour price lifts the Random Forest to 41.98%. Much of the accuracy reported for this problem class therefore reflects location identity rather than location description.**

***Keywords*—land price prediction, spatial cross-validation, target encoding, flood hazard, Random Forest**

## I. Introduction

Land valuation in Sri Lanka is conducted manually by officers of the Valuation Department, using comparative, contractor's, residual and investment approaches [1]. No publicly accessible transaction price database exists, so price evidence is scattered across web listing platforms and automated valuation research is built on asking prices rather than transacted ones. Internationally, such models have moved toward tree-based ensembles [2]–[5], but the integration of climate hazard risk remains largely unaddressed. Hedonic pricing theory holds that rational buyers discount property in hazardous locations, and evidence from the United States indicates that flood-exposed residential property is systematically overvalued where risk is not capitalised into price [6]. Whether an equivalent discount is observable in a South Asian developing-country land market is open: the principal prior Sri Lankan land price model restricts its features to market and amenity variables [7].

A second gap is methodological, and this study treats it as the more consequential. Where location is represented by an administrative unit rather than a plot coordinate, every location-derived feature is constant within that unit, so a random train–test split places the same unit on both sides of it. The model may then recover a unit's price level from features that identify the unit rather than describe it, and the reported accuracy measures interpolation within known locations rather than prediction for new ones. None of the property valuation studies reviewed here reports a spatially-blocked design, so the size of this effect is undocumented.

The study centres on Colombo District at GN division resolution for three empirical reasons. First, the hazard signal is destroyed at coarse resolution: an earlier five-district construction, resolving 12,870 listings to 880 town-level points, returned grouped cross-validation $R^2$ = 0.4531 with the full feature set against 0.4489 without the climate variables. Second, GN resolution requires both a boundary register and sufficient listing density, which Ratnapura, the highest-hazard district, lacks: only 61 real listings across 12 distinct locations are available. Third, Colombo is not hazard-free: the Kelani Ganga floodplain runs through its eastern divisions, and the flood events of June 2021, October 2024 and Ditwah 2025 are observable in Sentinel-1 imagery. The trade-off is explicit: the district spans a narrower hazard spectrum, so any discount is measured across a compressed range of exposure. This paper makes three contributions. It screens satellite-derived hazard and terrain variables at GN division centroids under rules fixed in advance, rejecting two candidates on validity grounds even though they passed statistical screening. It reports, to our knowledge, the first spatially-blocked evaluation of a Sri Lankan land valuation model, holding model, features and a fold-internal preprocessing pipeline fixed so that the gap between seen-division and unseen-division accuracy is attributable to the validation design. And it shows that the target-encoded division identifier, carrying 75.40% of the model's Gini importance, loses its importance in an unseen division,

while describing location by the prices of neighbouring divisions recovers much of the lost accuracy.

## II. Related Work

Automated valuation models have moved toward tree-based ensembles [2]–[5], while a separate literature establishes that physical climate risk is progressively but incompletely priced into residential [6] and commercial [8] property. The two have developed largely independently. Hedonic discounts also require buyers to act on accurate risk information, yet Peiris *et al.* show that communities in the flood-prone Kelani River lower basin have limited awareness of formal hazard zone classifications [9]. A negative coefficient therefore does not by itself demonstrate risk capitalisation: low-lying land in Colombo is also older, denser and closer to the historic core.Sri Lankan property research is concentrated within Colombo: Pushpakumara *et al.* map spatial variation in land prices but carry no environmental risk variables [10], and Naotunna reaches state-of-the-art accuracy for Colombo using Random Forest, Extra Trees, SVR, XGBoost and multiple linear regression, but is restricted to market and amenity features [7]. Across the studies reviewed here [2]–[5], [7], [10], [11], none carries hazard variables, and every predictive study reports a random train–test split rather than a spatially-blocked design. The review is targeted rather than systematic, so this is a statement about these studies, not the whole literature.

That omission is consequential. Property prices are spatially autocorrelated [12], but the consequence for model evaluation has not transferred into machine-learning valuation. Roberts *et al.* argue that when observations are spatially structured, random cross-validation is optimistically biased [13], and Ploton *et al.* show ecological mapping models collapsing under spatial cross-validation [14]. Here the leak is stronger than proximity: every listing in a GN division carries that division's centroid, so all location features are exactly constant within it and the correct blocking unit is the division itself. Spatially aware alternatives such as geographically weighted regression, kriging and spatial-lag models need a point location per observation; with one centroid per division they reduce to division-level smoothing, which is tested here as a neighbour-price baseline. Both figures are reported here, since many real valuation queries concern divisions for which comparable sales already exist.Location also enters as a high-cardinality categorical. Micci-Barreca proposed replacing each category with a smoothed target statistic and identified the leakage risk directly [15]. Out-of-fold computation is the standard remedy, applied here inside every fold, but it is incomplete for spatial data: it stops a row seeing its own label, not the labels of other rows in the same location, which under a random split is nearly as informative.

Synthetic aperture radar is the primary flood-mapping sensor because it penetrates cloud [16], but under-detects inundation beneath dense built-up cover; the digital-surface-model [17] and interferometric-coherence [18], [19] remedies are not implemented here. Openshaw's modifiable areal unit problem [20] bears on the location unit: one large enough to contain both floodplain and adjacent high ground averages the hazard away before modelling begins. The Sri Lankan valuation literature [1], [7], [10], [11] carries no hazard variables, and the Sri Lankan SAR flood literature [16]–[19], [21] is not linked to land value; spatially-blocked validation is mature but resides in ecology [13], [14]. This study performs the transfer.

## III. Methodology

### A. Study Area and Data Sources

The study is confined to Colombo District, Sri Lanka. Price, land size, land type and location were scraped from Ikman.lk, the largest online property marketplace in Sri Lanka by listing volume, for which academic permission was obtained before collection. Amenity coordinates were obtained from the OpenStreetMap Overpass API across eight categories. Hazard and terrain variables were drawn from four raster sources and one vector source (Table I).

### B. Location Resolution

Listings are not geocoded to plot coordinates. Each is resolved to a Grama Niladhari (GN) division, the smallest administrative unit in Sri Lanka, and assigned that division's polygon centroid; 557 raw GN records from the level-4 national boundary dataset are consolidated to 399 named divisions after alias merging. Addresses are matched by division name and landmark string, with unresolved cases passed to a bounded geocoding fallback whose failures are logged rather than imputed. The 7,034 retained listings resolve to 194 distinct GN divisions. Because all listings in a division share one centroid, every geographic, amenity and hazard feature is constant within it: the data contain 7,034 listings but 194 distinct feature vectors. Divisions are identified by `Address`, a lower-cased GN division name, which is the grouping key throughout.

### C. Hazard Variable Extraction and Screening

Hazard and terrain variables are extracted at the 194 division centroids, each summarised as the mean over a 1 km circular buffer. Eleven candidates were screened under three rules fixed in advance: drop any variable with fewer than 30 distinct values across the 194 divisions; drop any correlating above 0.8 in absolute value with `Distance_from_fort`, already in the model; and where two survivors exceed that threshold, retain the more interpretable. Four were dropped on these statistics: `slope_deg` ($r$ = +0.912 with `elevation_m`), `built_up_fraction` and `tree_cover_fraction` (both restating `Distance_from_fort`), and NASA LHASA `landslide_susceptibility` (four distinct classes). MODIS flood frequency passed both rules but was rejected on validity: it was zero in 159 of the 194 divisions with every non-zero value coastal, measuring tidal rather than river water. JRC water recurrence, extracted outside this set, failed a pre-specified geography test: the Kelani floodplain divisions ranked a mean 60.9 of 194 against 124.8 for the southern uplands. Six variables were retained (Table I). None of the rules uses price, so screening on all divisions cannot carry target information into any fold.

TABLE I. RETAINED HAZARD AND TERRAIN VARIABLES. ALL SIX ENTER THE MODEL AND THE HAZARD ABLATION

| Variable | Source | Distinct values (of 194) | *r* vs `Distance_from_fort` |
|---|---|---|---|
| `elevation_m`† | Copernicus DEM GLO-30, 30 m | 194 | +0.673 |
| `water_occurrence_pct` | JRC Global Surface Water, occurrence band, 30 m | 126 | −0.180 |
| `dist_to_stream_km` | OpenStreetMap, nearest river/stream/canal vertex | 192 | +0.282 |
| `dist_to_kelani_km` | OpenStreetMap, nearest Kelani Ganga vertex | 194 | +0.172 |
| `hand_m` | MERIT Hydro, height above nearest drainage | 194 | +0.558 |
| `sar_flood_open_land` | Sentinel-1 change detection, three flood events | 107 | +0.202 |

†`elevation_m` correlates with `hand_m` at $r = +0.927$; because the ablation removes the six variables as one block, this collinearity does not affect it.

### *D. Features, Target and Model*

The feature set carries land size, `Distance_from_fort`, amenity counts and nearest-amenity distances from cKDTree indexing, the six hazard variables, five engineered amenity aggregates and 13 land-type indicators (48 raw features), plus a target-encoded division identifier, `Address_target_enc`, giving 49. No coordinates are present. The target is the natural logarithm of price per perch. `Address` is target-encoded out of fold with five shuffled inner folds and smoothing 10; a division absent from training receives the training mean.

Random Forest, Extra Trees, XGBoost and LightGBM were compared at the model-selection stage, tuned by `RandomizedSearchCV` (30 settings, five-fold CV, $R^2$ scoring). The tuned Random Forest (`n_estimators` = 500, `max_depth` = 30, `max_features` = `None`, `min_samples_split` = 6, `min_samples_leaf` = 2) is the specification used for every result in Sections IV-C to IV-F, where Section IV-D also adds or removes one feature. These hyperparameters were chosen once, on the seed-42 training rows, which include divisions later held out in grouped folds; the untuned 300-tree forest scores 3.17 points lower under `GroupKFold` with the original preprocessing, so tuning can account for at most about that much of the unseen-division figure.

### *E. Preprocessing and Validation Protocols*

Price filtering and trimming depend on the target, so they are fitted inside each training fold. For every split, in this order: (1) the 0.5% and 99.5% quantiles of price per perch are computed on the training rows, and training rows outside them are dropped; (2) training rows beyond 1.5 SD of their division's mean are dropped, first on price per perch, then on land size, using training rows only; (3) `Address` is target-encoded on the remaining training rows; (4) the model is fitted; (5) it is scored on the validation rows whose price lies inside the training quantiles of step 1, a validity range fixed by the training rows; scores on every validation row are also reported. No statistic computed from the validation rows is used at any step. The 1.5 SD constant was fixed beforehand, chosen from 1.5, 2.0, 2.5 and 3.0 by cross-validation within the training set.

Three protocols share this pipeline. *Seen divisions*: a random 80/20 split of listings, repeated over seeds 42, 1, 7, 123 and 2024. *Unseen divisions*: five folds grouped on `Address`, so every validation division is absent from training, both as scikit-learn's `GroupKFold` (deterministic; the five seeds vary only the forest) and as 25 independent randomised, size-balanced division-to-fold assignments, which vary the partition itself and give the uncertainty estimate. For comparison, all three are also run under the original global pipeline, in which steps 1–2 were applied once to all 7,034 rows, leaving 6,167 rows in 187 divisions.

The hazard block is assessed by removing all six hazard variables and re-running each of the 25 randomised partitions with the same folds and seed, so that each partition yields one paired difference. A 95% interval for the mean difference is obtained by bootstrap over the 25 partitions. Permutation importance (five repeats, $R^2$ scoring) is computed on the validation rows of each `GroupKFold` fold and of the seed-42 random split. All runs use Python 3.13 and scikit-learn 1.9; scripts and per-fold outputs are available from the authors.

Spatially aware alternatives are compared on the same 25 partitions and pipeline. Each division is given a *neighbour price*: the inverse-distance-weighted mean of the training mean log price of its $K = 5$ nearest training divisions by centroid distance, $K$ being fixed in advance. Used alone, it is a model-free baseline. It is also added to the Random Forest, with and without `Address_target_enc`; for a training row it is computed from the $K$ nearest *other* training divisions, so no row sees its own division's price.

## IV. RESULTS

### *A. The Modelling Dataset*

After the original global filtering, the data span 187 divisions and are severely unbalanced: the median division contributes 6 listings and the largest 518, while 108 of the 187 rest on fewer than 10 and 38 on a single listing. Because every geographic, amenity and hazard column is constant within a division, the effective sample for any location-based inference is 187 to 194 divisions, not thousands of listings. The raw target maximum is exactly Rs 10,000,000 per perch, with 16 rows at that value, indicating a ceiling rather than an observed maximum.

### B. Model Selection (Exploratory)

TABLE II. MODEL SELECTION: RANDOM 80/20 SPLIT, SEED 42, GLOBAL PREPROCESSING

| Stage | Model | $R^2$ (%) | RMSE (log) | MAPE (%) |
|---|---|---|---|---|
| Baseline, original | RF | 85.60 | 0.2918 | 18.45 |
| Tuned, original | RF | 85.80 | 0.2897 | 18.68 |
| Enriched, untuned | RF | 89.09 | 0.2442 | 15.19 |
| Enriched, untuned | ExtraTrees | 88.57 | 0.2499 | 15.55 |
| Enriched, untuned | XGBoost | 88.45 | 0.2513 | 15.74 |
| Enriched, untuned | LightGBM | 89.57 | 0.2387 | 15.62 |
| **Enriched, tuned** | **RF** | **89.43** | **0.2403** | **15.29** |
| Enriched, tuned | XGBoost | 89.68 | 0.2461 | 16.16 |
| Enriched, tuned | LightGBM | 89.16 | 0.2523 | 16.89 |
| Ensemble (mean) | RF+XGB+LGB | 90.23 | 0.2395 | 15.63 |

Exploratory. The tuned Random Forest (bold) is the specification used in Table III.

Table II records the model-selection stage, run once under the original global pipeline on a single seed-42 split; it guides the choice of model and is not the confirmatory evaluation. On the 1,234 held-out rows the tuned Random Forest attains $R^2$ = 89.43% with a median absolute percentage error of 6.5%. Tuned XGBoost reaches 89.68% and an equal-weight ensemble 90.23%; the Random Forest is kept because every later analysis is conducted on it, and no claim is made that it is the best algorithm in general. Under this split the six hazard columns add a mean paired 0.13 percentage points over five seeds, within the noise of their own paired distribution (SD 0.17 pp).

In-fold: price quantiles, per-division trimming and target encoding fitted on training rows only (Section III-E). Global: the same steps applied once to all rows before splitting. $R^2$ on the log target; median APE on the rupee scale. The $\Delta$ row's CI is a bootstrap over the 25 partitions. Neighbour price: inverse-distance mean of the 5 nearest training divisions (Section III-E); the RF rows add it as a 50th feature, and the last also drops `Address_target_enc`.

TABLE III. SEEN- VERSUS UNSEEN-DIVISION ACCURACY. TUNED 500-TREE RANDOM FOREST AND IDENTICAL ENCODER UNLESS STATED

| Protocol | Preprocessing | Replicates | Mean $R^2$ (%) | SD | Range of replicates | Median APE (%) |
|---|---|---|---|---|---|---|
| Random 80/20 (seen divisions) | In-fold | 5 seeds | 78.97 | 1.21 | 77.09–80.16 | 8.99 |
| GroupKFold (unseen) | In-fold | 5 seeds | 31.79 | 3.42 | 27.94–35.76 | 33.86 |
| Randomised grouped partitions (unseen) | In-fold | 25 partitions | **27.90** | 6.03 | 11.22–39.85 | 33.00 |
| Random 80/20 (seen divisions) | Global (original) | 5 seeds | 87.72 | 1.29 | 85.99–89.43 | 7.15 |
| GroupKFold (unseen) | Global (original) | 5 seeds | 43.55 | 2.14 | 40.74–45.70 | 28.40 |
| Randomised grouped partitions (unseen) | Global (original) | 25 partitions | 30.28 | 6.72 | 14.68–42.06 | 31.63 |
| No hazard variables, paired $\Delta$ vs in-fold partitions (pp) | | | −0.37 | 0.97 | 95% CI −0.74 to −0.01 | 17/25 negative |
| Neighbour price only, $K$ = 5 (unseen) | In-fold | 25 partitions | 45.57 | 2.76 | 38.78–50.05 | — |
| RF + neighbour price (unseen) | In-fold | 25 partitions | 29.73 | 5.84 | 14.10–40.89 | — |
| RF + neighbour price, no identity (unseen) | In-fold | 25 partitions | 41.98 | 6.24 | 29.22–53.81 | — |

### C. Seen- Versus Unseen-Division Accuracy

Table III holds the model, features, encoder and pipeline fixed and varies only the validation design. With in-fold preprocessing, mean $R^2$ is 78.97% on seen divisions and 27.90% on unseen divisions across 25 partitions, a gap of 51.07 percentage points; the central 95% of partitions lies between 15.93% and 38.21%, and single folds range from −61.17% to 70.46%, so accuracy for any one unseen division is highly uncertain. Scoring every validation row, including prices outside the training range, gives 75.67% and 27.92%. Under the original global pipeline the gap is larger (87.72% against 30.28%, 57.44 points), so it is not an artefact of preprocessing order. In-fold preprocessing lowers both figures, chiefly because validation rows are no longer trimmed on their own division statistics: trimming them as before restores 87.09% on seen divisions. The deterministic `GroupKFold` partition scores above 76% of the randomised partitions, so a single fixed partition is not a reliable estimate on its own.

### D. Spatially Aware Comparison

The neighbour price alone reaches a mean $R^2$ of 45.57% on unseen divisions, 17.67 points above the Random Forest and higher in 25 of 25 partitions (bootstrap 95% interval of the gain 15.66 to 19.79); with $K$ = 3 and $K$ = 10 it gives 35.78% and 50.43%. Added to the Random Forest alongside

`Address_target_enc`, it adds only 1.83 points, because in training the identity column already explains each division's price and the forest has no reason to split on the neighbour price. Without the identity column the same forest reaches 41.98%, 14.09 points above the paper's specification, though still 3.58 points below the neighbour price alone. For divisions without listings, the prices of nearby divisions are the most useful location information, and the identity feature prevents the forest from learning to use them.

### *E. Ablation of the Hazard Block*

Removing all six hazard variables changes unseen-division $R^2$ by a mean −0.37 percentage points across the 25 paired partitions (SD 0.97; bootstrap 95% interval −0.74 to −0.01), with the difference negative in 17 of 25; on the fixed `GroupKFold` partition it is −0.11 points. The bootstrap interval treats the 25 partitions as independent although they reuse the same divisions, so it is optimistic. The effect fails the study's pre-specified decision rule (mean paired difference beyond two SD of its own distribution, with a consistent sign), so the hazard block adds at most a small and non-robust amount, well under one percentage point. An earlier analysis on the fixed partition with an untuned forest had reported a 1.00-point loss that met the rule; it does not survive the controlled design and is withdrawn. Individual variables are not tested, since the block as a whole is already near the noise floor.

### *F. Feature Importance*

Gini importance of the tuned Random Forest of Table II ranks `Address_target_enc` first at 75.40% of the total across the 49 features, followed by `Distance_from_fort` (5.23%) and land size (5.11%); the six hazard variables together carry 1.83%. Permutation importance under the in-fold pipeline separates the two settings. On the random split, `Address_target_enc` ranks 1st with a mean $R^2$ drop of 1.0541; on unseen `GroupKFold` folds it ranks 34th of 49 with a drop of 0.0000, and `min_dist_semigovtschools` leads at 0.1409. The mechanism is structural: a validation division is absent from the training half, so the encoder holds no statistic for it and assigns the training mean, leaving the column constant across the validation fold. The feature that dominates the model's apparent explanatory power on seen divisions carries no information for an unseen one.

### *G. Exploratory Extension to Ratnapura District*

As an exploratory check, not re-run under the revised protocol, the original global pipeline was applied to a merged Colombo–Ratnapura file with 1,000 CTGAN-generated synthetic Ratnapura rows, all assigned to training, and 65 features including five landslide variables. On real listings only, the random split gave $R^2$ = 88.90% and `GroupKFold` 49.84%. The direction of the gap reproduces, but because the training augmentation is synthetic this is not a validated out-of-district result.

### *H. Direction of the Surface-Water Effect*

SHAP values for `water_occurrence_pct`, computed with TreeExplainer on held-out rows of the model-selection split and regressed on the underlying occurrence (Fig. 1), give an overall slope of −0.000144 and negative slopes in all four quartiles of `Distance_from_fort` (−0.000100, −0.000222, −0.000122, −0.000111; $n$ = 334, 271, 311, 284). This describes the fitted model only: given the ablation result it is not evidence that the variable materially improves prediction.

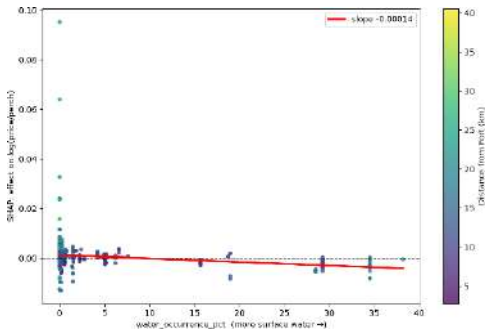



Fig. 1. SHAP values for `water_occurrence_pct` against the underlying surface-water occurrence for held-out rows of the model-selection split (tuned Random Forest, Table II), coloured by distance from Fort.

## V. Discussion

The central result is the distance between two accuracy figures obtained from one model and one pipeline. On seen divisions the model behaves substantially as a lookup on division identity: `Address_target_enc` carries three-quarters of Gini importance because listings from the same division appear on both sides of the split. On unseen divisions that encoding carries no division-specific information. Feature importance from a randomly split model is therefore misleading for any model with a location-identity feature; what matters is whether a feature identifies a place or describes it. Both figures are valid answers to different questions: the first applies to divisions with existing listings, the second to divisions without them. Section IV-D shows the cost of the identity feature directly: replacing it with a neighbour price, which describes location rather than naming it, raises unseen-division accuracy by 14.09 points.

The hazard evidence is weak. Neither the random split nor the repeated grouped design shows a hazard contribution that meets the decision rule, the paired difference on unseen divisions is a fraction of a percentage point, and the SHAP direction for surface water, though negative, is small. This is consistent with three explanations the design cannot separate: the estimates are descriptive associations from listed rather than transacted prices; low-lying land in Colombo is also older, denser and closer to the historic core, and hazard awareness in the Kelani lower basin is limited [9]; and Sentinel-1 under-detects inundation beneath dense built-up cover, so `sar_flood_open_land` is blind where flood damage concentrates.

The number of independent units is the main limitation. Every location feature is constant within a division, so location effects are estimated from at most 194 divisions, of which many contribute only a few listings; this is why unseen-division accuracy varies so widely between folds and partitions, and why the hazard ablation has little power. The results therefore generalise only to Colombo GN divisions similar to those with listings; they say nothing about the right-censored high-value core above Rs 10,000,000 per perch, nor about other districts.

Within-division variation in exposure is not represented, and GN divisions are administrative rather than hydrological units. Coherence-based urban flood detection [18], plot-level geocoding that would permit point-based spatial models such as kriging, and real Ratnapura listings are the natural extensions.

## VI. Conclusion

This paper evaluated a land price model for Colombo District at Grama Niladhari division resolution with one model, one feature set and one fold-internal preprocessing pipeline under three validation designs. The tuned Random Forest attains mean $R^2$ of 78.97% on seen divisions and 27.90% on unseen divisions across 25 randomised grouped partitions, a gap of 51.07 percentage points that persists under the original preprocessing. The mechanism is identified: the target-encoded division identifier dominates importance on seen divisions and carries no division-specific information on unseen ones, while a neighbour-price description of location lifts unseen-division accuracy to 41.98% in the forest and 45.57% on its own. On the hazard question the evidence is weak: removing all six satellite-derived hazard and terrain variables lowers unseen-division accuracy by 0.37 percentage points on average, an effect that does not meet the study's decision rule. The contribution is the transfer of spatially-blocked validation from ecological modelling into property valuation, with estimates that rest on fewer than 200 effective units and listed rather than transacted prices.

## References


B. V. S. M. Athurupana and R. M. M. M. Pradeep, "Methods using land valuation for Sri Lanka," in *Proc. Fac. Comput. Student Symp.*, Gen. Sir John Kotelawala Defence Univ., Sri Lanka, 2021, p. 16.

R. T. Mora-Garcia, M. F. Cespedes-Lopez, and V. R. Perez-Sanchez, "Housing price prediction using machine learning algorithms in COVID-19 times," *Land*, vol. 11, no. 11, art. 2100, Nov. 2022.

L. H. T. Choy and W. K. O. Ho, "The use of machine learning in real estate research," *Land*, vol. 12, no. 4, art. 740, Mar. 2023.

M. Yazdani, "Machine learning, deep learning, and hedonic methods for real estate price prediction," arXiv:2110.07151, Oct. 2021.

N. S. Ja'afar, J. Mohamad, and S. Ismail, "Machine learning for property price prediction and price valuation: A systematic literature review," *Plan. Malaysia*, vol. 19, no. 17, Oct. 2021.

J. D. Gourevitch *et al.*, "Unpriced climate risk and the potential consequences of overvaluation in US housing markets," *Nature Clim. Change*, vol. 13, no. 3, pp. 250–257, Mar. 2023.

R. D. Rajasinghe, "Generative AI and Its Implications for the Authenticity and Legal Use of Video Evidence," 2025 1st International Conference on Data Science and Intelligent Network Computing (ICDSINC), Raipur, India, 2025, pp. 17-22, doi: 10.1109/ICDSINC66221.2025.11448275.

J. Clayton, S. Devaney, and J. van de Wetering, "Climate risk and commercial property values: A review and analysis of the literature," SSRN preprint, 2025, doi: 10.2139/ssrn.5227387.

M. T. O. V. Peiris, N. Randeniya, and N. Bopitiyegedara, "How do anthropogenic factors define flood risk perception of vulnerable communities? Evidence from Kelani River Lower Basin, Colombo, Sri Lanka," *Int. J. Disaster Risk Reduct.*, vol. 120, art. 105352, Apr. 2025.

T. A. G. S. Pushpakumara, K. G. P. K. Weerakoon, and W. H. T. Gunawardhana, "Spatial variation of land values in the Colombo District, Sri Lanka using geographic information system," in *Proc. Int. Conf. Real Estate Manage. Valuation*, vol. 7, Dec. 2023.

O. Rathnayaka and P. Rathnayaka, "Graph-based multitask learning framework for property price prediction and data imputation in the Sri Lankan real estate market," in *Proc. 7th Int. Conf. Image, Video Process., Artif. Intell. (IVPAI)*, Aug. 2025.

S. Basu and T. G. Thibodeau, "Analysis of spatial autocorrelation in house prices," *J. Real Estate Finance Econ.*, vol. 17, no. 1, pp. 61–85, Jul. 1998.

D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, Aug. 2017.

P. Ploton *et al.*, "Spatial validation reveals poor predictive performance of large-scale ecological mapping models," *Nature Commun.*, vol. 11, no. 1, art. 4540, Sep. 2020.

D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *ACM SIGKDD Explor. Newsl.*, vol. 3, no. 1, pp. 27–32, Jul. 2001.

M. Tazmul Islam and Q. Meng, "An exploratory study of Sentinel-1 SAR for rapid urban flood mapping on Google Earth Engine," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 113, art. 103002, Sep. 2022.

D. C. Mason, S. L. Dance, and H. L. Cloke, "Floodwater detection in urban areas using Sentinel-1 and WorldDEM data," *J. Appl. Remote Sens.*, vol. 15, no. 3, Feb. 2021.

J. Zhao, Y. Li, P. Matgen, R. Pelich, R. Hostache, W. Wagner, and M. Chini, "Urban-aware U-Net for large-scale urban flood mapping using multitemporal Sentinel-1 intensity and interferometric coherence," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–21, 2022.

J. Zhao, Z. Xiong, and X. X. Zhu, "UrbanSARFloods: Sentinel-1 SLC-based benchmark dataset for urban and open-area flood mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 419–429.

S. Openshaw, *The Modifiable Areal Unit Problem*. Norwich, U.K.: Geo Books, 1984.

H. Zhang, Z. Qi, X. Li, Y. Chen, X. Wang, and Y. He, "An urban flooding index for unsupervised inundated urban area detection using Sentinel-1 polarimetric SAR images," *Remote Sens.*, vol. 13, no. 22, art. 4511, Nov. 2021.

# Short-Term Residential Load Forecasting: A Machine Learning Evaluation Against a Persistence Baseline Using Fused Smart-Meter, Weather, and Socio-Economic Data in Sri Lanka

Y. D. K. Soysa
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
heethsoysa005@gmail.com

R. W. K. T. Rajapaksha
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
kavishka.r@nsbm.ac.lk

**Abstract—Short-term residential load forecasting increasingly relies on smart-meter data, but studies rarely establish whether a proposed model beats the trivial alternative of repeating the last reading, or whether household socio-economic attributes add anything once consumption history is known. We address both questions on a fused dataset from the Residential Electricity Consumption Dataset for Sri Lanka (RECON-SL): 15-minute smart-meter readings for 131 households in seven Lanka Electricity Company branches, NASA POWER weather and wave-1 survey attributes (49,387 readings, 42 features). A regularised linear model and two gradient-boosted tree implementations share a residual target and a chronological split, are selected on validation and are scored once on a held-out quarter. Both boosted models beat naive persistence by a modest margin (CatBoost 5.56%, 95% CI 3.89–7.39%; LightGBM 5.84% lower mean absolute error), a gain that holds at two earlier forecast origins and, at 3.57%, for households never seen in training. Methodologically, the linear model is 6.86% worse than persistence yet has the highest coefficient of determination at every origin: on strongly autocorrelated 15-minute load, $R^2$ ranks models incorrectly. Substantively, rows with survey data gain 8.17% and rows without 1.99%, but blanking the attributes on the same rows removes only 2.61 percentage points, re-training without them 1.19, and for unseen households their contribution is indistinguishable from zero: the attributes help forecast households already observed rather than new ones.**

**Keywords—gradient boosting, persistence baseline, short-term load forecasting, smart meters, socio-economic data**

## I. INTRODUCTION

Sri Lanka's distribution utilities have deployed smart metering at scale over the last decade, and the Lanka Electricity Company (LECO) now records residential consumption at 15-minute resolution for large numbers of households. At the same time, growing air-conditioning penetration, rooftop solar net metering and time-of-use tariffs have made household demand more volatile and more weather-sensitive than the monthly billing data on which Sri Lankan residential energy research has historically depended [1], [2]. Short-horizon, household-level forecasts are the input that peak management and net-metering planning require, and they are the input utilities have least experience producing. Two gaps in how this problem is usually studied motivate this work, and neither is about model architecture.

The first is evaluative. A large share of the load-forecasting literature reports accuracy without establishing that the proposed model beats the cheapest possible alternative: predicting that the next 15-minute reading equals the last one [21]. At this horizon the persistence baseline is strong, because residential consumption is dominated by its own autocorrelation; in our data the correlation between a reading and its predecessor is 0.964. A model that does not clear this bar has not shown that its complexity buys anything, and a metric that cannot detect the difference is not a useful metric.

The second is substantive. Smart-meter data describe what a household consumed; survey data describe who the household is: its income band, floor area, occupancy and appliance stock. The energy-poverty literature uses such attributes to characterise deprivation rather than to forecast consumption [3], [4]. Whether they help a short-term forecasting model once lagged consumption is available is rarely tested, in part because few datasets link the two for the same households. The 2025 release of RECON-SL [5] (4,063 households, three survey waves and over 50 million smart-meter readings from LECO's service area) makes the question answerable for a developing-economy setting. This paper makes four contributions:

1. A reproducible pipeline, released with the fused dataset, that joins 15-minute smart-meter readings, NASA POWER weather and wave-1 survey attributes and benchmarks three model families against naive persistence under a chronological split, with selection on validation, one scoring pass on test, household-clustered confidence intervals, rolling origins and a household-held-out evaluation.
2. A methodological result: $R^2$ ranks models incorrectly here. The linear control, the worst of the three families by mean absolute error (MAE), has the highest $R^2$ at every origin, the do-nothing baseline scores 0.9395, and we trace the inversion to the training loss.
3. A substantive result, kept separate from the first: what socio-economic attributes contribute, measured by blanking them on fixed rows, by re-training without them and on held-out households. The uncontrolled subset comparison overstates the effect more than twofold, and the attributes do not measurably help forecast unseen households.
4. A deployed service, with the operational consequences of these modelling choices.

## II. RELATED WORK

Deep learning and tree ensembles. Kong et al.'s LSTM established recurrent architectures as a strong default for individual household load [6], and Dong et al.'s survey confirms deep learning, from recurrent to attention-based models, as the dominant approach of the last decade while noting its larger data requirements [7]. Gradient boosting has nonetheless proved highly competitive on tabular formulations: Zhao et al. report an optimised XGBoost [26] model for regional demand [8], Rubattu et al. a probabilistic LightGBM with temporal hierarchies [9], and Miller et al.'s post-mortem of the ASHRAE Great Energy Predictor III competition found that gradient boosting machines with careful pre-processing worked best across a large heterogeneous field [10]; Mansoor et al. evaluate tree ensembles on advanced metering infra-

structure data [11]. Whether the choice among boosted implementations matters is not settled, and we test it directly. Weather as a driver. Maia-Silva et al. show that temperature-only models understate cooling demand by 10–15% relative to models that include humidity [12], and Al-Musaylh et al. combine ground climate with ECMWF reanalysis predictors [13]. Most such work uses ground stations; open satellite and reanalysis products such as NASA POWER [14] remain comparatively unexploited, despite being the practical option where ground networks are sparse.

Sri Lankan work. Sri Lanka-specific research has largely operated on aggregate data. Abeywickrama and Asanka integrate rainfall, humidity and temperature into district-level monthly demand models, finding vector autoregression to outperform LSTM [15]. Fernando and Hewapathirana forecast sector-aggregate demand using socio-economic and weather variables with LSTM, SARIMA and a hybrid [16]. Household-level studies have been cross-sectional and survey-based [1], [2]. We therefore phrase our contribution as a measurement rather than a claim of precedence: not whether anyone has fused these sources before, but what the fusion is worth against a baseline that a forecasting model must actually beat.

## III. DATA

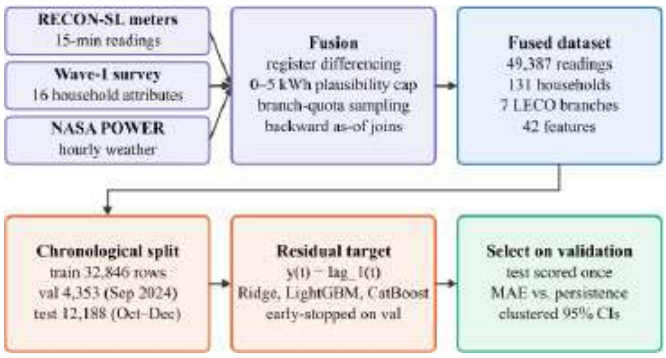



*Fig. 1. Data-fusion and evaluation pipeline. Three sources are fused into one model-ready dataset, split chronologically and used to train three model families on a residual target; the reported model is selected on validation and scored once on the held-out test period.*

### *A. Source and construction*

The source is RECON-SL [5], released publicly in 2025, which links LECO smart-meter records for 4,063 households to a three-wave household survey covering demographics, dwelling characteristics, appliance ownership and perceptions of energy security. We use the 15-minute series. Meters record a cumulative import register, so the series is first differenced within each household, which exposes negative differences where a register was reset or replaced and implausibly large positive ones from counter rollover (one household shows four consecutive 15-minute readings above 20,000 kWh). We drop negative differences and apply a physical plausibility cap of 5 kWh per 15-minute window immediately after differencing, before any lag or rolling feature is built; filtering only at the sampling stage would leave the outliers embedded in every rolling statistic computed over them.

### *B. Branch-preserving sampling*

Household coverage in RECON-SL is heavily skewed across LECO branches, and proportional sampling would eliminate the smaller branches entirely. We instead draw a capped quota of 6–30 households per branch (Galle 30; Kalutara, Kelaniya, Moratuwa and Nugegoda 20 each; Kotte 15; Negombo 6), and every household contributes the same 377 readings, so that none dominates the training loss. The 131 households and 49,387 readings (2 October 2023 to 23 December 2024) are thus a design choice within a budget of about 50,000 model-ready rows, not a property of RECON-SL, and the sample is not representative of the branch population: a deliberate trade of external validity for the ability to report disaggregated results at all. Four integrity conditions are asserted at every run (no duplicate household– timestamp pairs, no negative consumption, no reading above the cap, a defined baseline for every scored row), and all hold. Consumption is zero-inflated (17.3% of readings are exactly zero and a further 10.5% fall in (0, 0.01] kWh), which is why MAPE is floored, and strongly autocorrelated, which is why persistence is a demanding baseline rather than a formality.

### *C. Uneven survey coverage*

Wave-1 survey data is missing for 36 of the 131 households (27.5% of rows), and not at random: Negombo has no surveyed households and Kotte three of fifteen, while Kalutara and Nugegoda are fully covered. Dropping these rows would remove a branch, and imputing the attributes would manufacture characteristics that were never measured, so we retain the rows and expose the gap as an explicit has_survey_data indicator. Because 24.6% of training rows carry no survey record, the trained model can also operate on households whose attributes are unknown, which allows the attributes to be blanked at inference (Section V-E) and makes the deployed service usable without a household profile (Section VI).

### *D. Weather*

Hourly temperature, relative humidity, all-sky downward shortwave irradiance and corrected precipitation are retrieved from NASA POWER [14] for each branch centroid across the full study period. The join to consumption is a backward-only as-of join: each reading is matched to the most recent weather observation at or before its timestamp, never to a later one, so no future information enters the feature set.

### *E. Features*

Forty-two features are supplied to every model: four weather variables, six calendar features including cyclical hour encodings, six lag and rolling-history features, sixteen socio-economic attributes (including has_survey_data), a branch identifier and nine engineered features. Of the history features, lag_1 and the 1-hour and 24-hour rolling means come from the continuous 15-minute series; lag_2, lag_3 and the rolling standard deviation were computed over each household's preceding sampled readings, so they summarise its recent level (lag_2 reaches a median of 16 hours back) rather than the last 30–45 minutes. The engineered set comprises trends (lag_1 − lag_2, lag_2 − lag_3), a three-lag mean, the deviation of the last reading from the household's own 24-hour mean, the ratio of short- to long-window means, a recent coefficient of variation, a zero-consumption flag, and temperature and humidity each multiplied by air-conditioner count. Every feature uses strictly earlier observations, so none introduces look-ahead leakage; nine of the forty-two are categorical.

### *F. Constraint on model class*

The fused dataset is a row sample rather than a continuous series: consecutive rows of a household are a median of 390

minutes apart, and only 3.1% follow the previous row by exactly 15 minutes. The lag_1 column is nevertheless the true prior 15-minute reading, computed before sampling; on all 1,507 contiguous pairs it equals the preceding row's consumption. The tabular models and the persistence baseline are therefore valid, but a sequence model, which needs contiguous windows, is not. This is a property of the data as constructed, not a preference, and it is why recurrent and attention-based architectures [6], [7] appear in Section VIII as future work rather than as an experiment that was run and rejected.

## IV. METHOD

### *A. Residual formulation*

Predicting the consumption level asks a model to re-derive information the persistence baseline already holds. Every model here instead predicts the correction to the last observed reading:

target = period_kwh(t) − lag_1(t) (1) prediction = max(0, lag_1(t) + f(x_t)) (2) A model that learns nothing therefore degrades to persistence rather than to noise, and capacity is concentrated on the genuinely uncertain component. Because all models share the formulation, it cannot account for differences between them.

### *B. Model families and implementation*

Three families are compared on the same features and target:

- Ridge regression [19] as a classical control, included so that the non-linearity of the problem is measured rather than asserted; the penalty is chosen on validation from {0.01, 0.1, …, 1000} ($\alpha = 1000$).
- LightGBM [18], leaf-wise gradient boosting with native categorical splitting, trained on an L1 objective (learning rate 0.03, 63 leaves, at least 20 rows per leaf, row and column subsampling 0.8, L2 penalty 1; 497 trees).
- CatBoost [17], which encodes categorical variables by ordered target statistics, relevant because nine features are categorical household attributes for which one-hot encoding is wasteful and label encoding would impose a false ordering (MAE loss, learning rate 0.03, depth 6, L2 penalty 3; 1,646 trees).

Both boosted models stop when validation MAE has not improved for 100 rounds (at most 3,000 trees). Each family receives the encoding it requires on the same features: pandas categorical dtype for LightGBM, which routes missing values natively; string levels with a MISSING category for CatBoost; and median imputation, standardisation and one-hot encoding for Ridge, fitted on training data only. A supplementary feed-forward network (MLP; ReLU layers of 128 and 64 units, Adam, epochs chosen on validation) uses the Ridge encoding as the closest neural comparator the data allow. We use Python 3.12, scikit-learn 1.9, LightGBM 4.7 and CatBoost 1.2 with seed 42; code, data and trained models are available at https://github.com/ydksoysa/leco-load-forecasting.

### *C. Evaluation protocol*

The split is chronological [22], since a random split would leak future consumption through each household's lag and rolling features. Training covers 2 October 2023–31 August 2024 (32,846 readings, 66.5%, 131 households), validation September 2024 (4,353, 8.8%, 94 households) and test 1 October–23 December 2024 (12,188, 24.7%, 112 households). Every test household also appears in training, so the headline figures measure forecasting the future of known households; Section V-D adds unseen ones. Ridge's penalty, the tree counts and the reported model are all chosen on validation MAE, and the test slice is scored once, at the end. No information passes from validation or test into training: features look strictly backward, the split is by date, and every fitted transformation uses training rows only.

MAE is the primary metric; RMSE, $R^2$ and MAPE are reported alongside, with MAPE excluding readings at or below 0.01 kWh [20]. Uncertainty comes from a household-clustered bootstrap (2,000 resamples of households) [30], and paired comparisons use a Diebold–Mariano-type test [28] on the absolute-error differential with standard errors clustered by household.

## V. RESULTS

### *A. Forecast accuracy and model selection*

Table I reports validation and test performance. CatBoost was selected on validation MAE (0.022946 against LightGBM's 0.022956) before the test period was examined, and is the reported model.

TABLE I. FORECAST ACCURACY ON THE VALIDATION AND HELD-OUT TEST SLICES (MAE AND RMSE IN KWH; SELECTION ON VALIDATION MAE BEFORE THE TEST PERIOD WAS EXAMINED)

| Model | Val. MAE | Test MAE | Test RMSE | Test $R^2$ | Test MAPE | vs. naive |
|---|---|---|---|---|---|---|
| Naive persistence | 0.024072 | 0.020262 | 0.0610 | 0.9395 | 36.38% | — |
| Ridge | 0.025648 | 0.021652 | 0.0577 | 0.9458 | 38.83% | −6.86% |
| MLP (suppl.) | 0.025644 | 0.021756 | 0.0581 | 0.9452 | 40.24% | −7.37% |
| LightGBM | 0.022956 | 0.019079 | 0.0579 | 0.9456 | 31.47% | +5.84% |
| CatBoost (sel.) | 0.022946 | 0.019134 | 0.0580 | 0.9454 | 31.46% | +5.56% |

Both boosted models beat persistence on test, CatBoost by 5.56% (95% CI 3.89–7.39%) and LightGBM by 5.84% (4.21–7.64%), and on validation (4.68% and 4.63%); the DM-type test rejects equal accuracy for both ($p < 0.001$). The means of the last four and last 96 readings are 19.5% and 158.6% worse than persistence, the strongest of these naive baselines. The margin is modest, as expected when consumption is dominated by its own autocorrelation.

Ridge is 6.86% worse than doing nothing on test (CI of the gain −9.85 to −4.15%) and 6.55% worse on validation, and the MLP is 7.37% worse. This negative result turns the assumption that classical models are unsuitable for household-

level panel data from a premise into a measurement, and shows that the residual formulation is not sufficient on its own: a model handed the baseline as its starting point can still be driven below it.

Selection and reporting. The boosted models differ by 0.000010 kWh in validation MAE (0.04%) and 0.000055 kWh in test MAE (about 5% of either's improvement over persistence), and neither difference is meaningful (DM-type $p = 0.86$ on validation and 0.24 on test; test-difference CI −0.73 to 0.18 points). CatBoost remains the reported model because the selection rule was fixed before testing: switching to LightGBM after seeing its marginally lower test MAE would be the test-set selection the protocol excludes, and no conclusion depends on the choice. Once the target is a residual and history features are supplied, any competent boosted implementation reaches the same accuracy, while a linear model does not reach it at all.

*B. $R^2$ ranks these models incorrectly*

Ridge attains $R^2 = 0.9458$, the highest in Table I, while producing larger absolute errors than doing nothing; persistence itself scores 0.9395, and the MLP repeats the pattern ($R^2$ 0.9452). A reader shown only $R^2$ would rate every row a strong model, including the two that are worse than no model at all.

The inversion has a precise cause: $R^2$ and RMSE reward forecasting the conditional mean, MAE the conditional median [31], and the two differ sharply here (test mean 0.079 kWh, median 0.030 kWh, skewness 10.8). Ridge and the MLP minimise squared error, trimming the rare large errors (both beat persistence on RMSE) at the cost of many small ones: Ridge over-forecasts 61.8% of readings and lands within ±0.005 kWh of only 38.2%, against 54.4% for persistence. The same LightGBM trained on squared error instead of L1 stops after 9 trees and beats persistence by only 0.96%. The inversion recurs at both earlier origins (Section V-D), so $R^2$ should be reported for completeness [23] but never used to rank models on this problem, and a result quoting $R^2$ without a baseline comparison cannot be assessed from the number alone

.

*C.* Accuracy in interpretable terms and practical significance Regression has no accuracy in the classification sense, and computing one inverts the ranking: only 6.98% of CatBoost's forecasts match the actual reading exactly, against 20.27% of the baseline's, because whenever a household is idle the previous reading is zero and repeating it is exactly right. Table II reports the more useful formulation, the share of forecasts within a stated tolerance: the model is marginally worse than persistence at ±0.005 kWh and better at every wider tolerance, by 0.6 to 1.2 percentage points.

Practical significance. In absolute terms the gain is 1.13 Wh per 15-minute reading, 1.4% of the mean reading (0.079 kWh), and it is uneven: readings whose previous value lay between 0.15 and 0.4 kWh (0.6–1.6 kW of average demand) are 7.8% of the test set but carry 54% of the total error reduction (a 12.2% gain), whereas idle readings below 0.017 kWh gain 0.8%. The gain is largest in the 18:30–22:30 peak-tariff window (6.3%), and forecasts missing by more than 0.1 kWh fall from 5.14% to 4.50%. The model earns its keep in the high-load readings that peak management cares about; on average the gain is modest, which is why the service shows the baseline beside every forecast.

TABLE II. SHARE OF TEST FORECASTS WITHIN A STATED ABSOLUTE ERROR

| **Absolute error within** | **CatBoost** | **Naive** |
|---|---|---|
| ± 0.005 kWh | 53.95% | 54.36% |
| ± 0.010 kWh | 67.76% | 66.57% |
| ± 0.020 kWh | 81.53% | 80.73% |
| ± 0.050 kWh | 91.12% | 90.53% |
| ± 0.100 kWh | 95.50% | 94.86% |

*D. Robustness across periods and households*

Rolling origins. Fig. 2 repeats the protocol at two earlier origins [29]: R1 trains to April 2024, validates on May and tests on June–July (9,451 readings); R2 trains to June, validates on July and tests on August–September (6,645), covering south-west monsoon months absent from the reported quarter. The model selected on validation beats persistence at both (LightGBM 3.70%, CI 1.84–5.38%; CatBoost 3.46%, CI 1.85–5.33%), and Ridge loses at both (−11.51% and −6.88%) while again recording the highest $R^2$.

Unseen households. To measure generalisation beyond known households, all three families were re-trained in five folds that hold out whole households (26–27 per fold, stratified by branch) on the same dates, so each test reading is forecast by a model that never saw its household. CatBoost still beats persistence by 3.57% (CI 1.92–5.34%; $p < 0.001$) and LightGBM by 3.63%, while Ridge falls to −9.19%. About two-thirds of the gain thus transfers to new households; the other 2.00 points (CI 1.16–2.98) depend on having observed the household, and branch gains fall accordingly (Moratuwa from 12.36% to 8.16%, Galle from 2.58% to −0.32%).

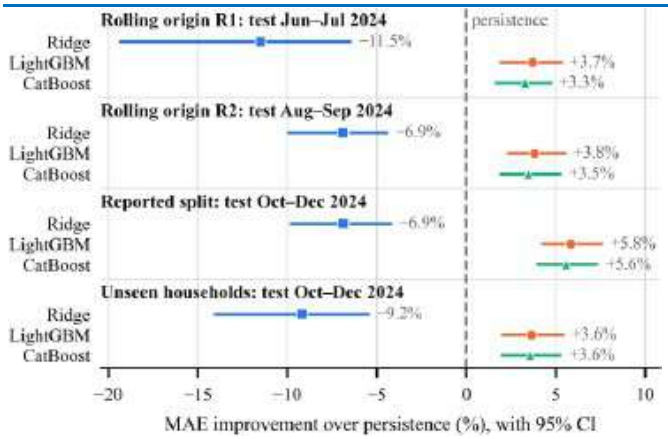



Fig. 2. MAE improvement over persistence with household-clustered 95% bootstrap CIs, by model and evaluation setting: two earlier rolling origins (R1, R2), the reported split, and households held out of training. Points right of the dashed line beat persistence.

*E. The contribution of socio-economic features*

Uneven survey coverage invites a natural comparison: rows with survey data improve on persistence by 8.17% and rows without by 1.99% (Table III, rows a and d). That contrast is confounded, because survey availability correlates with branch and the unsurveyed households have higher baseline error (25.84 against 17.50 Wh). Table III therefore scores fixed rows under controlled treatments of the sixteen socio-economic features: blanked sets them to unknown at inference (numeric to missing, categorical to MISSING, has_survey_data to 0), a state the model learned from its unsurveyed training rows; dropped removes them, with the two air-conditioner interactions derived from them, and re-trains the model. Each gain is 100 × (MAE_naive − MAE_model) / MAE_naive over the rows in scope.

On the 8,153 surveyed rows, blanking lowers the gain from 8.17% to 5.56%, a contribution of 2.61 points (CI 1.76– 3.56), not the 6.18 points the subset contrast suggests, which overstates the effect more than twofold. Blanking measures how much the trained model relies on the attributes; re-training measures what they add, a smaller 1.19 points (CI0.65–1.84), because a model fitted without them recovers part of the information from consumption history. Over all test rows the figures are 1.56 and 1.17 points.

Because the attributes are constant within a household, a tree can also use them to identify the household. Held-out households separate this identity effect from transferable signal: for unseen households the attributes change CatBoost's gain by 0.68 points (Table III, rows h and i; CI −0.35 to 1.76; p = 0.21) and LightGBM's by −0.23 points. In this sample they help forecast households already observed but add no measurable accuracy for new ones, so 2.61 points is an upper bound on their transferable contribution.

TABLE III. MAE GAIN OF CATBOOST OVER PERSISTENCE BY ROWS SCORED AND TREATMENT OF THE SURVEY ATTRIBUTES (TEST PERIOD; MAE IN WH; HOUSEHOLD-CLUSTERED 95% CI)

| Rows scored | n | Attributes | Naive | Model | Gain (%) | 95% CI |
|---|---|---|---|---|---|---|
| **Households seen in training (reported split)** | | | | | | |
| (a) Surveyed | 8,153 | present | 17.50 | 16.07 | 8.17 | 6.38, 10.13 |
| (b) Surveyed | 8,153 | blanked | 17.50 | 16.53 | 5.56 | 4.32, 6.85 |
| (c) Surveyed | 8,153 | dropped | 17.50 | 16.28 | 6.98 | 5.41, 8.64 |
| (d) Unsurveyed | 4,035 | none | 25.84 | 25.33 | 1.99 | −0.50, 4.50 |
| (e) All | 12,188 | present | 20.26 | 19.13 | 5.56 | 3.89, 7.39 |
| (f) All | 12,188 | blanked | 20.26 | 19.45 | 4.00 | 2.71, 5.45 |
| (g) All | 12,188 | dropped | 20.26 | 19.37 | 4.39 | 2.66, 6.13 |
| **Households unseen in training (household-grouped folds)** | | | | | | |
| (h) All | 12,188 | present | 20.26 | 19.54 | 3.57 | 1.92, 5.34 |
| (i) All | 12,188 | dropped | 20.26 | 19.68 | 2.89 | 0.81, 4.94 |

*F. Feature-group importance and disaggregation*

Table IV reports importance by feature group. The families disagree substantially (weather 1.09–17.16%, socio-economic 9.86–19.59%); the shares are comparable in rank within a column rather than in magnitude across columns [24], and they record reliance rather than contribution, which is why Section V-E uses controlled comparisons.

TABLE IV. SHARE OF TOTAL MODEL IMPORTANCE BY FEATURE GROUP (%)

| Feature group | Ridge | LightGBM | CatBoost |
|---|---|---|---|
| Lag / rolling history | 41.32 | 22.71 | 34.97 |
| Engineered | 34.48 | 24.92 | 33.30 |
| Socio-economic | 17.41 | 19.59 | 9.86 |
| Weather | 1.09 | 17.16 | 12.52 |
| Calendar | 2.17 | 13.83 | 7.51 |
| Branch / location | 3.53 | 1.80 | 1.84 |

12.36%, Nugegoda 8.01%, Kelaniya 5.98%, Negombo 4.65%, Kotte 4.11% and Galle 2.58%. Kalutara is the only branch where the model loses (−17.15%), on 201 test readings with the lowest baseline error of any branch (0.0044 kWh), too few to support a conclusion; a deployment would need per-branch validation rather than one aggregate figure.

*G. Companion classification*

A Low/Medium/High classifier over training-set consumption terciles (cut points 0.0170 and 0.0540 kWh) reproduces the pattern: LightGBM reaches 0.8425 accuracy and CatBoost 0.8416, while a logistic-regression control trails at 0.7868, with errors concentrated at adjacent class boundaries. Where a stakeholder asks for an accuracy figure, 84.2% is the number that exists; the regression task has none.

## VI. DEPLOYMENT

We implemented the selected model as a service, a REST API and single-page interface that turns a household's last four meter readings into a 15-minute-ahead forecast; each design problem that arose followed from a modelling decision. Reconstruction: the model predicts a correction (1), so serving must add it to the last reading and clip at zero (2), and the reconstruction step must travel with the model artefact [25]. Profiles: the service runs with or without a stored survey profile, with has_survey_data telling the model which applies; the ablation prices a missing profile at 1.56 points for known households and at no measurable cost for new ones (Section V-E), so a forecast is never blocked over a survey. Weather: NASA POWER is queried live, with a fallback to a training-period climatology (the median for that branch, month and hour), and every response names its weather source.

Error bands: absolute error grows strongly with consumption, so bands are bucketed by forecast magnitude; the 68th-percentile absolute error is 0.0017 kWh below 0.017 kWh, 0.0104 kWh at 0.017–0.054, 0.0191 kWh at 0.054– 0.15, 0.0574 kWh at 0.15–0.4 and 0.1361 kWh above 0.4 kWh, a factor of eighty between the extremes. Finally, the persistence baseline is displayed beside every forecast, because a 5.56% improvement is small enough that the model's number alone would overstate what the system provides. The demonstration forecasts one step ahead and needs four consecutive readings, available from a meter feed but a barrier to manual entry.

## VII. DISCUSSION

Four findings generalise beyond this dataset. Baselines are not a formality: two of the four learners, the linear control and the MLP, underperform repeating the last reading while looking strong on $R^2$, and reporting a baseline separates a model that has been evaluated from one that has merely been measured.

Metric choice can invert conclusions: the inversion of Section V-B recurs at every origin and follows from the loss, so wherever a target is autocorrelated, skewed and heterogeneous across units, $R^2$ will favour trivial predictors and squared-error learners that lose on absolute error.

Subset comparisons on non-randomly missing data measure composition as well as effect: rows with and without survey features differ by 8.17% against 1.99%, blanking on fixed rows gives 2.61 points, re-training 1.19, and unseen households show no measurable effect. Where data is missing structurally, as survey non-response usually is, the available-versus-missing comparison is an upper bound and the ablation should be run [27].

Generalisation must be measured, not assumed: two-thirds of the gain survives on unseen households but the socio-economic contribution does not, so evaluating only future readings of known households overstates what survey linkage buys for a new customer.

Threats to validity. The sample covers 131 of 4,063 RECON-SL households, is branch-stratified rather than representative, and each held-out fold has only 26–27 households, so the unseen-household estimates are imprecise and a larger sample might detect transferable socio-economic effects that this one cannot. The test windows span June to December 2024, not a full year; importance measures reliance, not causation; and the attributes are fixed at wave 1.

## VIII. CONCLUSION AND FUTURE WORK

We fused 15-minute smart-meter readings, satellite-derived weather and household survey attributes for 131 Sri Lankan households and benchmarked three model families against naive persistence under a chronological split, with selection on validation and a single test pass. Methodologically, gradient boosting improved on persistence by 5.56% in MAE (95% CI 3.89–7.39%), a gain that held at two earlier origins and, at 3.57%, on unseen households, while a regularised linear control was 6.86% worse than persistence yet had the highest $R^2$ at every origin. Substantively, socio-economic attributes contributed 2.61 percentage points when blanked at inference and 1.19 when the model was re-trained without them, less than half of what the uncontrolled comparison suggests, and nothing measurable for unseen households. The selected model runs as a service with magnitude-matched error bands and the baseline displayed alongside.

Three extensions follow. Re-training with each feature group removed would turn the reliance measures of Table IV into contributions. Re-extracting contiguous sequences from raw RECON-SL would enable the recurrent and attention-based comparison [6], [7] that the sampling precludes here, and a larger household sample would sharpen the unseen-household estimates. Wave-2 and wave-3 survey data would let the attributes vary in time, which matters because those that carry signal, appliance stock above all, are the most likely to change within fifteen months.

## REFERENCES


[1] Sisira Kumara and P. D. Nimal, "Drivers of household electricity
[2] expenditure: Insights from the highest electricity tariff-paying nation in South Asia," Energy Build., vol. 334, art. 115493, May 2025, doi: 10.1016/j.enbuild.2025.115493. [2] C. E. Kankanamge and P. S. A. D. Gunasekara, "Trends in appliance usage
[3] and electricity consumption in households: A case study in Hambantota District, Sri Lanka," Ruhuna J. Sci., vol. 13, no. 1, 2022, doi: 10.4038/rjs.v13i1.113. [3] S. Bouzarovski and S. Petrova, "A global perspective on domestic energy
[4] deprivation: Overcoming the energy poverty-fuel poverty binary," Energy Res. Social Sci., vol. 10, pp. 31-40, 2015. [4] P. Nussbaumer, M. Bazilian, and V. Modi, "Measuring energy poverty:
[5] Focusing on what matters," Renewable Sustainable Energy Rev., vol. 16, no. 1, pp. 231-243, 2012. [5] C. Algama, I. Fernando, M. Chandana, D. Dissanayake, N. Habaragamuwa,
[6] P. Ranage, J. Ramos, and K. Amarasinghe, "Residential Electricity Consumption Dataset for Sri Lanka (RECON-SL)," IEEE Dataport, Feb. 2025, doi: 10.21227/n1dk-q860. [6] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term
[7] residential load forecasting based on LSTM recurrent neural network," IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 841-851, Jan. 2019, doi: 10.1109/TSG.2017.2753802. [7] Q. Dong, R. Huang, C. Cui, D. Towey, L. Zhou, J. Tian, and J. Wang, "Short-
[8] term electricity-load forecasting by deep learning: A comprehensive survey," Eng. Appl. Artif. Intell., 2025. [Online]. Available: arXiv:2408.16202 [8] Q. Zhao, W. Xiang, B. Huang, J. Wang, and J. Fang, "Optimised extreme
[9] gradient boosting model for short term electric load demand forecasting of regional grid system," Sci. Rep., vol. 12, 2022, doi: 10.1038/s41598-022-22024-3. [9] N. Rubattu, G. Maroni, and G. Corani, "Electricity load and peak forecasting:
[10] Feature engineering, probabilistic LightGBM and temporal hierarchies," in Advanced Analytics and Learning on Temporal Data, Lecture Notes in Computer Science, vol. 14343. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-49896-1_18. [10] C. Miller et al., "The ASHRAE Great Energy Predictor III competition:
[11] Overview and results," Sci. Technol. Built Environ., vol. 26, no. 10, 2020,
[12] doi: 10.1080/23744731.2020.1795514. [11] H. Mansoor, S. Ali, I. Khan, N. Arshad, M. A. Khan, and S. Faizullah, "Short-
[13] term load forecasting using AMI data," IEEE Internet Things J., vol. 10, no. 24, pp. 22040-22050, Dec. 2023. [12] D. Maia-Silva, R. Kumar, and R. Nateghi, "The critical role of humidity in
[14] modeling summer electricity demand across the United States," Nature Commun., vol. 11, art. 1686, 2020, doi: 10.1038/s41467-020-15393-8. [13] M. S. AL-Musaylh, R. C. Deo, J. F. Adamowski, and Y. Li, "Short-term
[15] electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF Reanalysis atmospheric predictors in southeast Queensland, Australia," Renewable Sustainable Energy Rev., vol. 113, art. 109293, 2019, doi: 10.1016/j.rser.2019.109293. [14] NASA Langley Research Center, "POWER: Prediction of Worldwide
[16] Energy Resources," NASA Applied Sciences Program, Hampton, VA, USA. [Online]. Available: https://power.larc.nasa.gov/ [15] S. Abeywickrama and P. P. G. D. Asanka, "Integrating weather patterns into
[17] machine learning models for improved electricity demand forecasting in Sri Lanka," in Proc. 2023 Int. Res. Conf. Smart Computing and Systems Engineering (SCSE), Kelaniya, Sri Lanka, Jun. 2023. [16] G. Fernando and I. U. Hewapathirana, "Sector-specific electricity demand
[18] forecasting in Sri Lanka using deep learning and hybrid models," in Proc. 2025 Int. Res. Conf. Smart Computing and Systems Engineering (SCSE), 2025, pp. 1-7, doi: 10.1109/SCSE65633.2025.11030954. [17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin,
[19] "CatBoost: Unbiased boosting with categorical features," in Proc. 32nd Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Montreal, QC, Canada, 2018, pp. 6639-6649. [18] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu,
[20] "LightGBM: A highly efficient gradient boosting decision tree," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 3149-3157. [19] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for
[21] nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55-67, 1970. [20] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast
[22] accuracy," Int. J. Forecasting, vol. 22, no. 4, pp. 679-688, 2006. [21] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and
[23] Practice, 3rd ed. Melbourne, Australia: OTexts, 2021. [22] C. Bergmeir and J. M. Benitez, "On the use of cross-validation for time series
[24] predictor evaluation," Inf. Sci., vol. 191, pp. 192-213, 2012. [23] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination
[25] R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., vol. 7, art. e623, 2021. [24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model
[26] predictions," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 4765-4774. [25] D. Sculley et al., "Hidden technical debt in machine learning systems," in
[27] Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Montreal, QC, Canada, 2015, pp. 2503-2511. [26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in
[28] Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, 2016, pp. 785-794. [27] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd
[29] ed. Hoboken, NJ, USA: Wiley, 2019. [28] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," J. Bus.
[30] Econ. Statist., vol. 13, no. 3, pp. 253–263, 1995. [29] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: An analysis and
[31] review," Int. J. Forecasting, vol. 16, no. 4, pp. 437–450, 2000. [30] A. C. Cameron, J. B. Gelbach, and D. L. Miller, "Bootstrap-based
[32] improvements for inference with clustered errors," Rev. Econ. Statist., vol. 90, no. 3, pp. 414–427, 2008. [31] T. Gneiting, "Making and evaluating point forecasts," J. Amer. Statist. Assoc.,
[33] vol. 106, no. 494, pp. 746–762, 2011.

# Leakage-Aware Multi-Horizon PM2.5 Forecasting at a Single New Delhi Station: Model Comparison, Conformal Uncertainty, and Actual-Model Explanations

AHK Thiwanka
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
thiwanka.ahk@gmail.com

RWKT Rajapaksha
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
kavishka.r@nsbm.ac.lk

***Abstract*—Short-term PM2.5 studies often compare forecasting models under incompatible timestamps, input assumptions, and temporal partitions. This single-station case study evaluates a leakage-aware multi-horizon framework using OpenAQ New Delhi location 8118 and coordinate-matched NASA POWER meteorology. Its contribution is methodological and evaluative rather than a new forecasting architecture. Models were initially fitted on 2017–2023, tuned using development data through 2024, and refitted through 2024 with frozen configurations before testing on common timestamps from 2025 to May 2026. Paired moving-block bootstrap comparisons, feature ablation, severe-event analysis, split-conformal intervals, and fidelity-checked explanations were evaluated using stored point forecasts. At one hour, GRU achieved RMSE 19.018 µg/m³ and was statistically indistinguishable from Extra Trees. At six hours, the equal-weight XGBoost–GRU ensemble achieved RMSE 37.810 µg/m³ and showed a statistically supported improvement over its GRU component and Extra Trees. At 24 hours, an ensemble GRU component achieved RMSE 41.292 µg/m³, but leading alternatives remained statistically indistinguishable; its subsequent diagnostic selection was post-hoc. Nominal 90% conformal intervals achieved only 64.6–69.4% coverage during severe events. TreeSHAP and Integrated Gradients described changing predictive dependence, with substantial baseline sensitivity at 24 hours. Short-gap interpolation was retrospective and could use post-issue observations; observed-history sensitivity results are therefore reported, and complete operational leakage prevention is not established. The findings support horizon-specific evaluation at this station, without establishing Delhi-wide generalization or reliable extreme-event calibration.***

***Keywords—PM2.5 forecasting, multi-horizon forecasting, conformal prediction, explainable artificial intelligence, New Delhi***

## I. INTRODUCTION

Fine particulate matter (PM2.5) is a major environmental health risk because particles can penetrate deeply into the respiratory system, and the burden is especially important in densely populated urban regions [1], [2]. Short-horizon forecasts may support earlier awareness, but a precise point estimate can still be misleading when its uncertainty, severe-event behaviour, and evidential basis are hidden. New Delhi was selected as a practical South Asian case study because the available station record covered several years and contained wide seasonal and concentration variation, including severe pollution episodes. These characteristics support meaningful evaluation across 1-, 6-, and 24-hour horizons.

PM2.5 forecasting research includes statistical, tree-based, recurrent, graph-based, and hybrid approaches. New Delhi studies have reported benefits from optimized hybrid learning [3] and multi-station spatiotemporal modelling [4]. However, published scores are rarely directly comparable because studies differ in monitoring sites, input availability, missing-data treatment, forecast horizons, temporal splits, and eligible timestamps. Random splitting and retrospectively supplied future weather can produce optimistic estimates that do not represent issue-time forecasting. Complex models are also sometimes reported without strong persistence baselines or uncertainty around performance differences.

This study asks to what extent a common, leakage-aware multi-model framework can provide accurate, uncertainty-aware, and explainable 1-, 6-, and 24-hour PM2.5 forecasts at one New Delhi monitoring station. The novelty is methodological and evaluative, not a new forecasting architecture. It makes three contributions: (1) a broad comparison under common chronological splits and common test timestamps; (2) paired block-bootstrap inference and temporally separated conformal calibration and evaluation, including severe-event reliability; and (3) faithful explanations of the actual selected tree, recurrent, and hybrid forecasts. The framework improves upon accuracy-only comparisons by controlling temporal partitions and timestamp eligibility, testing whether performance differences are statistically credible, and separately evaluating uncertainty and explanation reliability. AirAware implements a bounded evidence-translation layer, but the scientific claims derive from the locked forecasting experiment rather than from untested claims of user benefit.

## II. RELATED WORK AND RESEARCH GAP

Classical autoregressive models remain valuable because they expose temporal assumptions and provide credible comparators. Tree ensembles such as Random Forest and Extra

Trees represent nonlinear effects in engineered pollutant, weather, and calendar features [5], [6], while XGBoost adds regularized boosting and efficient interaction modelling [7]. LSTM and GRU networks learn from ordered sequences and can represent longer temporal dependencies [8], [9]. Recent work has also evaluated deep-learning time-series models across diverse PM2.5 forecast horizons [10]. Hybrid systems may combine complementary tabular and sequential behaviour, but additional complexity is justified only when evaluated improvements exceed sampling variation.

Recent reviews find that explainability is still less consistently evaluated than predictive accuracy [11]. Applied SHAP-based studies demonstrate useful feature-level summaries [12], but post-hoc importance can be unstable when inputs are correlated or a reconstructed model does not reproduce the reported predictions. SHAP provides an additive framework for tree explanations [13], whereas Integrated Gradients attributes neural predictions relative to a chosen baseline. Both require fidelity checks, and neural explanations require baseline-sensitivity analysis before strong interpretation.

Conformal prediction offers distribution-free marginal coverage under exchangeability [14], while newer time-series methods seek to address dependence and non-stationarity [15], [16]. Uncertainty-aware PM2.5 models have also been proposed [17], yet nominal coverage alone can hide failure in high-concentration regimes. The unresolved gap is therefore not simply the absence of another architecture. Limited evidence combines broad horizon-specific comparison, common timestamp eligibility, time-respecting selection, paired inference, temporally separated conformal calibration and evaluation, severe-event assessment, and validated explanations of the actual selected models under one New Delhi protocol.

The model families also make different assumptions. Persistence is a demanding short-horizon baseline because urban PM2.5 often changes gradually. Linear and autoregressive methods provide transparent temporal comparators but have limited capacity for nonlinear interactions. Tree ensembles can use lagged pollution, weather, and calendar features without imposing linear relationships, while recurrent networks learn ordered representations directly from multivariate sequences. The XGBoost–GRU combination was included to test whether these tabular and sequential views provide complementary errors rather than to assume that a hybrid is automatically superior.

## III. RESEARCH METHODOLOGY AND PROPOSED APPROACH

### *A. Data and preprocessing*

The repository contains hourly PM2.5 downloaded from OpenAQ New Delhi location 8118 (28.63576° N, 77.22445° E) [18] and coordinate-matched NASA POWER meteorology [19]. The merged hourly grid contains 82,488 timestamps: 69,084 original PM2.5 observations, 2,896 values interpolated only across one-to-three-hour internal gaps, 10,464 longer-gap missing states, and 44 nearby off-grid missing states. The model-ready table contains 71,980 non-missing PM2.5 rows. Meteorological predictors comprise pressure, temperature, dew point, relative humidity, corrected precipitation, shortwave radiation, wind speed, and cyclic wind direction. Repository documentation identifies location 8118 and the coordinates above; the raw CSV contains timestamps and PM2.5 but no embedded request metadata. The historical endpoint/API version, complete request parameters, acquisition date, and contemporaneous licence snapshot could not be recovered; current metadata cannot establish the historical licence. Raw extracts and quality flags are retained, but exact reacquisition is not assured.

Quality flags distinguish observed, interpolated, and unavailable PM2.5. Time-linear interpolation fills only internal gaps of one to three hours without nearby off-grid observations; longer gaps remain missing. Interpolated values can enter inputs, but training and evaluation targets must be original observations. Because interpolation uses both endpoints, it can depend on an observation after issue time: this occurs in 34/11,090, 45/11,030, and 42/10,865 final cases at 1, 6, and 24 hours. Thus, the primary benchmark is retrospective rather than fully issue-time leakage-free. Off-grid observations are not silently shifted. NASA sentinel values are checked; PRECTOTCORR is retained as supplied without an additional unit conversion. Wind direction uses sine/cosine encoding.

### *B. Forecast protocol and models*

Separate targets occur exactly 1, 6, or 24 hours after issue time. Input timestamps do not exceed issue time; future observed weather is excluded from primary models. Initial fitting uses 2017–2023; tuning uses development data ending in 2024. After configuration selection, the selected models and preprocessing statistics are refitted using eligible 2017–2024 development data and frozen before 2025–May 2026 testing. Benchmarks include persistence, seasonal-naive and historical-mean baselines; Ridge, SARIMA, explicitly labelled oracle SARIMAX; tree ensembles and XGBoost; LSTM, GRU, corrected 72-hour CNN–LSTM; and XGBoost–GRU ensembles.

Splits follow local target time; final tests also require issue time in 2025 or later. Exact timestamp offsets prevent missing hours from altering the forecast horizon. Common eligible timestamps yield 11,090, 11,030, and 10,865 test cases at 1, 6, and 24 hours. Training-only scaling is used during validation; final scaling is fitted through 2024. Primary features are predefined without full-period correlation screening. The weather-screened ablation uses the full record and is exploratory. The corrected CNN–LSTM enforces 72 complete hourly steps and remains non-selected.

### *C. Evaluation, uncertainty, and explanation*

MAE, RMSE, and $R^2$ quantify point accuracy, with RMSE used for validation-based selection. Paired moving-block bootstrap comparisons use 168-observation blocks and 2,000 resamples to preserve weekly error dependence. Severe events are observations $\geq 150\ \mu g/m^3$. After point forecasts were locked, residuals from 2025 calibrated symmetric and asymmetric 90% split-conformal intervals; January–May 2026 was reserved for interval evaluation. Coverage, width, interval score, and severe-event coverage were reported. TreeSHAP explains the six-hour XGBoost component, while Integrated Gradients explains the one-hour GRU and the recurrent components at six and 24

hours. Six-hour attributions are combined using the locked 0.5/0.5 ensemble weights.

Bootstrap blocks contain 24 or 168 consecutive eligible errors (nominally daily/weekly); missing-hour exclusions mean these are not always contiguous clock hours. Point comparisons use non-circular moving blocks; interval diagnostics use circular blocks, with 2,000 resamples. Symmetric intervals use the finite-sample 90% quantile of absolute 2025 residuals; asymmetric intervals use signed residual 5%/95% quantiles. Lower endpoints are clipped at zero. Exchangeability and severe-event conditional coverage are not guaranteed for this dependent series.

For conformal evaluation, 2025 residuals alone determine the interval offsets and January–May 2026 observations are used only to assess them. Prediction parity verifies that each reconstructed XAI model reproduces the locked forecast before it is explained. Completeness checks compare the attributed contribution sum with the model-output difference, resampling checks rank stability, and alternative Integrated-Gradients baselines test dependence on the chosen reference input. The 24-hour GRU component was chosen for diagnostic reporting after inspecting final-test errors; its interval and XAI results are conditional, post-hoc analyses, not an independent confirmation of a validation-selected winner.

## IV. SYSTEM AND MODEL DESIGN AND IMPLEMENTATION

The implemented research workflow separates evidence gates. Raw provider files are preserved; preprocessing scripts create quality flags and aligned hourly tables; horizon-specific feature builders construct exact timestamp targets; validation code tunes configurations without viewing the final period; and final evaluation writes immutable predictions, metrics, significance tests, interval results, XAI outputs, and reproducibility manifests. Fixed random seeds, versioned configurations, dependency pins, and explicit sample-eligibility checks reduce accidental variation. A corrected CNN–LSTM rerun enforces its full 72-hour input requirement and remains non-selected, preventing a discovered eligibility defect from changing the locked conclusions.

Tabular models use 26 features: current PM2.5; lags of 1, 2, 3, 6, 12, 24, 48, 72, and 168 hours; nine weather channels including cyclic wind direction; and target-time hour, weekday, month sine/cosine pairs plus a weekend flag. Recurrent inputs contain 16 channels per step: PM2.5, the same weather channels, and issue-history calendar sine/cosine pairs. Ridge searches $\alpha \in \{0, 0.1, 1, 10, 100\}$, selecting 0, 0, and 100 by 2024 RMSE. SARIMA/SARIMAX orders are fixed at $(1,0,0)\times(1,0,0,24)$; SARIMAX uses future weather and is an oracle comparator.

Focused tree tuning compares six candidates with 100 trees, using mean RMSE over expanding 2022, 2023, and 2024 validation folds. Extra Trees candidates (depth, minimum leaf size, feature fraction) are (unlimited,2,1.0), (unlimited,1,0.8), and (30,2,0.8); Random Forest candidates are (unlimited,2,sqrt), (unlimited,1,0.8), and (30,2,0.8). The first Extra Trees configuration is retained at 1/6 hours and the first Random Forest configuration at 24 hours. Initial fixed RF/Extra Trees benchmarks use 200 trees. Neural models use one 32-unit recurrent layer, a linear output, Adam (0.001), MSE, batch size 256, gradient clipping at 1, at most 30 epochs, patience 5, and seed 42. Initial LSTM/GRU/CNN–LSTM windows are 72 hours; the CNN has 32 filters of width 3 and ReLU. Window tuning compares 24/72/168 hours on common 2024 timestamps, retaining 24 hours. Locked standalone configurations are GRU/15 epochs (1 hour), GRU/11 (6 hours), and LSTM/4 (24 hours).

For the ensemble, both components are initially trained on 2017–2023. January–June 2024 selects stopping iterations and the XGBoost weight from {0,0.05,…,1}; July–December 2024 evaluates the frozen blend. XGBoost uses learning rate 0.05, depth 6, minimum child weight 3, subsample/column fractions 0.8, squared-error loss, and a 1,000-round cap with patience 50. The retained 6-hour blend has weights 0.5/0.5, 49 boosting rounds and 13 GRU epochs; the 24-hour blend has weights 0.85/0.15, 47 rounds and 5 GRU epochs. Both use 24-hour sequences. Fixed configurations are refitted through 2024. “Component” denotes a constituent fitted predictor, not the separately tuned recurrent model. In particular, the reported 24-hour GRU component is distinct from the validation-selected standalone LSTM.

The selected recurrent models were deterministically reconstructed for XAI and compared with locked CSV predictions. Maximum discrepancies were approximately 0.000015–0.000021 μg/m³, well below the 0.1 μg/m³ parity tolerance. Attributions were converted into PM2.5 output units and grouped into current PM2.5, recent history, older history, meteorology, wind, and calendar/seasonal information. Completeness, rank stability, and Integrated-Gradients baseline sensitivity were tested. These checks prevent an attractive explanation from being reported when it describes a different fitted function or depends strongly on an arbitrary reference.

AirAware exposes the verified outputs through typed FastAPI interfaces and a React presentation layer with a PostgreSQL audit trail. It separates historical research replay from a guarded operational pilot and distinguishes observations, point forecasts, intervals, explanations, provenance, and limitations. No official AQI, medical advice, causal conclusion, or hyperlocal claim is generated. The application demonstrates evidence translation; it is not treated as participant evidence that decisions or health outcomes improved.

## V. RESULTS AND DISCUSSION

### *A. Point forecasting and statistical evidence*

Table I summarizes the reported horizon-specific predictors; the 24-hour row is a post-hoc component diagnostic. At one hour, the tuned GRU obtained RMSE 19.018 μg/m³ and $R^2$ 0.9424, while Extra Trees obtained RMSE 19.170 μg/m³. The GRU's 0.152 μg/m³ RMSE improvement had a 95% bootstrap interval of [-0.222, 0.564]; therefore, no difference is detected by this bootstrap comparison. At six hours, the equal-weight XGBoost–GRU ensemble obtained RMSE 37.810 μg/m³. Its improvement over the GRU component was 0.865 [0.255, 1.513] and over Extra Trees was 0.914 [0.115, 1.612], making this the clearest evidence that hybridization added value. At 24 hours, the GRU component achieved RMSE 41.292 μg/m³, but

its differences from the retained ensemble and Random Forest included zero. All selected approaches improved credibly on persistence.

TABLE I. FINAL POINT-FORECAST PERFORMANCE (MAE AND RMSE IN µG/M³)

| Horizon | Model | MAE | RMSE | $R^2$ | Inference |
|---|---|---|---|---|---|
| 1 h | GRU | 10.989 | 19.018 | 0.942 | Tied with Extra Trees |
| 6 h | XGB-GRU | 23.908 | 37.810 | 0.771 | Credible hybrid gain |
| 24 h | GRU component | 26.416 | 41.292 | 0.722 | Leading models tied |

### *B. Predictor value and difficult conditions*

Tree-model ablation showed that adding all issue-time weather and calendar inputs to PM2.5 history reduced RMSE by approximately 4.53%, 15.16%, and 5.13% at 1, 6, and 24 hours, respectively. Weather alone performed poorly, confirming that pollutant history remains indispensable. Severe-event RMSE rose to 35.242, 70.836, and 76.322 µg/m³. During one episode, the observation reached approximately 616 µg/m³ and all selected forecasts strongly underpredicted the peak. This negative result is important because average accuracy would otherwise conceal the behaviour most relevant to warning communication.

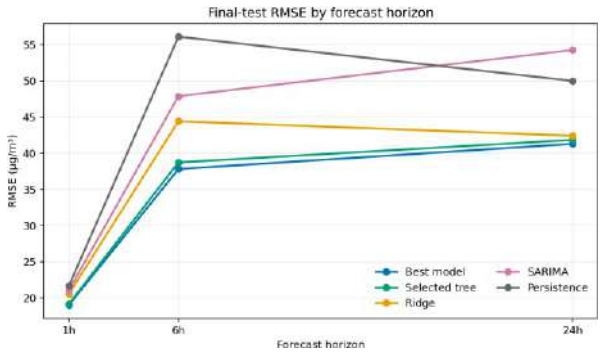



Fig. 1. Final-test RMSE on common eligible timestamps. Numerical rankings differ by horizon; the 24-hour GRU component is reported post-hoc.

The ablation comparison used the same eligible timestamps within each horizon. For the representative tree models, adding all issue-time weather and calendar inputs to PM2.5 history reduced RMSE from 20.080 to 19.170 µg/m³ at one hour, from 45.646 to 38.724 µg/m³ at six hours, and from 44.100 to 41.835 µg/m³ at 24 hours. The largest relative gain therefore occurred at six hours. Correlation-screened weather results remain post-hoc because the screening step used the full period and are not used as confirmatory evidence.

In the existing sensitivity subset requiring original observations throughout the model history and baseline inputs, the same fixed predictors achieve RMSE 18.537, 36.327, and 39.177 µg/m³ (n = 10,254, 10,185, and 10,058). These different subsets do not isolate an interpolation effect, remove interpolated training inputs, or establish unchanged rankings.

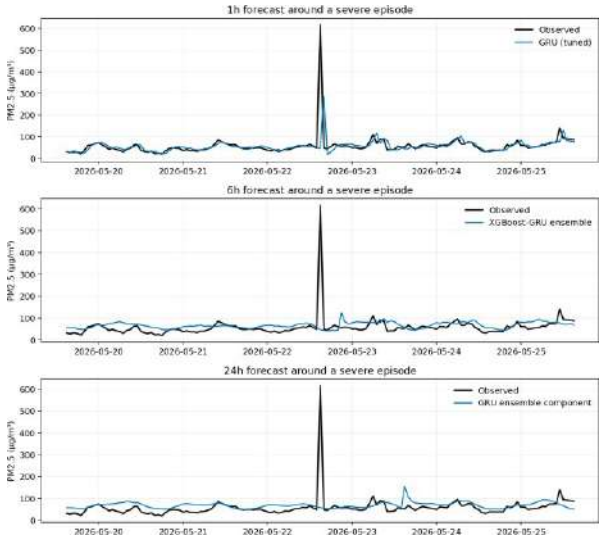



Fig. 2. Observed and selected-model forecasts during one severe episode. This case illustrates peak underprediction and is not a claim about every severe event.

Fig. 2 presents one documented severe episode in which the observed series rose to approximately 616 µg/m³ while every selected forecast remained substantially lower. It illustrates the practical form of the aggregate severe-event errors but should not be interpreted as the behaviour of every episode. Across all severe cases, RMSE increased to 35.242, 70.836, and 76.322 µg/m³ for the 1-, 6-, and 24-hour forecasts, respectively.

### *C. Uncertainty reliability*

Table II reports the symmetric interval baseline. Overall coverage was 85.8%, 87.2%, and 88.2% at the three horizons, below the nominal 90%. Asymmetric overall coverage was 85.6%, 86.9%, and 88.4%, with severe coverage 64.6%, 67.2%, and 68.7%. Severe-event coverage fell to 64.6–69.4%, while intervals also became much more costly under the interval score. Calibration on 2025 therefore did not fully transfer across the January 2026 boundary. Pooled residual quantiles need not represent the larger, heteroskedastic errors during severe episodes, and temporal regime changes can further weaken calibration. These are plausible explanations, not established causes. The intervals are not reliably calibrated for extreme events. Adaptive or locally scaled methods require a new untouched period if designed after observing these failures.

TABLE II. SYMMETRIC 90% CONFORMAL INTERVALS (JANUARY–MAY 2026)

| Horizon | Coverage | Width | Severe n | Severe coverage |
|---|---|---|---|---|
| 1 h | 85.8% | 45.1 | 588 | 64.6% |
| 6 h | 87.2% | 101.7 | 588 | 67.7% |
| 24 h | 88.2% | 113.4 | 588 | 69.4% |

[a] Width is the mean clipped interval width in µg/m³; severe means observed PM2.5 ≥150 µg/m³. Calibration: 2025. Severe n = 588 at each horizon.

The interval sample counts were 3,277, 3,272, and 3,254 for the three horizons, including 588 severe observations at each horizon. Under 168-observation moving-block resampling, the symmetric overall-coverage intervals were approximately [81.7%, 89.4%], [82.1%, 91.9%], and [83.1%, 92.9%]. These wider intervals reflect temporal dependence and include 90% at the longer horizons, but observed pointwise coverage and severe-event coverage remain below target. In particular, the corresponding severe-coverage intervals were approximately [58.4%, 69.9%], [56.9%, 75.9%], and [57.8%, 79.2%], providing no basis for claiming reliable severe-event calibration. For 24-observation blocks, overall-coverage 95% intervals are [83.9%,87.7%], [84.6%,89.6%], and [85.3%,90.8%]; severe-coverage intervals are [59.6%,69.4%], [60.3%,74.2%], and [60.9%,77.2%].

### *D. Actual-model explanations*

Prediction parity passed for every explained component, and maximum attribution-completeness error remained below 0.00018 µg/m³. Fig. 3 shows that current PM2.5 represented 71.8% of grouped absolute attribution at one hour. At six hours, current PM2.5 (27.8%) and meteorology (27.4%) were balanced, with older history contributing 21.3%. At 24 hours, meteorology (30.3%) and recent history (23.9%) led. These are predictive associations, not environmental causes. Integrated-Gradients six-group baseline rank correlation declined from 0.943 at one hour to 0.600 at six hours and 0.257 at 24 hours, so the long-horizon ranking is not uniquely determined by the fitted model alone.

Resampled factor-rank stability was high for the reported explanation pipelines, with median Spearman correlation of at least 0.986. This stability does not remove reference dependence: changing the Integrated-Gradients baseline for the six displayed groups produced correlations of 0.943, 0.600, and 0.257 at 1, 6, and 24 hours. The six-hour weighted explanation is especially informative because it combines TreeSHAP and neural attributions using the same fixed 0.5/0.5 weights as the forecast. At 24 hours, only broad dependence patterns should be communicated because the exact ordering changes substantially with the baseline.

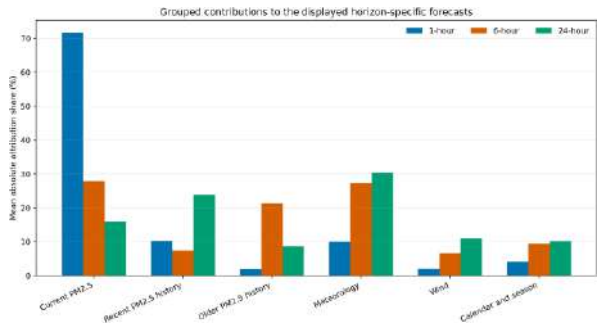



Fig. 3. Mean grouped absolute attribution shares for the displayed horizon-specific forecasts. Values describe model dependence, not causation.

### *E. Threats to validity*

The study uses one station-level temporal series and gridded meteorology, so results do not establish Delhi-wide or hyperlocal performance. The historical OpenAQ request metadata are incomplete; the location identification is confirmed, but the acquisition and licence snapshot must remain disclosed. The correlation-selected ablation is post-hoc, severe events are relatively sparse, the final period represents one temporal regime, and the 24-hour neural explanation is baseline-sensitive. AirAware has technical verification but no participant evidence of improved decision-making. Retrospective interpolation can cross issue times and split boundaries; the observed-history subset cannot remove its influence on fitted models. Archived NASA weather also does not establish provider publication latency at issue time. The 24-hour diagnostic predictor was chosen after test inspection. These qualifications limit operational and confirmatory claims.

Internal validity is strengthened by chronological selection, locked tests, common timestamps, explicit oracle labelling, and paired inference, but it is not absolute. Hyperparameter choices are limited to the explored search spaces, and residual temporal dependence may not be fully represented by either block length. Construct validity is limited because the ≥150 µg/m³ threshold is a research segmentation rule rather than an official hourly AQI category. External validity is constrained by one monitoring location and one final temporal period. The numerical results should therefore be interpreted as station- and protocol-specific evidence rather than universal rankings of forecasting algorithms.

## VI. CONCLUSION AND FUTURE WORK

At this single New Delhi station, a retrospective comparison across common timestamps shows that PM2.5 model choice should be horizon-specific. GRU and Extra Trees are effectively tied at one hour; the equal-weight XGBoost–GRU ensemble provides the only statistically credible hybrid improvement at six hours; and no separation is detected among leading 24-hour approaches; the GRU component is a post-hoc diagnostic. Conformal intervals reveal substantial severe-event undercoverage, while validated TreeSHAP and Integrated-Gradients explanations show changing predictive dependence and an important long-horizon sensitivity limitation. Subject to the interpolation and post-hoc-selection limitations, the framework addresses the research question by combining point-forecast accuracy with explicit evidence about uncertainty, statistically credible model differences, and explanation reliability. Its principal advantage over accuracy-only approaches is the reliability and completeness of this evidence, rather than universal numerical superiority. Future work should recover request-level provenance, evaluate adaptive intervals on a new untouched period, and test broader spatial generalization without weakening the current leakage controls. A planned extension will apply the same leakage-aware evaluation framework to Sri Lankan monitoring locations using locally collected PM2.5 and meteorological data. This extension should involve site-specific data-quality assessment, model retraining, uncertainty calibration, and evaluation on an untouched local period rather than directly transferring the New Delhi models.

## REFERENCES

[1] World Health Organization, “Ambient (outdoor) air pollution,” World Health Organization. Accessed: Sep. 10, 2026. [Online]. Available:

https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] Health Effects Institute, "State of Global Air 2024," Boston, 2024. Accessed: Sep. 10, 2026. [Online]. Available: https://www.stateofglobalair.org/resources/report/state-global-air-report-2024

[3] A. Masood *et al.*, "Improving PM2.5 prediction in New Delhi using a hybrid extreme learning machine coupled with snake optimization algorithm," *Sci. Rep.*, vol. 13, no. 1, p. 21057, Dec. 2023, doi: 10.1038/s41598-023-47492-z.

[4] M. Panja, T. Chakraborty, A. Biswas, and S. Deb, "E-STGCN: Extreme Spatiotemporal Graph Convolutional Networks for Air Quality Forecasting," Jun. 2025. Accessed: Nov. 02, 2026. [Online]. Available: https://arxiv.org/abs/2411.12258v2

[5] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Sep. 2001, doi: 10.1023/A:1010933404324.

[6] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[9] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Alessandro Moschitti, Walter Daelemans, and Bo Pang, Eds., Doha, Qatar: Association for Computational Linguistics, Sep. 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.

[10] L. Zeng, R. Dong, M. Yuan, L. Jing, and S. Jiao, "Evaluating deep learning time series models for PM2.5 forecasting across diverse horizons," *iScience*, vol. 29, no. 2, p. 114770, Feb. 2026, doi: 10.1016/j.isci.2026.114770.

[11] A. Houdou *et al.*, "Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review," *Aerosol Air Qual. Res.*, vol. 24, no. 1, p. 230151, Jan. 2024, doi: 10.4209/aaqr.230151.

[12] S. M. Malakouti, "From accurate to actionable: Interpretable PM2.5 forecasting with feature engineering and SHAP for the Liverpool–Wirral region," *Environmental Challenges*, vol. 21, p. 101290, Dec. 2025, doi: 10.1016/j.envc.2025.101290.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Curran Associates, Inc., 2017, pp. 4765–4774. Accessed: Sep. 10, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[14] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World*, 1st ed. New York: Springer-Verlag, 2005. doi: 10.1007/b106715.

[15] A. Cini, A. Jenkins, D. Mandic, C. Alippi, and F. M. Bianchi, "Relational Conformal Prediction for Correlated Time Series," Jun. 2025. Accessed: Nov. 02, 2026. [Online]. Available: https://arxiv.org/abs/2502.09443v2

[16] R. Neglia, A. Cini, M. M. Bronstein, and F. M. Bianchi, "ResCP: Reservoir Conformal Prediction for Time Series Forecasting," Mar. 2026, [Online]. Available: http://arxiv.org/abs/2510.05060

[17] Q. Li, H. Lu, H. Xu, and D. Sun, "Flipout Bayesian LSTM with Residual Attention for Uncertainty-Aware PM2.5 Forecasting and Anomaly Detection," *Sustainability (Switzerland)*, vol. 18, no. 4, p. 1718, Feb. 2026, doi: 10.3390/su18041718.

[18] OpenAQ, "About us," OpenAQ. Accessed: Sep. 10, 2026. [Online]. Available: https://openaq.org/about/

[19] National Aeronautics and Space Administration, "NASA POWER Project," NASA POWER. Accessed: Sep. 10, 2026. [Online]. Available: https://power.larc.nasa.gov/

# Development and Evaluation of a Context-Aware Multimodal Machine Learning Framework for Real-Time Stress Inference Using Smartphone Camera PPG-Derived HRV, Keystroke Dynamics and Environmental Sensing

W. G. G. A. Sahan Kaushalya Dias
*Dept. of Software Engineering & Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
askdias@students.nsbm.ac.lk

Nimesha Hewawasam
*Dept. of Software Engineering & Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
nimesha.h@nsbm.ac.lk

***Abstract*—Stress among university students is usually measured with retrospective questionnaires or with wearable sensors that few students own. This paper presents a context-aware multimodal framework that predicts self-reported stress using only a standard Android smartphone. Heart rate variability (HRV) is derived from photoplethysmography (PPG) recorded with the rear camera and flash, behavioral features come from a copy-typing task and the previous day's phone use, and environmental context comes from ambient noise, device motion, location type and time of day. A Flutter application captures each session, Cloud Firestore stores it, and a FastAPI service scores it with one Random Forest per modality combined by late fusion. The modelling approach was first tested on the public WESAD dataset, where leave-one-subject-out accuracy rose from 0.702 with heart rate alone to 0.908 with the full multimodal feature set (Wilcoxon p = 0.0001). The framework was then evaluated on 300 sessions from 100 IT undergraduates at one Sri Lankan university, using participant-grouped cross-validation and a held-out set of 20 unseen participants. Single-modality test ROC AUC was 0.961 for PPG, 0.889 for behavior and 0.892 for environment, and late fusion raised test F1 from 0.837 to between 0.917 and 0.927. Integrity diagnostics, however, found almost no participant-level structure in the session data (residual intraclass correlation of 0.092 or lower), so these figures confirm that the pipeline works end to end but cannot be taken as the accuracy the system would achieve with new students. Camera-derived HRV was not validated against a beat-to-beat reference. The study shows that a smartphone alone can host the complete stress-prediction workflow, and confirmation on independently collected, supervised data with a reference ECG is still required.**

***Index Terms*—stress detection, heart rate variability, photoplethysmography, keystroke dynamics, late fusion**

## I. Introduction

Academic stress is linked to lower motivation, weaker performance, poorer sleep and a higher risk of mental health problems, and university students are especially exposed to it [1]. Students of information technology (IT) carry a steady load of programming assignments, group projects with fixed deadlines and practical examinations, often working late at a screen. This study targets IT undergraduates at NSBM Green University in Sri Lanka, a population for which sensor-based stress research is scarce.

Stress is normally measured in one of two ways. Questionnaires such as the Perceived Stress Scale are cheap and well validated but retrospective, so they cannot tell a student that the last hour of work has pushed them into a stressed state. Physiological monitoring captures the autonomic response directly [2], but research systems depend on chest-worn electrocardiography (ECG) belts or research wristbands [3] that ordinary students do not own. A smartphone, however, already contains most of the sensors needed. Its rear camera and flash can record photoplethysmography (PPG), from which heart rate variability (HRV) can be derived [2], [4]. Its touchscreen records typing, its operating system logs screen use, and its microphone and accelerometer measure loudness and movement. Passive phone data has been linked to student stress [5], [6], yet most of this work treats the phone as a behavioral sensor and leaves physiology to a separate wearable [7].

Three further gaps motivate this work. Many published models report high accuracy under validation schemes in which the same person appears in training and testing, which overstates performance for new users [8], [9]. Few systems are carried through to an application that returns a prediction to the user within one session. And none of the datasets found in this review describe Sri Lankan university students.

This paper asks whether a standard smartphone, without any wearable, can support a multimodal framework that predicts self-reported stress for people it has never seen. It makes four contributions. The first is an Android application that runs a five-minute physiological, behavioral and environmental session. The second is a pipeline with one model per modality and late fusion, evaluated under participant-grouped cross-validation and a single-use held-out set. The third is a feasibility study on the public WESAD dataset showing that the chosen feature types carry stress information across people. The fourth

is a set of dataset integrity diagnostics which show that the collected session data cannot yet support claims about stress detection in new students. Section II reviews related work, Section III describes the method, Section IV reports and discusses the results, and Section V concludes.

## II. Related Work

### A. HRV and Smartphone PPG

Stress withdraws parasympathetic activity, which lowers HRV. A meta-analysis by Kim *et al.* concluded that reduced time-domain HRV, such as RMSSD, is one of the most practical objective markers of psychological stress [2]. Time-domain measures are preferred for short recordings because frequency-domain measures need longer, more stationary signals [4]. PPG estimates the pulse from changes in reflected light, and the interval between successive peaks approximates the RR interval. Tsai *et al.* reported that RR intervals from a wearable PPG device correlated with ECG at $r = 0.992$ [10], although their device was not a phone camera. A phone camera samples at about 30 frames per second, which limits peak timing to roughly 33 ms, so camera HRV must be treated as a noisy signal and gated for quality.

### B. Behavioral and Environmental Sensing

In the StudentLife study, passive phone data from 48 students was related to stress, depression and academic performance [5], and a review of 40 digital phenotyping studies found that phone use, activity and sleep were each associated with stress, anxiety or mild depression [6]. Keystroke timing distinguished emotional states with 77 to 88% accuracy in Epp *et al.* [11]. Noise is a documented environmental stressor [12]. The closest antecedent is Aqajari *et al.*, who raised stress classification F1 from 56% to 70% by adding phone context to smartwatch PPG [7]. Their eleven participants still needed a smartwatch, and the phone supplied only context.

### C. Machine Learning and Evaluation Practice

Stress detection from engineered features is usually treated as supervised classification, with tree ensembles and support vector machines most common [3]. Random Forest [13] needs little tuning and resists overfitting on small tables, while Gradient Boosting [14] can be more accurate but is more sensitive to hyperparameters. Early fusion concatenates all features into one model, which hides each source's contribution and fails when a modality is missing. Late fusion trains one model per modality and combines their outputs, which makes each contribution measurable. Evaluation choices matter as much as models. Saeb *et al.* showed that record-wise validation lets a model recognize the person rather than the condition [9], and Zhang *et al.* found overfitting and cross-user shift to be the main reasons stress models fail to generalize [8]. Precision-recall analysis is more informative than ROC analysis on imbalanced data [15]. Table I summarizes representative systems.

TABLE I. Representative Reviewed Studies

| Study | Sensing / algorithm | Best result | Main limitation |
|---|---|---|---|
| Al-Alim [16] | ECG, skin temp., EDA / KNN | 98.0% | Wearable, no PPG or context |
| Rashid [17] | Wrist BVP, EDA, ACC / ensemble | 94.12% | Wearable required |
| Moser [18] | E4 EDA, temperature / RF | 92.74% | Wearable, 16 subjects |
| Aqajari [7] | Watch PPG, phone context / RF | F1 70% | Watch needed, $n = 11$ |
| Izzah [19] | ECG HRV / SVM, KNN, RF | 54 to 62% | Low cross-subject accuracy |
| Dahal [20] | ECG HRV / RF, mRMR | 99 to 100% | Per-subject split inflates result |
| Seo [21] | ECG, respiration, face / DL | 73.3% | Laboratory only |

No reviewed system combines physiological, behavioral and environmental evidence using a smartphone alone, many high scores come from optimistic validation, and few systems return a result to the participant. This work addresses these points, with the evaluation designed to expose rather than hide weaknesses in the data.

## III. Methodology

### A. Research Design

The study follows Design Science Research [22] within a post-positivist stance. An artefact was built and then evaluated deductively against criteria fixed in advance, namely accuracy clearly above the majority-class baseline and ROC AUC above 0.90. The work ran in five phases, namely a feasibility study on public data, construction of the application and data layer, session collection, model training and fusion, and deployment with a validity analysis of the collected data.

### B. Feasibility Study on WESAD

Before any data was collected with the phone, the modelling idea was tested on the wrist signals of the WESAD dataset [23], which comprise blood volume pulse (64 Hz), electrodermal activity (EDA), skin temperature and three-axis acceleration from 15 participants. Stress induced by the Trier Social Stress Test formed the positive class, and baseline and amusement the negative class. Signals were cut into 60 s windows with a 30 s step, and a window was kept only when over 80% of its samples shared one label, giving 1,187 windows (362 stress). The pulse was band-pass filtered between 0.5 and 4 Hz, peaks were detected, and heart rate, SDNN, RMSSD, pNN50, heart rate range and the LF/HF ratio were computed, together with EDA mean, standard deviation and slope, mean temperature, and the mean, standard deviation and range of acceleration magnitude. Four nested feature sets (Table III) were compared with Random Forest, an RBF-kernel support vector machine (SVM) and XGBoost under leave-one-subject-out (LOSO) validation, and the weakest and strongest configurations were compared with a Wilcoxon signed-rank test on the 15 paired per-participant accuracies.

### C. Session Data, Participants and Label

Each of 100 IT undergraduates installed the application on their own Android phone, signed up with their own account and recorded three sessions between July and September 2026. Some sessions were supervised on campus and the rest were completed independently. Three researcher test accounts used during development were removed before export. Each session begins with a self-report taken before any measurement, so the rating cannot be influenced by a displayed heart rate. It records

current stress, academic context (exam, assignment, normal day or holiday), sleep the previous night in five bands, and caffeinated drinks that day. The participant then types a fixed sentence, records 60 s of PPG and leaves the phone still for a 10 s environmental measurement.

Early builds collected the rating with a slider from 1 to 10, and later builds offered five labelled options stored as 1, 3, 5, 7 and 9. A session is labelled stressed when the rating is 7 or higher, so the model predicts how a participant would rate their own stress, not a verified physiological state. The data was split by participant before modelling into 80 participants (240 sessions, 35.0% stressed) for training and 20 participants (60 sessions, 36.7% stressed) for a held-out test, giving majority-class baselines of 0.650 and 0.633. Ethical clearance was obtained through the Faculty of Computing research ethics procedure, supported by a participant information sheet and a data protection agreement under Sri Lanka's Personal Data Protection Act No. 9 of 2022. Consent is recorded in the application before any data is collected, and no audio, contacts or message content is captured.

### *D. System Architecture and Feature Extraction*

The framework has three tiers (Fig. 1), namely a Flutter application that runs the session and extracts all features on the device, Firebase authentication and a Cloud Firestore database, and a Python tier on the researcher's machine that trains the models and hosts a FastAPI scoring service with a batch scorer for sessions the service misses. Only numeric features leave the phone. Firestore rules let a participant read only their own documents and make readings append-only, and the scoring service repeats the ownership check because it runs with administrator credentials.

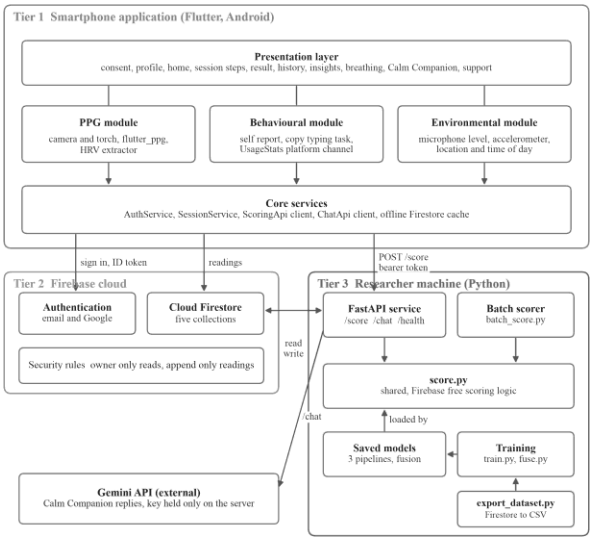



Fig. 1. Three-tier architecture of the framework, showing the smartphone application, the Firebase cloud and the Python scoring and training tier.

For PPG, the rear camera opens at low resolution with the torch on, and the flutter_ppg package extracts the mean red-channel intensity per frame, filters it and detects peaks at 30 frames per second, accepting RR intervals between 300 and 1500 ms with at most 30% change between neighbours. An interval is kept only while the frame rate is stable, no more than 20% of recent intervals have been rejected, signal quality is not poor and at least 150 frames have passed. A reading with fewer than 10 intervals is rejected. Behavioral features come from the typing task (inter-key intervals below 30 ms or above 5 s discarded) and from screen and unlock events over the previous 24 hours, which are omitted rather than recorded as zero when usage access is refused. The microphone level is sampled for 10 s and its trimmed mean stored without recording audio. Table II lists the 24 features.

TABLE II

FEATURES PER MODALITY

| Modality | Features |
|---|---|
| PPG (8) | meanRR, meanHR, SDNN, RMSSD, pNN50, HR range, quality flag, RR count |
| Behavioral (12) | mean inter-key interval, its variance, typing speed, backspace rate, first-key latency, long pauses (>2 s), keystroke count, screen-on time, unlock count, academic context, sleep band, caffeine count |
| Environmental (4) | noise level (dB), accelerometer variance, location (home, library, computer lab, examination hall, other), time-of-day band |

### *E. Machine Learning Pipeline and Evaluation*

Seven columns are removed before training, namely the document, participant and session identifiers, the timestamp, the stress level, the binary label and the free-text note, in which participants often describe their stress. An automated test fails if any of them reaches a feature list, and any modality with cross-validated ROC AUC above 0.98 is flagged for manual review. Each modality is one scikit-learn [24] pipeline with median or most-frequent imputation and one-hot encoding refitted on every training fold. Random Forest (300 trees, balanced class weights) and Gradient Boosting (100 trees, equivalent sample weights) were compared by five-fold StratifiedGroupKFold on the training set, with all three sessions of a participant in the same fold (16 participants per fold), a test confirming that no participant is split, and selection by mean ROC AUC alone.

Late fusion uses a 240 × 3 out-of-fold probability matrix built with identical folds for all modalities. Three rules were compared. The first is a simple average. The second is a weighted average with $w_k = (\mathrm{AUC}_k - 0.5) / \Sigma_j(\mathrm{AUC}_j - 0.5)$. The third is logistic-regression stacking with balanced class weights, chosen in advance for deployment. Headline results use a threshold of 0.5, and an F1-optimal threshold was chosen from the pooled out-of-fold probabilities only. The held-out set was scored once, after every choice was fixed. Accuracy is read against the majority baseline, alongside precision, recall, F1, ROC AUC and average precision [15].

Three integrity diagnostics were run on the collected data before interpretation. Test 1 computes each feature's mean at every stress level, its Spearman correlation with the level and the number of direction reversals. Test 2 removes the rating's effect from each feature and computes the intraclass correlation (ICC) by participant on the residual, since repeated measurements from one person should resemble each other [25]. Test 3 compares the mean distance between a participant's own sessions with that between different participants in standardized feature space. All runs use seed 42 and scikit-learn 1.9.0 with

pinned library versions. Eighty-five automated tests (45 for the pipeline, 40 for the backend) cover integrity rules, leakage, fold separation, fusion, authentication and a replay of the 60 held-out sessions through the deployed scoring code.

## IV. Results and Discussion

### A. Feasibility on WESAD

Table III gives the LOSO results. Heart rate alone reached between 0.702 and 0.729 accuracy with F1 below 0.50 for all classifiers. Adding the other HRV features raised accuracy to between 0.840 and 0.890, and the full multimodal set gave the best accuracy (0.908) and ROC AUC (0.981) with Random Forest and the best F1 (0.834) with XGBoost. The spread across participants also narrowed, from about 0.12 to between 0.056 and 0.075. The Wilcoxon test comparing heart rate only (Random Forest) with the full set (XGBoost) gave a statistic of 0 and $p = 0.0001$, which means no participant was classified less accurately with the multimodal features. These results come from real recordings and agree with the up to 93% binary accuracy reported by the dataset's authors [23]. HRV alone performed far above the 54 to 62% of Izzah *et al.* [19], probably because the Trier test provokes a stronger response than mild cognitive tasks.

TABLE III

LOSO Accuracy on WESAD, Mean (SD) Across 15 Participants

| Feature set | RF | SVM | XGBoost | Best AUC |
|---|---|---|---|---|
| A, heart rate (1) | 0.702 (0.117) | 0.729 (0.125) | 0.725 (0.117) | 0.790 |
| B, full HRV (6) | 0.853 (0.104) | 0.890 (0.073) | 0.840 (0.091) | 0.947 |
| C, B + activity (9) | 0.852 (0.099) | 0.882 (0.073) | 0.860 (0.074) | 0.957 |
| D, C + EDA, temp. (13) | 0.908 (0.056) | 0.887 (0.074) | 0.897 (0.075) | 0.981 |

### B. Pipeline Verification and Classifier Comparison

On the engineering side, every session produced all three modality records, 94.7% of PPG readings passed with the highest quality flag, and a session took about three minutes. All 85 automated tests passed, the deployed service reproduced the offline held-out predictions and metrics exactly for all three fusion rules, and a request for another participant's session was refused with status 403. Of the 300 sessions, 106 (35.3%) were labelled stressed.

Table IV reports cross-validation on the training set. Random Forest had the higher mean ROC AUC, accuracy and F1 on every modality and was selected for all three, but its ROC AUC margin (0.022 for PPG, 0.026 for behavior, 0.017 for environment) was smaller than one standard deviation across folds, so the two candidates should be regarded as close. No modality exceeded the 0.98 review threshold. On the held-out participants (Table V), all three modalities exceeded the 0.633 baseline. PPG reached 0.883 accuracy, F1 0.837 and ROC AUC 0.961, and the behavioral and environmental models each reached 0.817 accuracy.

TABLE IV

Participant-Grouped Cross-Validation, Mean (SD) Over Five Folds

| Modality | Model | Accuracy | F1 | ROC AUC |
|---|---|---|---|---|
| PPG | RF | 0.871 (0.040) | 0.821 (0.056) | 0.938 (0.033) |
| PPG | GB | 0.842 (0.043) | 0.769 (0.071) | 0.916 (0.039) |
| Behavioral | RF | 0.867 (0.019) | 0.800 (0.034) | 0.916 (0.035) |
| Behavioral | GB | 0.829 (0.031) | 0.742 (0.060) | 0.890 (0.042) |
| Environmental | RF | 0.842 (0.073) | 0.780 (0.106) | 0.911 (0.066) |
| Environmental | GB | 0.829 (0.065) | 0.753 (0.123) | 0.894 (0.052) |

### C. Late Fusion

All three fusion rules exceeded the best single modality on accuracy, F1 and ROC AUC (Table V). The two averaging rules gave identical predictions at 0.5, with no false positives and three missed stressed sessions (F1 0.927), while stacking detected all 22 stressed sessions at the cost of four false positives (F1 0.917). The same pattern appeared out of fold on the training data (fusion ROC AUC 0.989 for all rules, with F1 of 0.939, 0.933 and 0.941), which did not depend on the test set. The AUC weights (0.345, 0.329, 0.326) and stacking coefficients (3.210, 3.255 and 3.284, with an intercept of −4.387) were nearly equal, and out-of-fold probabilities correlated between 0.480 and 0.596 (mean 0.544), so averaging cancels part of each model's errors. Tuned thresholds fell between 0.43 and 0.56 and changed test F1 by −0.021 to +0.010, and at those thresholds all three rules produced the same confusion matrix. Because each misclassified session moves accuracy by about 1.7 points, the rules are best read as performing equally well, and the fusion benefit is claimed for this dataset only.

TABLE V

Held-Out Test Performance at Threshold 0.5 (20 Participants, 60 Sessions)

| System | Acc. | Prec. | Recall | F1 | ROC AUC |
|---|---|---|---|---|---|
| PPG | 0.883 | 0.857 | 0.818 | 0.837 | 0.961 |
| Behavioral | 0.817 | 0.762 | 0.727 | 0.744 | 0.889 |
| Environmental | 0.817 | 0.720 | 0.818 | 0.766 | 0.892 |
| Simple average | 0.950 | 1.000 | 0.864 | 0.927 | 0.993 |
| Weighted average | 0.950 | 1.000 | 0.864 | 0.927 | 0.996 |
| Stacking | 0.933 | 0.846 | 1.000 | 0.917 | 0.993 |

The relative fusion gain mirrors that of Aqajari *et al.* [7] without a smartwatch, but the absolute values are far above daily-life cross-user evaluations such as Zhang *et al.* [8], which rarely report F1 above 0.9 for new users. That gap is consistent with the data concerns in Section IV-E.

### D. Feature Importance and Context

In the PPG model, impurity importance was spread across meanRR (0.217), SDNN (0.196), pNN50 and meanHR (0.169 each) and RMSSD (0.141). The behavioral model was led by keystroke timing, with mean inter-key interval (0.160),

backspace rate (0.126) and first-key latency (0.114), while the two phone-usage features did not reach the top five. The environmental model relied on noise (0.397) and movement (0.375). The self-reported context was strongly associated with the label, with 83% of exam-period sessions were labelled stressed against 13% on a normal day, and 71% after under five hours of sleep against 15% after eight hours or more. The movement result needs care, because the phone rests on a surface during the measurement, so higher variance reflects handling of the device, which may indicate restlessness or a less careful protocol. Location also encodes the situation, since examinations take place in examination halls. Impurity importance favors continuous features, so these rankings describe the fitted models, not causal effects.

### *E. Dataset Integrity Diagnostics*

Table VI summarizes the diagnostics. In Test 1, feature means moved in the expected direction across stress levels but with several reversals (Fig. 2). The strongest correlations were SDNN −0.573, accelerometer variance +0.540 and mean inter-key interval +0.394, and accelerometer variance stayed roughly flat from level 1 to 6 before rising at level 7. In Test 2, the rating explained 27 to 37% of the variance of the five main heart rate and HRV features, but after its effect was removed the ICC by participant was 0.031 or lower for every PPG feature, 0.092 or lower for every behavioral feature and 0.031 or lower for both environmental features, far below the 0.5 under which reliability is considered poor [25]. In Test 3, a participant's own sessions were 96.8 to 99.3% as far apart as sessions from different people.

TABLE VI

DATASET INTEGRITY DIAGNOSTICS

| Test | PPG | Behavioral | Environ. |
|---|---|---|---|
| Test 1, strongest Spearman ρ | SDNN −0.573 | meanIki 0.394 | accVar 0.540 |
| Test 1, features with ≤ 1 reversal | 1 of 7 | 0 of 10 | 0 of 2 |
| Test 2, variance explained by rating | 0.011 to 0.367 | 0.054 to 0.189 | 0.301 to 0.326 |
| Test 2, residual participant ICC | 0.000 to 0.031 | 0.000 to 0.092 | 0.015 to 0.031 |
| Test 3, within/between distance ratio | 0.985 | 0.993 | 0.968 |

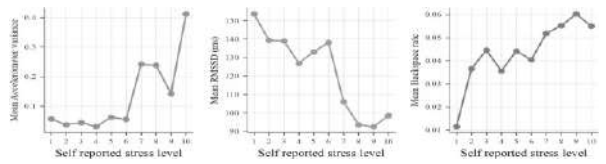



Fig. 2. Mean accelerometer variance, RMSSD and backspace rate at each self-reported stress level.

Two other observations point the same way. Sessions labelled not stressed had a mean RMSSD of 135 ms and pNN50 of 78%, well above values usually reported for healthy young adults at rest [4]. A likely cause is timing noise, since a peak can be misplaced by up to one frame (about 33 ms), which inflates beat-to-beat differences. And 160 of the 300 sessions recorded exactly zero screen time and unlocks, which probably reflects missing usage access recorded as zero in earlier builds or on some devices. Together these results mean the dataset cannot be assumed to reflect genuine, independent, repeated measurement of 100 individuals. Possible explanations include sessions recorded under shared or test accounts, sessions recorded by different people on one account, or values that were generated or altered rather than measured. This study cannot tell them apart and does not assume the most favorable one. Two consequences follow. Grouping folds by participant protects less than intended when individuals are not distinguishable, and the session-data figures show that the framework, pipeline and fusion work end to end on the data collected, not how accurately the system would detect stress in new students. The WESAD results remain the more trustworthy evidence that the chosen features carry stress information across people.

### *F. Limitations*

Beyond dataset validity, five limitations apply. The held-out set holds only 60 sessions from 20 participants, all from one institution and cohort. The label is a single self-report, reduced to two classes and presented differently in early and later builds. No reference ECG was recorded, so the accuracy of camera HRV is unknown. More than half of the sessions lack valid phone-usage values, and no session used the holiday context. Finally, usage statistics exist only on Android, and live scoring currently needs the researcher's machine on a local network over plain HTTP.

## V. CONCLUSION

This paper presented a smartphone-only framework that predicts self-reported stress in IT undergraduates by fusing camera-derived HRV, keystroke and usage behavior, and environmental context. The phone carried the whole workflow, from a session of about three minutes to a returned result, without wearable hardware. On real WESAD recordings, HRV and motion features raised accuracy for unseen people from 0.70 with heart rate alone to 0.91. On the collected sessions, late fusion raised test F1 from 0.837 to between 0.917 and 0.927, but the integrity diagnostics showed that the data lack the individual consistency of real repeated measurements, and camera HRV values were implausibly high, so these figures cannot yet be taken as the accuracy the system would achieve with new students.

The most important next step is therefore a verified, supervised data collection in which every session is recorded by a verified participant with a chest-strap ECG recorded alongside the camera, so that camera HRV can be validated and the diagnostics and participant-grouped evaluation repeated on data of known origin. The pipeline re-runs on such data without code changes. Further work includes more sessions per participant with personal baselines, sub-frame peak interpolation or higher frame rates, recording missing usage as missing, hosted HTTPS scoring, additional classifiers such as SVM, and calibration of the fused probabilities.

## ACKNOWLEDGMENT

The authors thank the Faculty of Computing of NSBM Green University for its support and the students who volunteered for the study sessions.

## REFERENCES


[1] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 104–112, 2020, doi: 10.1080/02673843.2019.1596823.

[2] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investig.*, vol. 15, no. 3, pp. 235–245, 2018, doi: 10.30773/pi.2017.08.17.

[3] G. Taskasaplidis, D. A. Fotiadis, and P. D. Bamidis, "Review of stress detection methods using wearable sensors," *IEEE Access*, vol. 12, pp. 38219–38246, 2024, doi: 10.1109/ACCESS.2024.3373010.

[4] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Front. Public Health*, vol. 5, Art. 258, 2017, doi: 10.3389/fpubh.2017.00258.

[5] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2014, pp. 3–14, doi: 10.1145/2632048.2632054.

[6] A. Choi, A. Ooi, and D. Lottridge, "Digital phenotyping for stress, anxiety, and mild depression: Systematic literature review," *JMIR mHealth uHealth*, vol. 12, Art. e40689, 2024, doi: 10.2196/40689.

[7] S. A. H. Aqajari et al., "Context-aware stress monitoring using wearable and mobile technologies in everyday settings," *medRxiv*, 2023, doi: 10.1101/2023.04.20.23288181.

[8] P. Zhang, G. Jung, J. Alikhanov, U. Ahmed, and U. Lee, "A reproducible stress prediction pipeline with mobile sensor data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 3, Art. 143, 2024, doi: 10.1145/3678578.

[9] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, no. 5, pp. 1–9, 2017, doi: 10.1093/gigascience/gix019.

[10] Y.-Y. Tsai, Y.-J. Chen, Y.-F. Lin, F.-C. Hsiao, C.-H. Hsu, and L.-D. Liao, "Photoplethysmography-based HRV analysis and machine learning for real-time stress quantification in mental health applications," *APL Bioeng.*, vol. 9, no. 2, Art. 026103, 2025, doi: 10.1063/5.0256590.

[11] C. Epp, M. Lippold, and R. L. Mandryk, "Identifying emotional states using keystroke dynamics," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI)*, 2011, pp. 715–724, doi: 10.1145/1978942.1979046.

[12] M. Basner et al., "Auditory and non-auditory effects of noise on health," *Lancet*, vol. 383, no. 9925, pp. 1325–1332, 2014, doi: 10.1016/S0140-6736(13)61613-X.

[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[15] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Art. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[16] M. A. Al-Alim, R. Mubarak, N. M. Salem, and I. Sadek, "A machine-learning approach for stress detection using wearable sensors in free-living environments," *Comput. Biol. Med.*, Art. 108918, 2024, doi: 10.1016/j.compbiomed.2024.108918.

[17] N. Rashid, T. Mortlock, and M. A. Al Faruque, "Stress detection using context-aware sensor fusion from wearable devices," arXiv:2303.08215, 2023.

[18] M. K. Moser, B. Resch, and M. Ehrhart, "An individual-oriented algorithm for stress detection in wearable sensor measurements," *IEEE Sensors J.*, vol. 23, no. 19, 2023, doi: 10.1109/JSEN.2023.3304422.

[19] N. Izzah, A. P. Sutarto, and M. Hariyadi, "Machine learning models for the cognitive stress detection using heart rate variability signals," *J. Tek. Ind.*, vol. 24, no. 2, pp. 83–94, 2022, doi: 10.9744/jti.24.2.83-94.

[20] K. Dahal, B. Bogue-Jimenez, and A. Doblas, "Global stress detection framework combining a reduced set of HRV features and random forest model," *Sensors*, vol. 23, no. 11, Art. 5220, 2023, doi: 10.3390/s23115220.

[21] W. Seo, N. Kim, C. Park, and S.-M. Park, "Deep learning approach for detecting work-related stress using multimodal signals," *IEEE Sensors J.*, vol. 22, no. 12, 2022, doi: 10.1109/JSEN.2022.3170915.

[22] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[23] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interaction (ICMI)*, 2018, pp. 400–408, doi: 10.1145/3242969.3242985.

[24] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[25] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

# Adaptive Confidence-Based Fusion of IoT Sensor Networks and Transformer-Based Visual Detection for Real-Time Fire Identification

Bommera Sharath Chandra
*Vasavi College of Engineering*
*Hyderabad, India*
sharathbommera@gmail.com

Rampalli Shiva Pallavi
*Vasavi College of Engineering*
*Hyderabad, India*
pallavishiva04@gmail.com

G. Rajashekhar
*Vasavi College of Engineering*
*Hyderabad, India*
rajashekhar gattu@staff.vce.ac.in

***Abstract*—Fire detection with a single sensing mode is unreliable: cameras struggle with poor visibility and glare, whereas gas and temperature sensors raise many false alarms. We present a decision-level fusion framework that combines a transformer-based visual detector (RT-DETR-L) with a Random Forest on environmental sensor data. Both outputs are Platt-calibrated, weighted by per-sample reliability estimates (tree-vote entropy for the sensors, detection count and confidence variance for the camera), and passed through a conflict rule, exponential smoothing, and a hysteresis alarm state machine with a stricter single-channel fallback. RT-DETR-L reaches 0.9786 mAP@50 on D-Fire and outperforms YOLOv8m, YOLOv10m and YOLOv11m trained with the same settings. An eight-feature Random Forest reaches an F1-score of 0.9999 on an offline benchmark, a figure we show to be optimistic because of temporal correlation in the data; the deployed node runs a three-feature model. The fusion layer is demonstrated on an ESP32-and-host prototype and on constructed scenarios. It has not been evaluated on synchronised camera and sensor recordings, and the camera calibrator was fitted on synthetic scores, so its effect on false alarms remains to be measured.**

***Keywords—fire detection, multimodal fusion, IoT sensors, RT-DETR, random forest, Platt scaling, adaptive weighting, ESP32***

## I. Introduction

Fire kills more than 180,000 people a year and causes billions of dollars of damage [1]. Smoke detectors go off when toast burns or someone showers [2], yet may stay silent during the smouldering stage of an electrical fire; false alarms teach people to ignore the alarm. Vision-based recognition has advanced quickly with deep learning [3], and low-cost microcontrollers such as the ESP32 allow continuous gas, temperature and humidity monitoring [4]. Neither channel suffices alone: cameras fail in darkness, thick smoke or glare, while gas and temperature readings spike when an oven door opens. Because the two fail for different reasons, a system that decides, for every reading, which branch deserves more trust should outperform either one.

Our contributions are: (1) a comparison of YOLOv8m, YOLOv10m, YOLOv11m and RT-DETR-L on two fire datasets, in which RT-DETR-L is best on both; (2) an offline study of an eight-feature Random Forest against three-feature Platt scaling, entropy-based confidence, exponential smoothing and hysteresis are established techniques and we do not present them as new. The contribution lies in combining them into a per-sample, reliability-weighted decision layer for a camera and a gas-and-climate node, in the controlled detector comparison, and in a working prototype. The fusion layer has not been validated on synchronised recordings (Section VII).

## II. Related Work

*Visual detection.* Early fire detection thresholded pixel colour in HSV or YCbCr and added simple motion cues [9]; these rules broke down under uneven lighting and clutter. Convolutional networks learned features from annotated images and cut false positives on clean benchmarks [10], and YOLO variants have since become the usual real-time choice [21], with YOLOv8 often reported as the best speed–accuracy trade-off, sometimes by a narrow margin [11], [12]. RT-DETR [6], a descendant of DETR [7], uses a hybrid encoder that separates intra-scale interaction from cross-scale fusion, aiming at YOLO-level latency with a global receptive field. Two recent studies adapt it to fire imagery: Yang et al. [26] add attention and feature-pyramid modules for smoke detection in surveillance video, and Tran et al. [27] build a fire and smoke detector for smart-building and IoT settings. Both address the camera alone; we compare RT-DETR-L with YOLO variants on two datasets, including the D-Fire corpus [8], and couple it with a sensor branch.

*IoT-based detection.* Sensor nodes usually combine temperature, humidity, gas and smoke transducers on an ESP32 [4]. Random forests [20] and decision trees dominate classification because they fit in kilobytes of RAM and their decisions can be inspected [5]. A public smoke-detection dataset [15] with eightsensor channels is widely used, although most work exploits only temperature, humidity and one gas channel.

*Multimodal fusion.* Wang et al. [16] and Ding et al. [17] apply Dempster–Shafer evidence theory to temperature, CO and smoke readings, Jia et al. [18] fuse vision with several sensors, and Desikan et al. [19] treat fault-tolerant sensor fusion for fire-risk mitigation. In these works, as far as we can tell, confidence weights are fixed at design time or set by offline calibration rather than recomputed for each sample, and temporal behaviour such as noise and chattering near a threshold receives little attention. Our engine instead estimates reliability per sample from the tree-vote entropy of the forest and the detection consistency of the camera, and stabilises the output with smoothing and hysteresis.

## III. System Architecture

with $p_c = n_c/N_{\text{trees}}$, and map it to a reliability in $(0, 1)$:

$$r_{\text{IoT}} = \frac{1}{1 + \exp\big(\alpha_{\text{ent}}(H_{\text{IoT}} - 0.5)\big)}, \quad \alpha_{\text{ent}} = 2.0. \tag{2}$$

*Vision branch.* Reliability depends on how many boxes were detected, $n_{\text{det}}$ (fire and smoke), and on how consistent their confidences are (variance $\sigma^2_{\text{conf}}$):

$$r_{\text{det}} = \frac{1}{1 + \exp\big(-\alpha_{\text{det}}(n_{\text{det}} - \mu_{\text{det}})\big)}, \tag{3}$$

$$r_{\text{vis}} = r_{\text{det}} \cdot \max\big(1 - \beta_{\text{var}} \min(\sigma^2_{\text{conf}}, 1),\, 0.1\big), \tag{4}$$

with $\mu_{\text{det}} = 1.5$, $\alpha_{\text{det}} = 2.0$ and $\beta_{\text{var}} = 1.0$. The count term $H_{\text{IoT}} = -\sum_{c\in\{0,1\}} p_c \log_2 p_c$ IoT branch. A Random Forest yields a vote distribution, not only a class: 98 of 100 trees voting "fire" is confident, a 55–45 split is ambiguous. We measure the disagreement by the Shannon entropy of the votes, is deliberately conservative: a lone low-confidence box is the $r_{\text{det}} = 1/(1+e) \approx 0.27$ typical signature of a reflection or glare spot, so the camera should not dominate on that evidence. The price is that a single box, which is also what an early fire may produce, gives. so the camera is under-weighted; the conflict rule below keeps the alarm from being suppressed (Fig. 2). Conversely, with no detection $r_{\text{det}} \approx 0.05$ and the sensor branch receives most of the weight, however confident the camera's "no fire" is. Both cases are common: of the 1,281 frames in our live fusion log (two self-test entries excluded), 39% had exactly one detection and 38% none. The form and constants of these terms were not tuned on synchronised data (Section VII).

An ESP32 node senses the chemical and thermal signatures of combustion, a host computer runs RT-DETR on a camera feed, and a fusion layer turns the calibrated outputs of both arms into one fire probability and an alarm state (Fig. 1).

### A. IoT Sensor Node

The node pairs an ESP32 with an AHT10 temperature-humidity module (I2C, GPIO 21/22) and an MQ-2 gas sensor on GPIO 34. A three-feature Random Forest (temperature, humidity, gas; 10 trees, depth 5), compiled to C++ with micromlgen, runs on the node; it was trained on the public dataset of Section V-C, whose TVOC column stands in for the gas reading, and receives the MQ-2 ADC count averaged over five samples. The node cannot measure the other five channels of the dataset. Every second it sends a serial message at 115200 baud, for example T:25.30, H:60.12, G:512, PRED:1, CONF:10. A 30 s watchdog restarts the node if its main loop stalls, and the MQ2 gets 20 s of warm-up.

### B. Vision Subsystem

The camera arm uses RT-DETR-L [6]. To keep latency practical we reduce the input from 640 × 640 to 416 × 416, process every second frame and reuse the previous result in between; FP16 is available on CUDA hardware but disabled on our CPU-only host, where it crashed. Detections above a confidence of 0.25 are kept. The raw fire score s is the highest fire-box confidence plus a spatial-consistency bonus of 0.1 per additional fire box (capped at 0.2, result clipped to 1); with smoke only, s is 0.6 times the highest smoke confidence, and s = 0 without detections. We also record the number of boxes and the variance of their confidences.

### C. Platt Scaling

$$w_k = \frac{w_k^{(0)}\, r_k}{w_{\text{IoT}}^{(0)} r_{\text{IoT}} + w_{\text{vis}}^{(0)} r_{\text{vis}}}, \quad p_{\text{fused}} = \sum_{k\in\{\text{IoT,vis}\}} w_k\, p_k^{\text{cal}}. \tag{5}$$

Neither forest vote fractions nor detector confidences are probabilities [13], [14], [22], so we fit a Platt scaler for each branch:

$$P(\text{fire} \mid s) = \frac{1}{1 + \exp\big(-(As + B)\big)}. \tag{1}$$

A and B come from unregularised logistic regression on a calibration set. For the sensor branch, we reserve a stratified 20% of the 62,630 rows (12,526) and train the forest on the rest, giving A = 16.86, B = −7.66. For the camera branch no labelled set of real RT-DETR scores had been collected when the prototype was built, so the calibrator was fitted on 350 synthetic score–label pairs drawn from assumed Beta distributions (150 fire-like, 150 safe-like, 50 ambiguous), giving A = 12.92, B = −5.41. The camera probabilities are therefore not calibrated against real data (Sections V-D and VII). Calibration quality is measured by the Brier score [23],

$$\tfrac{1}{N}\textstyle\sum_i (p_i - y_i)^2.$$

## IV. Adaptive Fusion Engine

Most multimodal detectors give each input a fixed weight. Our engine instead recomputes, for every sample, how much each branch should count.

### A. Reliability Estimation

### B. Weights and Fusion

The baseline weights w(0) IoT = 0.35 and w(0) vis = 0.65 reflect the empirical finding in our calibration experiment that the camera is generally the more discriminative branch. They are scaled by the reliabilities and renormalised:

### C. Conflict Handling

$$|p_{\text{IoT}}^{\text{cal}} - p_{\text{vis}}^{\text{cal}}| > \tau_{\text{conflict}} = 0.50, \tag{6}$$

If one branch reports high fire probability and the other none, a weighted average lands in a middle that has no physical meaning. We therefore declare a conflict when

and then apply a safety-first rule: if either branch is confident, its score, slightly discounted, replaces the fused score; otherwise neither branch is trusted, and the fused score is damped:

$$p_{\text{conflict}} = \begin{cases} \kappa \max(p_{\text{IoT}}^{\text{cal}}, p_{\text{vis}}^{\text{cal}}) & \text{if } \max(\cdot) > \rho, \\ \gamma\, p_{\text{fused}} & \text{otherwise,} \end{cases} \tag{7}$$

with $\rho = 0.8$, $\kappa = 0.9$, $\gamma = 0.7$; in the second case the frame is also flagged UNCERTAIN. The rule is asymmetric by

Fig. 1. System architecture. Each branch is processed and calibrated independently, and the fusion engine weights, reconciles, smooths and thresholds the two probabilities.

design: it accepts more false alarms for fewer missed fires. We have not measured the false-alarm rate it induces; the

evidence is limited to the constructed scenarios of Table V and the live example of Fig. 2.

### *D. Temporal Smoothing and Alarm State Machine*

$$\text{State}_t = \begin{cases} \text{ALARM} & \text{if } \bar{p}_t \geq \theta_{\text{high}} = 0.70, \\ \text{SAFE} & \text{if } \bar{p}_t \leq \theta_{\text{low}} = 0.40, \\ \text{State}_{t-1} & \text{otherwise.} \end{cases} \quad (8)$$

Real readings are noisy, so the fused score is smoothed by an exponential moving average, $\bar{p}_t = \alpha\, p_{\text{fused},t} + (1 - \alpha)\bar{p}_{t-1}$ with $\alpha = 0.3$ (values below 0.2 felt sluggish and above 0.4 let too much noise through). To avoid chattering around one threshold, we use two:

The implementation adds a third state, UNCERTAIN, for out-put that should be reviewed manually: it is entered from SAFE when the conflict condition of Eq. (6) holds and $\bar{p}_t > \theta_{\text{low}}$, and left for ALARM when $\bar{p}_t \geq \theta_{\text{high}}$ or for SAFE when $\bar{p}_t < \theta_{\text{low}}$. Eq. (8) describes the SAFE–ALARM transitions.

### *E. Single-Modality Fallback and Parameters*

A branch counts as lost when no sensor packet has arrived for 5 s or no camera frame can be read. The remaining branch is then smoothed and thresholded in the same way, with $\theta_{\text{high}}$ raised to 0.80 (sensor only) or 0.75 (camera only) to compensate for the missing cross-modal confirmation. Table I lists every constant. The entropy slope and EMA constant were chosen from a few candidate values, and the baseline weights from the calibration experiment; the remaining values were set by hand and were not tuned on labelled, synchronised data.

## V. Experiments And Results

### *A. Datasets and Training*

D-Fire [8] is a fire corpus with bounding-box annotations (21,527 images at 416 × 416 in the original release); we used the single-class (fire) YOLOv8-format export on Roboflow Universe [24], which has no predefined validation split, so a subset of the training set served for validation. Fire-SmokeIndoor (FSI) is a two-class indoor collection with predefined training, validation and test splits (Roboflow export [25]). Smoke Detection IoT [15] has 62,630 sensor rows with eight channels (temperature, humidity, TVOC, eCO2, raw H2, raw ethanol, pressure, PM2.5) and a binary alarm label. Its rows appear to come from continuous event sequences, so temporal autocorrelation inflates metrics under stratified k-fold validation. We report the stratified results as they are and treat them as upper bounds; an event-wise split would be more conservative and was not run.

All four detectors were trained with the Ultralytics frame work from pretrained weights with the same settings: 50 epochs, SGD, input size 416, batch size 8, initial learning rate 0.01, mixed precision and default augmentation (mosaic, random erasing, flips). Each detector was trained once, and the tables report single-run values without confidence intervals.

### *B. Visual Model Comparison*

RT-DETR-L ranks first on both datasets (Table II). Against YOLOv8m, the best YOLO variant, it gains 0.4 mAP@50 points on D-Fire and 3.2 points on FSI, at a cost in latency (38.8 ms vs. 26.5 ms per frame). The “ms” columns are perimage inference times from the validation runs; their hardware was not recorded, so they support only a relative comparison and are not comparable with the 731 ms measured on the CPUonly host of Section VI. We did not expect the gap on FSI. Fire changes shape between frames and smoke is translucent with soft edges, which suits a grid of local receptive fields poorly. On FSI the YOLO models often returned nothing on frames where thin smoke had spread across the ceiling, while RTDETR flagged them. A plausible reading is that self-attention correlates colour shifts in one part of the frame with haze in another; on D-Fire, where fires are more distinct and localised, the gap is smaller. These observations are qualitative, and we ran no ablation isolating self-attention, so the explanation remains a hypothesis.

### *C. IoT Sensor Model Comparison*

Table III is an offline benchmark on a public dataset. The three-feature baselines are small models (a depth-5 decision tree, a 10-tree forest of depth 5 and a linear SVM) trained on a single random 80/20 split. The eight-feature forest has 100 trees of depth 12 and was configured for stratified five-fold cross-validation; the partition behind the reported row was not recorded separately. The comparison is therefore not like for like, and the gain cannot be attributed to the extra channels alone. The eight-feature model needs channels our node does not measure (Section VI), so its scores describe the benchmark, not the running prototype. Because adjacent rows come from the same events, the near-perfect metrics (F1 = 0.9999) are upper bounds. Gini importance ranks TVOC first (27.7%) and atmospheric pressure second (19.8%). Volatile organic compounds are among the first products of thermal decomposition; for pressure we can only hypothesise that even a small fire creates convection that shifts local pressure by a fraction of a hectopascal, and we ran no controlled burn to test it.

TABLE I
PARAMETERS OF THE FUSION PIPELINE

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $\alpha_{\text{ent}}$, entropy slope (2) | 2.0 | $\tau_{\text{conflict}}$ (6) | 0.50 | Platt $(A, B)$, sensor (1) | 16.86, −7.60 |
| $\alpha_{\text{fire}}$, fuser (2) | 1.5, 2.0 | $p_0$, $w_s$, $\gamma$ (7) | 0.8, 0.9, 0.7 | Platt $(A, B)$, camera (synthetic) | 12.92, −5.41 |
| $\beta_{\text{ema}}$ (4) | 1.0 | $\alpha$, EMA | 0.3 | Camera box threshold | 0.25 |
| $\theta_{\text{high}}^{(s)}$, $\theta_{\text{high}}^{(c)}$ (5) | 0.35, 0.65 | $\theta_{\text{high}}$, $\theta_{\text{low}}$ (8) | 0.70, 0.40 | Frame skipping | every 2nd |
| Trees, depth (offline 8-feature RF) | 100, 12 | Fallback $\theta_{\text{high}}$ (IoT, vision) | 0.80, 0.75 | Platt hold-out share (sensor) | 20% |

TABLE II
DETECTION PERFORMANCE ON D-FIRE AND FIRE-SMOKE-INDOOR (FSI)

| | D-Fire | | | | | FSI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Prec. | Rec. | F1 | mAP@50 | ms | Prec. | Rec. | F1 | mAP@50 | ms |
| YOLOv8m | 0.9496 | 0.9256 | 0.9380 | 0.9751 | 26.47 | 0.6717 | 0.6423 | 0.6566 | 0.6929 | 26.34 |
| YOLOv10m | 0.8716 | 0.8588 | 0.8651 | 0.9323 | 25.99 | 0.6458 | 0.6282 | 0.6369 | 0.6355 | 25.97 |
| YOLOv11m | 0.8879 | 0.8805 | 0.8779 | 0.9333 | 31.81 | 0.6620 | 0.6257 | 0.6454 | 0.6561 | 31.91 |
| **RT-DETR-L** | **0.9660** | **0.9613** | **0.9637** | **0.9786** | **38.79** | **0.7374** | **0.7137** | **0.7246** | **0.7246** | **37.14** |

TABLE III
CLASSIFIER COMPARISON ON THE SMOKE DETECTION IOT DATASET

| Model | Feat. | Acc. | Prec. | Sens. | F1 | Kappa |
|---|---|---|---|---|---|---|
| Decision Tree | 3 | 0.9610 | 0.9727 | 0.9727 | 0.9727 | 0.9048 |
| Random Forest | 3 | 0.9630 | 0.9804 | 0.9674 | 0.9739 | 0.9104 |
| SVM | 3 | 0.7472 | 0.7462 | 0.9782 | 0.8466 | 0.1966 |
| **Enhanced RF** | **8** | **0.9999** | **1.0000** | **0.9999** | **0.9999** | **0.9998** |

TABLE IV
BRIER SCORES BEFORE AND AFTER PLATT SCALING (LOWER IS BETTER, IN-SAMPLE)

| Branch (calibration data) | Before | After |
|---|---|---|
| IoT, Random Forest (hold-out) | 0.0002 | 0.0001 |
| Vision, RT-DETR (synthetic scores) | 0.0694 | 0.0549 |

### *D. Calibration Results*

Brier scores before and after Platt scaling are 0.0002 and 0.0001 for the sensor branch (12,526-row hold-out set) and 0.0694 and 0.0549 for the camera branch (350 synthetic pairs). Both are in-sample, i.e. computed on the data that fitted the calibrator (Table IV). The sensor head hardly moved: it was already close to calibrated on a benchmark where a classifier is almost error-free. The camera result shows only that the fitting procedure works; because the scores are synthetic, it is no evidence that RT-DETR is over- or under-confident on real

images. The learned map is also steep: a raw box confidence of 0.71 becomes 0.98 (Fig. 2), so the camera probabilities in the prototype are more confident than the raw scores.

*E. Fusion System Behaviour*

Table V lists the outputs of the implemented engine for constructed first-frame inputs (smoothing starts at the input, the IoT entropy is the binary entropy of $p_{IoT}$, and the detection variance is 0.01 for two boxes and 0 otherwise). The inputs are constructed, not recorded.

In the cooking scenario (oily stir-frying) the gas sensor gives $p_{IoT} = 0.30$ and the camera sees nothing, $p_{vis} = 0.02$. The fused score is 0.24, well below the alarm threshold. The camera's reliability is only about 0.05, so the sensor branch receives 78% of the weight and the camera's low score pulls the result down less than a fixed-weight average would (0.12 with the baseline weights). In the two conflict rows the safety-first rule replaces the fused score by 0.9 times the more alarming branch score (0.81 and 0.79), so a real fire signal is not averaged away. This table illustrates the decision logic and is not an evaluation: it yields no detection or false-alarm rate and cannot show that adaptive weighting beats fixed weights, single-modality operation or D-S and AND/OR combination. Calibration is quantified, and the entropy slope and EMA constant were chosen from a few candidates, but adaptive weighting, conflict handling, hysteresis and the fallback have not been ablated.

TABLE V
FUSION ENGINE OUTPUT FOR CONSTRUCTED SINGLE-FRAME INPUTS (ILLUSTRATIVE, NOT MEASURED)

| Scenario | $p$IoT | $p$vis | $n$det | $p$fused | State |
|---|---|---|---|---|---|
| Both detect fire | 0.95 | 0.92 | 2 | 0.93 | ALARM |
| Sensor only (no camera) | 0.88 | — | — | 0.88 | ALARM |
| Camera only (no IoT) | — | 0.82 | 2 | 0.82 | ALARM |
| Conflict (IoT high) | 0.90 | 0.15 | 1 | 0.81 | ALARM |
| Conflict (vision high) | 0.10 | 0.88 | 1 | 0.79 | ALARM |
| Both safe | 0.05 | 0.03 | 0 | 0.05 | SAFE |
| Cooking false alarm | 0.30 | 0.02 | 0 | 0.24 | SAFE |

## VI. PROTOTYPE IMPLEMENTATION

Hardware. The node is an ESP32 (dual-core Xtensa LX6, 240 MHz, 520 KB SRAM, 12-bit ADC) with the AHT10 and MQ-2 sensors of Section III. The host is a laptop with an Intel Core i5-1335U, 16 GB of RAM and integrated Intel Iris Xe graphics, no discrete GPU, running Windows 11, Python 3.10 and a CPU build of PyTorch 2; since the processor lacks AVX-512, tensor operations are restricted to AVX2 to avoid illegal-instruction faults. A 720p, 30 fps USB webcam is the video source, and the node is connected over USB serial at 115200 baud.

Host-side sensor model. The node transmits only T , H and G. On the host, G is placed in the TVOC input of the eight-feature forest and the other five inputs are fixed defaults (eCO2 400 ppm, raw H2 12000, raw ethanol 18000, pressure 1013 hPa, PM2.5 0). The forest was trained on public-dataset readings, not MQ-2 counts, so the IoT probability in the prototype is a functional stand-in whose absolute values have no physical meaning. The sensor Platt calibrator is fitted offline but is not loaded at runtime, so the fusion consumes the raw forest probability.

Throughput. The node emits one packet per second. The fusion log records the main loop at 1.5 and 2.0 iterations per second in two dual-modality sessions (median period 0.51 s and 0.41 s) and 0.6 per second in a vision-only session (1.46 s). Sensor-to-alarm latency is thus at least one loop period, and smoothing and hysteresis add a few seconds of reaction delay by design. RT-DETR needs 731 ms per processed frame on this host (Fig. 2), so we do not claim real-time operation here. Deployment choices. Three choices reduce inference cost: processing every second frame, a 416 × 416 input instead of 640 × 640 (about 40% lower per-frame cost), and FP16 on CUDA hardware, which we keep disabled on our CPU host. In addition, when the on-device forest predicts fire, the firmware raises a GPIO pin for one second (five-second cooldown) intended to wake a low-power host such as a Raspberry Pi. Our desktop-class host does not use this signal, so the wake-up path and any energy saving were neither exercised nor measured.

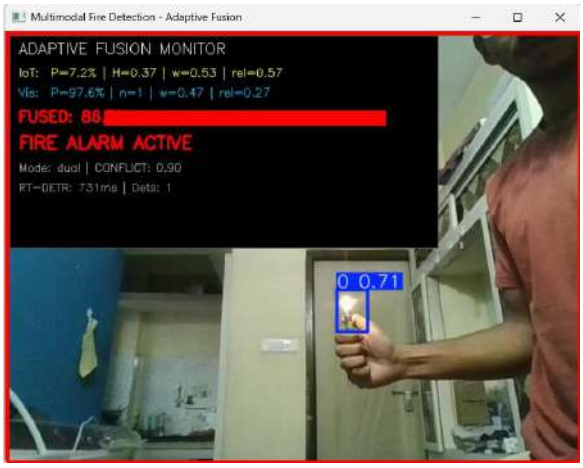



Fig. 2. Live prototype with a lighter flame as the test source. The camera reports $P_{vis} = 97.6\%$ from a single detection ($n = 1$, reliability 0.27; raw box confidence 0.71, mapped by the synthetic-data Platt scaler), the sensor branch 7.2%. The safety-first rule replaces the fused score by 0.9 times the camera score; after smoothing the fused score is 86% and the alarm is raised. RT-DETR inference takes 731 ms.

## VII. DISCUSSION AND LIMITATION

The biggest limitation is that the sensor and visual data come from separate datasets; no public corpus records camera frames and sensor readings together during the same fires, so the fusion engine has never been evaluated end-to-end on co-occurring data, and Table V uses constructed inputs. A synchronised dataset is the most important next step. Further limitations follow. First, the eight-feature model is an offline benchmark: the node measures three channels, and even the benchmark scores are upper bounds because of temporal leakage, which we have not countered with an event-wise split. Second, we did not compare the fused output with single-modality, fixed-weight, D-S or AND/OR baselines or ablate its stages, so the benefit of the adaptive design is not established. Third, the vision reliability term penalises a single detection, typical of an early fire, and the conflict rule trades false alarms for sensitivity; neither effect is quantified. Fourth, the 731 ms latency on the

deployed host far exceeds the figures of Table II, and RT-DETR-L is larger than the medium YOLO variants, so the speed comparison is not like for like. Fifth, in the running prototype the camera calibrator was fitted on synthetic scores and the sensor calibrator is not applied, so neither branch is calibrated on real data there, and the host-side forest receives five fixed inputs and an MQ-2 count in place of TVOC. Finally, the prototype was tested only in a controlled laboratory; kitchens, factory floors and stairwells remain untested.

Planned work includes recording a synchronised multi-modal dataset, refitting the camera calibrator on real detector scores, an event-wise evaluation of the sensor model, adding a BME688 module to measure more channels at the edge, and moving inference to a GPU edge device such as an NVIDIA Jetson.

## VIII. Conclusion

We combined Platt-calibrated probabilities, per-sample reliability weights, a conflict rule, EMA smoothing and a hysteresis state machine into a fusion layer for a camera and a sensor node, with a stricter fallback when one branch is lost. RT-DETR-L outperformed three YOLO variants on D-Fire (mAP@50 0.9786) and on FSI (0.7246), most clearly on cluttered indoor scenes. The eight-feature forest improves clearly on the three-feature baseline, but its near-perfect met-rics describe an offline benchmark with temporal correlation, not the deployed node. On an integrated-graphics host the vision branch is far from real time (731 ms per frame). The main gap in our evidence is the lack of synchronised multimodal recordings, without which the end-to-end benefit of the fusion engine remains an open question; building that dataset is our next step.

## References


[1] World Health Organization, “Burns—Key facts,” *WHO Fact Sheet*, Oct. 17, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/burns

[2] W. Yang, Y. Wu, and S. K. K. Chow, “Deep learning method for real-time fire detection system for urban fire monitoring and control,” *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 216, 2024, doi: 10.1007/s44196-024-00592-8.

[3] G. Cheng, X. Chen, C. Wang, X. Li, B. Xian, and H. Yu, “Visual fire detection using deep learning: A survey,” *Neurocomputing*, vol. 596, p. 127975, 2024, doi: 10.1016/j.neucom.2024.127975.

[4] Y. Subbarayudu, G. V. Reddy, J. Bhargavi, and K. Latha, “An efficient IoT-based novel approach for fire detection through ESP32 microcontroller in forest areas,” in *MATEC Web Conf.*, vol. 392, p. 01109, 2024, doi: 10.1051/matecconf/202439201109.

[5] S. K. Jayasingh, S. Swain, K. J. Patra, et al., “An experimental approach to detect forest fire using machine learning mathematical models and IoT,” *SN Comput. Sci.*, vol. 5, no. 1, p. 148, 2024, doi: 10.1007/s42979-023-02514-5.

[6] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, “DETRs beat YOLOs on real-time object detection,” 2024, arXiv:2304.08069 [cs.CV].

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” 2020, arXiv:2005.12872 [cs.CV].

[8] P. V. A. B. de Venaˆncio, A. C. Lisboa, and A. V. Barbosa, “An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices,” *Neural Comput. Appl.*, vol. 34, no. 18, pp. 15349–15368, 2022, doi: 10.1007/s00521-022-07467-z.

[9] T. C¸ elik and H. Demirel, “Fire detection in video sequences using a generic color model,” *Fire Safety J.*, vol. 44, no. 2, pp. 147–158, 2009, doi: 10.1016/j.firesaf.2008.05.005.

[10] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, “Convolutional neural networks based fire detection in surveillance videos,” *IEEE Access*, vol. 6, pp. 18174–18183, 2018, doi: 10.1109/ACCESS.2018.2812835.

[11] Catargiu, N. Cleju, and I. B. Ciocoiu, “A comparative per-formance evaluation of YOLO-type detectors on a new open fire and smoke dataset,” *Sensors*, vol. 24, no. 17, p. 5597, 2024, doi: 10.3390/s24175597.

[12] E. H. Alkhammash, “A comparative analysis of YOLOv9, YOLOv10, YOLOv11 for smoke and fire detection,” *Fire*, vol. 8, no. 1, p. 26, 2025, doi: 10.3390/fire8010026.

[13] Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[14] S. Dincer, “Smoke detection dataset,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset

[15] H. Wang, Y. Shi, L. Chen, and X. Zhang, “A tunnel fire detection method based on an improved Dempster-Shafer evidence theory,” *Sensors*, vol. 24, no. 19, p. 6455, 2024, doi: 10.3390/s24196455.

[16] Q. Ding, Z. Peng, T. Liu, and Q. Tong, “Multi-sensor building fire alarm system with information fusion technology based on D-S ev-idence theory,” *Algorithms*, vol. 7, no. 4, pp. 523–537, 2014, doi: 10.3390/a7040523.

[17] Y. Jia, X. Cui, B. Wang, Z. Mao, and Q. Zhang, “A fire detec-tion model based on vision and multi-sensor fusion,” in *Proc. 7th Int. Conf. Artif. Intell. Pattern Recognit. (AIPR)*, 2025, pp. 1–8, doi: 10.1145/3703935.3704042.

[18] J. Desikan, S. K. Singh, A. Jayanthiladevi, S. Bhushan, V. Rishiwal, and M. Kumar, “Hybrid machine learning-based fault-tolerant sensor data fusion and anomaly detection for fire risk mitigation in IIoT environment,” *Sensors*, vol. 25, no. 7, p. 2146, 2025, doi: 10.3390/s25072146.

[19] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[20] L. T. Ramos, E. Casas, C. Romero, F. Rivas-Echeverr´ıa, and E. Bendek, “A study of YOLO architectures for wildfire and smoke detection in ground and aerial imagery,” *Results Eng.*, vol. 26, p. 104869, 2025, doi: 10.1016/j.rineng.2025.104869.

[21] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1321–1330.

[22] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.

[23] R. D. Rajasinghe and D. Wijesinghe, "A Federated Multimodal Deepfake Detection Framework for Enhancing the Integrity of Video Evidence in Court Systems," 2026 9th International Conference on Electronics Technology (ICET), Chengdu, China, 2026, pp. 806-811, doi: 10.1109/ICET69987.2026.11659066.

[24] Roboflow Universe, “Fire-Smoke-Indoor,” project fire-smoke-indoor, version 5, workspace object-detection-7qn6l. [Online]. Available: https://universe.roboflow.com/object-detection-7qn6l/fire-smoke-indoor

[25] L. Yang, Y. Cheng, F. Xu, B. Li and X. Li, “Real-time smoke detection in surveillance videos using an enhanced RT-DETR framework with triplet attention and HS-FPN,” *Fire*, vol. 7, no. 11, art. 387, 2024, doi: 10.3390/fire7110387.

[26] P.-T.-T. Tran, N. P. Nguyen, K. A. Nguyen, N. H. Pham and H. T. Nguyen, “RT-DETR for intelligent fire detection: An applied framework for smart building and IoT systems,” *Appl. Comput. Syst.*, vol. 31, no. 1, pp. 138–154, 2026, doi: 10.2478/acss-2026-0012.

# CALM-Hallu: Calibration-Aware Claim-Level Hallucination Detection under Model, Language, and Decoding Shifts

Rahul Kumar
*Research Scholar School of Artificial Intelligence*
*Bennett University*
*Greater Noida, Uttar Pradesh*
*India*
mehrotrarahul55@gmail.com

Tapsi Nagpal
*Professor CSE Department*
*Rawal Institute of Engineering and Technology*
*Faridabad, India*
tapsiarora@gmail.com

Jatin Singh Rajawat
*Data Analyst & Power Platform Developer*
*Metaval Engineering Private Limited*
jatinsr310304@gmail.com

***Abstract*—Large language models produce factual errors that may be difficult to detect reliably under changes in generator, language or decoding policy. We introduce CALM-Hallu, a claim-level evaluation framework combining static uncertainty summaries, token-trajectory dynamics, source-only probability calibration and confidence-based abstention. On the released multilingual MUCH benchmark, combined XGBoost with Platt calibration reaches 0.7647 AUROC and 0.6110 AUPRC on 19,884 in-distribution claims, outperforming four simple token-uncertainty rankings. A length-controlled ablation establishes that dynamics improve AUROC by 0.0165 (95% prompt-bootstrap confidence interval [0.0111, 0.0221]) beyond static features and claim length. Relative dynamics deteriorate less under decoding-temperature transfer, especially 1.0 to 0.7, but this advantage is direction-dependent. Under held-out generators and joint generator-language shifts, expected calibration error increases by 0.1237 and 0.1495, even though AUROC remains stable or increases on average. Evaluation on 842 human-annotator-agreement claims and selective prediction further shows that an 80% coverage policy lowers accepted-set risk across evaluated shifts. These findings motivate joint reporting of discrimination, probability calibration and selective risk for claim-level hallucination detection.**

***Keywords—Index Terms—Hallucination detection, uncertainty estimation, calibration, deployment shift, selective prediction.***

## I. Introduction

Large language models can produce plausible but unsupported factual statements. An operational hallucination detector should distinguish unreliable claims and provide interpretable confidence, not merely rank samples under familiar conditions. Recent claim-level benchmarks enable controlled evaluation using cached generation scores [1]. Sampling-based self-consistency [2], semantic entropy [3] and claim-conditioned probability (CCP) [4] provide complementary viewpoints, but may require additional generation, semantic comparisons or richer access than a compact cached-token detector.

CALM-Hallu investigates whether static token uncertainty and within-claim uncertainty dynamics remain useful after a generator, language or decoding change. Its contribution is an integrated, group-disjoint evaluation of (i) dynamic-feature complementarity controlling for length, (ii) shift-direction asymmetry, (iii) calibration drift despite preserved AUROC, and (iv) abstention across automated and human annotation views. We use fixed source-only calibration and do not adapt to target-domain labels.

AUROC answers whether hallucinated claims are ranked higher than factual ones, whereas a calibrated probability should approximately track empirical error frequency. A change in class prevalence, generator behavior, prompt composition or decoding strategy can preserve pairwise ranking while distorting probability magnitudes. This distinction motivates our reporting of AUROC together with Brier and ECE, rather than relying on a single discrimination number.

## II. Related Work And Research Positioning

MUCH provides multilingual claim annotations and top-24 token-generation information for reproducible white-box uncertainty evaluation [1]. SelfCheckGPT [2] and semantic entropy [3] assess generation consistency or meaning-level uncertainty. CCP estimates uncertainty conditioned on the claim content [4], whereas Shifting Attention to Relevance (SAR) reweights semantically important tokens [5]. LM-Polygraph [6] systematizes uncertainty baselines. Our evaluated scalar scores are deliberately simpler and should not be interpreted as reproducing CCP, SAR or sampling-based detectors. Calibration methods [7] and studies of uncertainty under dataset shift [8] establish why good discrimination does not ensure reliable probability estimates. Selective classification explicitly trades coverage for risk [9].

The objective differs from retrieval-augmented fact checking. The detector does not access ground-truth documents at inference and cannot verify an isolated proposition by looking it up. Instead it uses the generation-time probability record to estimate hallucination risk. Compared with sampling-based approaches it avoids new generation calls, but it cannot be used when token-probability metadata is unavailable. This information-budget distinction is crucial when deciding which baselines are directly reproducible from the released MUCH fields.

## III. Data And Method

### A. Benchmark and group-disjoint protocol

We use the public orailix/MUCH release: 4,673 automatically labeled source responses and 200 official human-test responses, totaling 4,873 in the analyzed release; the benchmark publication reports 4,876 samples [1]. Excluding trailing auxiliary EOS-only segments gives 19,884 source claims and 867 official test claims. The release spans English, French, German and Spanish, four generators, and decoding temperatures 0.7 and 1.0. Group identifiers tie together claims from the same normalized question within language. Ten disjoint group folds are assigned to target, calibration, validation and the remaining seven training folds, with no prompt shared between roles.

The response-level corpus contains multiple claims per generation; those claims are statistically dependent. Splitting rows at random would create train/test overlap because several claims could originate from the same user question and generator. We therefore form a prompt-level group before assigning folds. Across E1, target question groups are disjoint and every claim receives one out-of-fold target prediction. Source-only calibration then uses a calibration role distinct from both the development selection and the test role. For the official human-test analysis, source questions overlapping the human evaluation questions are removed before forming the training cohort. This removes a second, subtler pathway through which familiar prompts could bias the human-label evaluation.

### B. Features and calibrated detector

For each claim, ten static statistics summarize sampled-token probabilities (mean, minimum, maximum and standard deviation), top-24 entropy (mean, maximum and standard deviation), probability margins (mean and minimum), and mean token negative log-likelihood. The second family contains entropy and confidence slopes, maximum entropy increase and confidence decrease, normalized entropy/confidence total variation, position of maximum uncertainty, entropy second-half contrast, number of relative entropy peaks, and a claim-length covariate. Thus the original 20-feature representation consists of 10 static statistics, nine trajectory attributes and one length feature. Robust within-claim normalization reduces sensitivity to nearly constant entropy trajectories.

XGBoost [10] selects maximum tree depth, learning rate and class-weighting mode using the source-only validation AUCPR criterion; early stopping determines the number of rounds. The selected model is refitted on source-training plus validation claims. A separate source calibration fold fits logistic (Platt) calibration to XGBoost margins m: $p(y=1 \mid m)=\mathrm{sigmoid}(am+b)$, and the fitted parameters remain fixed on all targets. We evaluate raw, Platt and isotonic probability outputs in E1 and use Platt for the primary shifts. LR, MLP and GRU are additional E1 discriminators, under the same target partitions.

The primary comparison is not merely a change in model type. For each E1 run, three folds are reserved: target fold r, source-calibration fold (r+1) mod 10 and validation fold (r+2) mod 10. The other seven folds provide training examples. XGBoost searches three tree depths (3, 5, 7), two learning rates (0.03 and 0.10), and two class-weighting modes. Validation AUPRC selects among twelve configurations; training uses at most 1,500 rounds with a 75-round early-stopping patience. The final model is refitted with its selected number of rounds on the source and validation partitions only. This arrangement fixes tuning decisions independently of the calibration and target annotations. Human-evaluation folds reuse the source-only fitting rule with additional cross-partition question exclusion. As an exploratory neural comparator, we also tested a fixed two-layer, four-head Transformer (64-dimensional states) over four token-probability channels, gated with an MLP of 19 engineered features excluding claim length. It uses the same E1 target

A claim of n token positions is associated with sampled-token probabilities $p_1,\ldots,p_n$, truncated distribution entropies $h_1,\ldots,h_n$, and token-margin values. Static p_mean and p_min summarize token confidence level; nll_mean computes the mean of $-\log(p_i)$ over positions. Dynamics summarize within-claim changes rather than only magnitude. For example, confidence slope fits a linear trend over normalized position $i/(n-1)$, maximum drop measures the largest consecutive decrease, and normalized total variation averages successive absolute changes. Entropy slopes, peaks and half-split contrasts are computed after robust within-claim scaling. Claim length is included as a tenth auxiliary covariate in the originally named "relative" family; hence the controlled 9-/11-/19-feature ablations are necessary to isolate the role of dynamics.

### C. Shift experiments and metrics

E1 is ten-fold in-domain evaluation. E2 holds out each generator and compares reference with unseen-generator claims. E3 holds out each language. E4 compares source/target temperature transfers 0.7→1.0 and 1.0→0.7 on shared question-generator combinations. E5 holds out each generator-language pair, forming reference, generator-only, language-only and joint-shift evaluations. E7 trains only on automated source claims excluding questions shared with the official human test, and then evaluates 867 human-test claims or the 842-claim annotator-agreement subset.

Metrics include AUROC, AUPRC, prevalence-normalized average precision, Brier score and ten-bin expected calibration error (ECE). Source-only calibration is evaluated on target claims without recalibration. Confidence for selective rejection is $|p-0.5|$; risk is the error rate on the highest-confidence retained claims. Claim-group-cluster bootstrap intervals use 2,000 replicates for E1–E5 and focused ablations and 5,000 for E7. Metrics are calculated within frozen evaluation cells and averaged across cells. Intervals are nominal, without multiple-testing correction.

E2 uses leave-one-generator-out transfer. E3 uses leave-one-language-out transfer, which does not automatically produce paired questions across languages because the group definition includes language. E4 compares matched question-generator combinations across both directions of decoding-temperature transfer. E5 holds out a generator-language pair and evaluates four domain types; the domains need not have identical prevalence or claim-length distributions. Thus shift-minus-reference differences are empirical contrasts for these cohorts, not causal treatment effects. E7 holds the official human target fixed while varying source training/calibration rotations. The fixed human subset is not ten independent experiments; its

prompt-cluster bootstrap samples official target questions jointly.

For probability calibration, ECE is computed with ten fixed-width probability bins, summing bin mass multiplied by absolute empirical-label minus mean-confidence discrepancy. Brier is the average squared probability error. Because ECE is non-smooth in the binning and the risk coverage operator is a sample quantile, their bootstrap means need not match point estimates exactly. Percentile confidence intervals are formed at the group-cluster level, and only prespecified paired contrasts should be interpreted as within-claim comparisons. The evaluated selective classifier orders claims by distance of predicted hallucination probability from 0.5 and accepts the most confident 80%. This is an observed-sample operating characteristic, not a proved risk bound under future shifts.

## IV. Experimental Results

### *A. Comparison with existing classifiers and scalar uncertainty*

TABLE I
E1 DISCRIMINATION COMPARISONS (HIGHER IS BETTER)

| Method | AUROC | AUPRC | Brier |
|---|---|---|---|
| LR + Platt | 0.7453 | 0.5944 | 0.1675 |
| MLP + temperature | 0.7510 | 0.5999 | 0.1661 |
| GRU + temperature | 0.7454 | 0.6022 | 0.1660 |
| Transformer fusion + Platt | 0.7526 | 0.6030 | 0.1650 |
| XGBoost + Platt | 0.7647 | 0.6110 | 0.1632 |
| Mean token NLL | 0.6876 | 0.4758 | — |
| Top-24 entropy | 0.6783 | 0.4711 | — |
| Negative min. token p | 0.7073 | 0.5282 | — |
| Negative mean token p | 0.6754 | 0.4451 | — |

All methods share the E1 target claims and group folds; classifier numbers are descriptive averages, not paired significance tests across architectures. Among four label-free scalar rankings, negative minimum token probability is strongest (0.7073 AUROC); the combined detector improves on it by 0.0574 AUROC (paired 95% CI [0.0486, 0.0666]) and 0.0828 AUPRC ([0.0717, 0.0956]). Simple scalar scores are not calibrated probabilities, so Brier/ECE is not reported for them. A Transformer–gated fusion comparator obtained 0.7526 AUROC, 0.6030 AUPRC and 0.1650 Brier. Its paired difference against XGBoost was −0.0121 AUROC (95% CI [−0.0174, −0.0072]) and −0.0080 AUPRC ([−0.0146, −0.0015]), favoring XGBoost. Hybrid ECE (0.0305) and XGBoost ECE (0.0308) did not differ significantly (paired CI spans zero).

We additionally checked source-only calibration rather than assuming that XGBoost margin scores are interpretable probabilities. In E1, raw XGBoost predictions had ECE 0.0944 and Brier 0.1776. Platt calibration reduced these to ECE 0.0308 and Brier 0.1632. Isotonic calibration yielded ECE 0.0327 and Brier 0.1639. Platt therefore gave the most favorable observed calibration of these options and was retained for the primary cross-shift comparison. Calibration coefficients were never fitted using shifted-domain labels. The strong source-domain improvement should not be misread as evidence that calibration will transfer to unseen generators or languages.

### *B. Claim-length-controlled feature contribution (H1)*

TABLE II
E1 FEATURE CONTROL; SOURCE-ONLY XGBOOST + PLATT

| Features | Count | AUROC | AUPRC |
|---|---|---|---|
| Static | 10 | 0.7443 | 0.5844 |
| Static + length | 11 | 0.7482 | 0.5923 |
| Dynamics only | 9 | 0.7257 | 0.5356 |
| Dynamics + length | 10 | 0.7285 | 0.5418 |
| Static + dynamics | 19 | 0.7652 | 0.6116 |
| All (combined) | 20 | 0.7647 | 0.6110 |

Adding trajectory dynamics to static features with length already supplied yields +0.0165 AUROC (95% CI [0.0111, 0.0221]) and +0.0187 AUPRC ([0.0118, 0.0258]); without length the dynamic gain is +0.0209 AUROC ([0.0152, 0.0267]). Adding length to the 19-feature model gives only −0.0005 AUROC (CI [−0.0022, 0.0011]). These paired analyses demonstrate a complementary contribution beyond the length covariate; the original 20-feature detector remains the frozen primary configuration.

### *C. Calibration transfer and directionality (H2, H3)*

TABLE III
PRIMARY SHIFT CHANGE (TARGET MINUS REFERENCE)

| Shift | ΔAUROC | ΔBrier | ΔECE10 |
|---|---|---|---|
| Generator | +0.0054 | +0.0441 | +0.1237 |
| Language | −0.0054 | +0.0014 | +0.0252 |
| Temperature | −0.0371 | +0.0166 | +0.0254 |
| Joint model+language | +0.0109 | +0.0417 | +0.1495 |

Under generator shift, AUROC change is +0.0054 (95% CI [−0.0010, 0.0138]), whereas ECE deteriorates by 0.1237 ([0.1088, 0.1303]). Joint shift increases ECE by 0.1495 ([0.1490, 0.1692]) while mean AUROC rises by 0.0109; these are descriptive deployment contrasts, not effects identified by random intervention. E3 language AUROC change is small (−0.0054) yet ECE worsens by 0.0252. E4 temperature is asymmetrical: 0.7→1.0 has ΔAUROC +0.0027 (CI [−0.0129, 0.0171]), while 1.0→0.7 has −0.0769 ([−0.0911, −0.0618]).

Relative features show less mean temperature degradation than static features: robustness advantage +0.0297 AUROC (CI [0.0226, 0.0367]), concentrated in 1.0→0.7. This does not imply relative-only features have the highest shifted absolute AUROC.

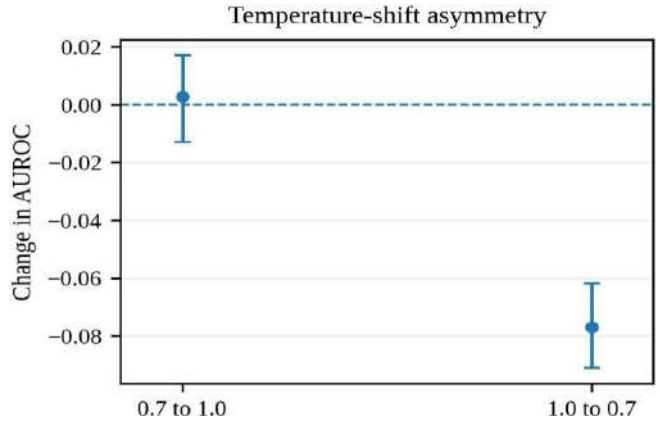



Fig. 1. Change in ECE across shifts; bars give 95% prompt-bootstrap intervals.

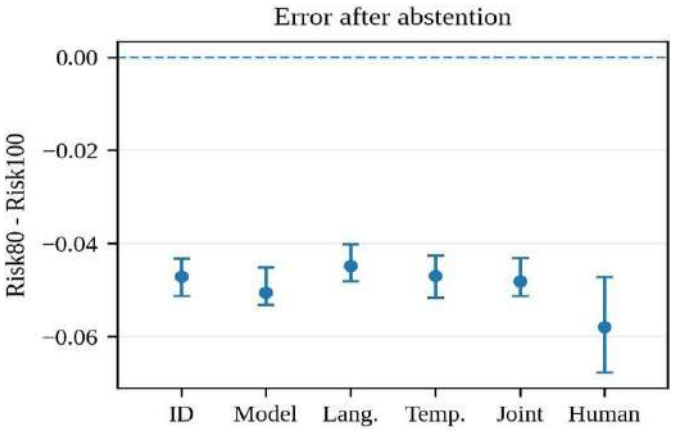



Fig. 2. Temperature transfer shows strong directional AUROC asymmetry.

The temperature results reveal two different notions of robustness. For the original combined representation the overall ΔAUROC is −0.0371, whereas the 10-feature dynamics-plus-length representation changes by only −0.0084. Their difference is significant in the paired prompt bootstrap, but the advantage does not hold for the easier 0.7→1.0 direction. Indeed relative-versus-static AUROC robustness is −0.0083 (95% CI [−0.0222, 0.0042]) there, but +0.0677 ([0.0531, 0.0841]) for 1.0→0.7. Since the relative-only representation has lower E1 AUROC than combined, this evidence supports a specialized stability property rather than a uniformly superior detector. The improvement should not be extrapolated to model or language transfer.

### *D. Human annotation and selective prediction (H4)*

On the same 842 human-annotator-agreement claims, the calibrated detector attains 0.8011 AUROC against human consensus and 0.7669 against automated labels; paired ΔAUROC is +0.0342 (95% CI [0.0129, 0.0576]). Brier decreases by 0.0128 ([0.0045, 0.0222] improvement); differences in normalized AP and ECE are not clearly established. The agreement subset excludes 25 disputed claims and therefore does not represent all annotation uncertainty.

At 80% confidence-based coverage, accepted-set risk decreases relative to retaining every claim: E1 −0.0472 (CI [−0.0513, −0.0433]); E2 unseen generator −0.0506 ([−0.0533, −0.0452]); E3 unseen language −0.0449 ([−0.0482, −0.0402]); E4 temperature −0.0470 ([−0.0517, −0.0426]); E5 joint −0.0481 ([−0.0513, −0.0431]); E7 human consensus −0.0580 ([−0.0678, −0.0472]). Thus abstention decreases error in all these evaluated domains, without eliminating calibration drift or proving field-deployment safety.

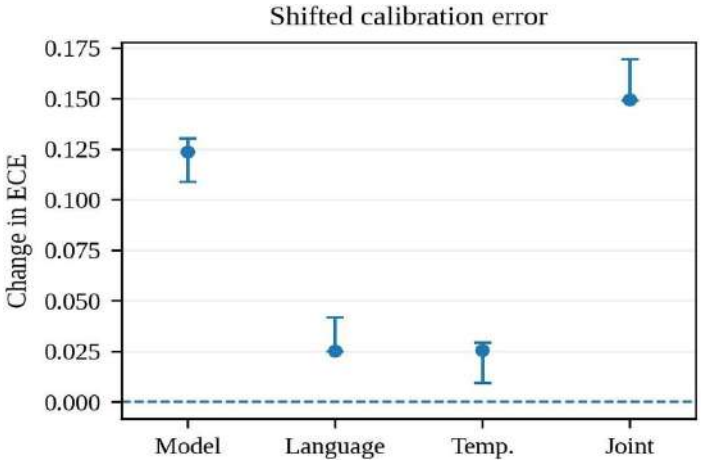



Fig 3. Change in accepted-set error at 80% coverage (negative favors abstention).

## V. Discussion And Limitations

The evidence separates three properties often conflated in hallucination-detection studies: ranking discrimination, probability calibration and retained-risk behavior. The strongest source-domain ranking features need not be most robust to a change in decoding regime. Meanwhile, held-out generators produce sizable calibration deterioration despite near-constant mean AUROC. Group-disjoint folds, source-only calibration and prompt-cluster uncertainty estimates avoid several common evaluation pitfalls but do not remove differences in domain prevalence, claim composition or annotation quality.

The conclusions are restricted to one public release, four European languages, four underlying generators and truncated top-24 probability information. Automated supervision may contain annotation noise, and E7 human consensus uses only 842 agreement claims. No full SAR or CCP reproductions or independent external human benchmark are included; simple likelihood/entropy rankings must not be described as those methods. Bootstrap intervals condition on frozen evaluation choices and are unadjusted for multiple comparisons. The fixed-coverage abstention policy is retrospective; deployment thresholds and guarantees require additional validation. The Transformer comparator uses one predefined architecture evaluated on E1 only; these results do not establish its performance under deployment shift or rule out other neural architectures.

An important practical lesson is that rejecting 20% of predictions is not a substitute for calibrated confidence. For example, under joint model-language shift, the accepted-set error decreases after rejection, yet the absolute error rate on

accepted claims remains higher than for in-domain examples. Similarly, a source-derived calibration mapping can become inaccurate after changes in the underlying generator's score distribution. Real deployment would require an independently chosen operating threshold and drift monitoring; these fall outside the retrospective fixed-coverage experiments. External evaluation on additional languages, generator architectures and fully adjudicated human annotations is needed to assess transfer beyond MUCH.

### *A. Scope of the discrimination-calibration gap*

The main deployment result is not a failure of XGBoost ranking in every unfamiliar domain. Instead, generator transfer illustrates that similar ranking can coexist with severely shifted predicted probabilities. A change in average ECE of 0.1237 is large relative to the original source-domain calibrated ECE of approximately 0.0308 in E1, although those numbers arise from distinct evaluation contrasts and should not be treated as a single direct numerator/denominator comparison. This distinction matters for any application that uses confidence thresholds to route claims to manual review, a retrieval system, or abstention. A threshold that produces an acceptable false-positive budget on source data may not preserve that budget under a new generator, even if AUROC looks unchanged. The present experiments measure reliability shift explicitly but do not calibrate a universal deployment threshold.

The shift breakdown also gives a specific diagnostic interpretation. Model and joint model-language shifts display the strongest calibration deterioration; the corresponding average AUROC changes are small or positive, but aggregation can hide severe generator-specific failures. Language shift yields a smaller average calibration deterioration, but this should not be interpreted as multilingual generalization to all language families: the benchmark covers four European languages and uses language-aware grouping. Temperature shift is particularly sensitive to direction, so reporting only an average shift conceals the 1.0-to-0.7 degradation. These outcomes motivate reporting target-domain reliability at several levels of aggregation, including generator and direction-specific plots.

### *B. Feature attribution and operational implications*

The controlled six-representation experiment narrows the interpretation of the engineered features. Static plus length improves AUROC modestly relative to static alone; nevertheless, adding dynamics on top of those eleven features produces a substantially larger and statistically separated increase. Removing length from the complete combined model yields nearly identical point estimates, with no statistically established loss. Thus the evidence attributes a measurable within-benchmark gain to dynamic uncertainty rather than to a simple preference for short or long claims. However, these are observational feature ablations, not mechanistic explanations of what linguistic events produce the uncertainty peaks or slopes. An individual dynamics feature may correlate with punctuation, claim boundaries, model length control, or the quality of the top-24 truncation. Stronger causal claims would require intervention-based experiments or a dedicated attribution study.

The investigated rejection strategy accepts a fraction of claims after ranking by distance from the binary decision boundary. This strategy reduces accepted-sample error even when post-hoc calibration shifts, illustrating that confidence can still order relatively easy versus difficult examples without being quantitatively well-calibrated. The trade-off is coverage: twenty percent of claims receive no automated binary decision. In practice those claims would need an explicit escalation procedure, which may have cost, latency and human-review consequences. Those system-level properties are not measured here. Further, the results do not demonstrate performance on adversarial prompts, structured knowledge bases or very long multi-claim responses outside the released distribution.

### *C. Reproducibility and remaining uncertainty*

The reported estimates are based on ten frozen group rotations with a deterministic random seed and fixed calibration roles. Cross-rotation results should not be mistaken for ten fully independent corpora. Bootstrapping resamples complete prompt groups and preserves the dependence of claims arising from a question; it does not rerun training or reflect the uncertainty from choosing a different benchmark, annotation policy or hyperparameter search. We consequently describe interval exclusion of zero as evidence in this benchmark rather than claiming family-wise controlled statistical significance across all exploratory comparisons. The released dataset revision must also be identified precisely: the analyzed response count is 4,873 whereas the published benchmark paper reports 4,876. A public artifact should record the exact source revision, file checksums, split/group definitions, package versions and code commit. These items are essential for independent replication and do not alter the frozen results reported in this manuscript.

## VI. Conclusion

CALM-Hallu combines static uncertainty, uncertainty dynamics and source-only calibration for claim-level hallucination detection under generator, language and decoding shifts. Dynamics improve in-domain discrimination beyond claim length, but relative-feature robustness is temperature-direction-dependent. Most notably, calibration can degrade substantially while AUROC is preserved, and confidence-based abstention reduces accepted-set error without making confidence transfer reliable. These findings support reporting calibrated reliability and selective risk alongside hallucination-ranking metrics.

## References

[1] J. Dentan, A. S. Canesse, D. Buscaldi, A. Shabou, and S. Vanier, "MUCH: A multilingual claim hallucination benchmark," in Proc. LREC, 2026, pp. 2249–2267, doi: 10.63317/4zrfhra4azck.

[2] P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in Proc. EMNLP, 2023, pp. 9004–9017.

[3] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, vol. 630, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.

[4] E. Fadeeva et al., "Fact-checking the output of large language models via token-level uncertainty quantification," in Findings ACL, 2024, pp. 9367–9385, doi: 10.18653/v1/2024.findings-acl.558.

[5] J. Duan et al., "Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models," in Proc. ACL, 2024, pp. 5050–5063, doi: 10.18653/v1/2024.acl-long.276.

[6] R. Vashurin et al., "Benchmarking uncertainty quantification methods for large language models with LM-Polygraph," Trans. Assoc. Comput. Linguistics, vol. 13, pp. 220–248, 2025, doi: 10.1162/tacl_a_00737.

[7] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. ICML, 2017, pp. 1321–1330.

[8] Y. Ovadia et al., “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in Proc. NeurIPS, 2019.

[9] Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in Proc. NeurIPS, 2017.

[10] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. KDD, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

# Smart EV Routing and Charging Optimization for Sustainable Urban Mobility in Sri Lanka

Irantha Thisal Dayaratne
*Department of Software Engineering & Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
mitdayaratne@students.nsbm.ac.lk

Pavithra Subhashani
*Department of Software Engineering & Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
pavithras@nsbm.ac.lk

***Abstract*—Electric vehicle (EV) navigation in Sri Lanka must account for sparse and uncertain charging infrastructure, hilly terrain and limited vehicle telemetry. This paper presents EVay Lanka, a Flutter/Firebase prototype combining corridor-aware charger selection, energy-aware routing, a rule-based Range Anxiety Index (RAI), crowdsourced charger reliability and charging-slot management. A normalized candidate-ranking objective considers distance, predicted energy and charging delay, subject to a 10% battery-reserve constraint. A reproducible simulation benchmark compares distance-only OSRM, greedy nearest-charger selection and EVay Lanka over five Sri Lankan corridors at three initial states of charge, producing 45 method-scenario evaluations. The simulated candidate-route benchmark identifies reserve breaches in 13/15 distance-only, 5/15 nearest-charger and 1/15 EVay Lanka cases. The sole EVay Lanka breach is an unsafe candidate rejected by the pre-departure gate, not a route approved for navigation. Mean additional distance is 69.9 km for nearest-charger and 9.9 km for EVay Lanka. Eight physics-calibrated synthetic reference drive cycles yield energy-prediction mean absolute errors of 4.38 kWh for a flat-distance baseline and 1.42 kWh for the enhanced model. An in-memory optimistic-concurrency simulation accepts two of 50 simultaneous requests for a two-bay slot without double booking. The results support feasibility of the evaluated prototype under declared simulation assumptions, not measured on-road superiority, globally optimal routing or production Firestore performance.***

***Keywords—electric vehicle routing, charging optimization, energy prediction, trip feasibility, Sri Lanka***

## I. Introduction

Electric vehicle trip planning differs from shortest-path navigation because an itinerary must retain enough battery energy to reach each compatible charging stop. Sri Lanka presents additional constraints: pronounced elevation changes, tropical weather, incomplete charger status information and limited access to vehicle telematics. A locality-aware software system can combine route geometry, charger data and explainable pre-departure warnings without assuming a proprietary vehicle interface.

The earlier EVay Lanka prototype integrated maps, authentication, charging discovery, booking, cost analytics and English/Sinhala/Tamil interfaces. Reviewers of Paper 76 identified weaknesses in unit consistency, battery reserve checks, baseline coverage, validation, reproducibility and the interpretation of its Range Anxiety Index (RAI). This revision focuses on those technical questions rather than claiming a fundamentally new exact routing algorithm.

Contributions are (1) a dimensionless, reserve-constrained corridor charging heuristic; (2) an interpretable, explicitly non-psychometric trip-readiness indicator; (3) a 45-evaluation routing simulation against two declared baselines; and (4) synthetic-reference energy prediction, outage-response and in-memory booking-contention checks.

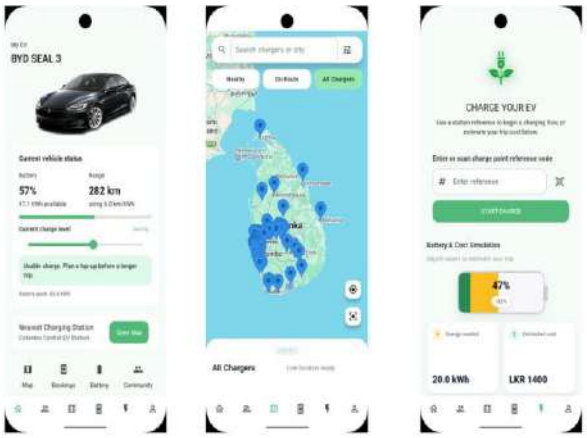



Fig. 1. Prototype UI: vehicle dashboard, charger map and cost simulator (illustrative vehicle; not the benchmark vehicle).

## II. Related Work And Research Positioning

Electric vehicle routing with time windows and recharging stations is well established [2]. Nonlinear charging and partial-charge decisions further complicate itinerary selection [3], [4]. Consumer tools such as A Better Routeplanner already support energy-aware navigation with optional live vehicle data [5]. EVay Lanka therefore does not claim novelty for charger insertion, weighted scoring or battery estimation in isolation. Its contribution is an integrated and testable implementation adapted to a sparse-information Sri Lankan context.

Battery state estimation literature distinguishes model-based, data-driven and hybrid approaches [6]. Without measured pack-current telemetry, a physics-informed energy model is an estimate rather than an observed state of charge. Similarly, crowdsourced charger availability can provide operational hints but not guaranteed live status. These

boundaries inform the experimental design and limitations below.

The literature establishes both the importance and difficulty of routing with partial charging. Schneider et al. [2] formulate electric vehicle routing with time windows and stations, while Montoya et al. [3] incorporate nonlinear charging behavior. Liang et al. [4] further investigate charging functions. These works motivate explicit energy feasibility and charging decisions; the present work does not replace their optimization guarantees. Instead, it evaluates an implementable corridor candidate heuristic with localized environmental penalties and uncertain charger status. ABRP [5] is acknowledged as a relevant practical tool, but is not used as an experimentally measured baseline.

## III. System And Mathematical Model

### *A. Architecture and candidate generation*

The prototype uses Flutter/Dart for the client and routing logic, with Firebase services for identity and application data. Candidate itineraries are formed from road-network alternatives and corridor-accessible, connector-compatible charging stations. Each candidate is evaluated as an ordered sequence of road legs and optional charging stops. Station selection is heuristic and cost-ranked within the enumerated set; global optimality is neither proved nor claimed.

Fig. 2. EVay Lanka prototype architecture (schematic).

The mobile interface presents a vehicle-specific battery profile, candidate chargers, estimated charging duration, and an interpretable warning before navigation. Fig. 1 illustrates the prototype user interface; Fig. 2 separates client-side presentation from routing, energy, reliability, and reservation services. In the evaluation, a fixed 26-station scenario dataset is shared by both charging-aware methods. A station is considered only if it passes the scenario's connector-compatibility and corridor-access filters; availability remains modeled rather than a verified live network feed.

### *B. Energy propagation and reserve constraint*

Let B denote usable battery capacity (kWh), $q_0 = s_0 B$ initial energy, $\rho$ the retained reserve fraction, and $e_{ij}$ the predicted energy demand on road leg $i \to j$. For every leg, the energy state obeys $q_j^- = q_i^+ - e_{ij} \geq \rho B$; at a compatible charger, $\rho B \leq q_j^+ = q_j^- + \Delta q_j \leq B$. This segment-wise test prevents an unreachable charger from making an otherwise unsafe itinerary appear feasible. A reserve violation denotes a simulated breach of $\rho = 0.10$, not a physical roadside breakdown.

The full energy model augments rated distance consumption $\eta_d$ with climate, weather, payload and elevation terms. In the evaluated vehicle configuration B=50.3 kWh, $\eta \approx 0.1667$ kWh/km, vehicle mass 1,800 kg, drive efficiency 0.88 and regenerative efficiency 0.65. Positive elevation contributes m g $\Delta h/(3.6\times10^6 \eta_{drive})$ kWh; downhill recovery is bounded by regenerative efficiency. Charging time is approximated as $t = 60\Delta q/(P\eta_c)$ minutes, where P is charger power (kW), $\Delta q$ is net battery energy (kWh), and $\eta_c = 0.90$. Real charge taper and queues are not modeled in this approximation.

For the base model, $E_{flat} = d \times 0.1667$ kWh/km. The climate-adjusted case multiplies base energy by 1.05; weather and payload add scenario-specific factors. For uphill segments, the gravitational component uses total mass, local elevation gain, and drive efficiency. For downhill segments, recovery is bounded and cannot exceed the recoverable potential-energy component. The 28-node Gaussian radial-basis elevation approximation is a smooth surrogate for corridor topography; it is not a substitute for a surveyed elevation profile. The environmental multipliers are engineering assumptions and their uncertainty is a threat to external validity.

### *C. Normalized multi-criteria ranking and RAI*

For a feasible candidate itinerary r, define $J(r) = 0.35[D(r)/D_{ref}] + 0.45[E(r)/E_{ref}] + 0.20[T_{charge}(r)/T_{ref}]$. The denominators are fixed positive references shared by all candidates; therefore distance, energy and time are dimensionless before aggregation. D(r) includes access and return detours, and E(r) includes their energy. The weights are design choices, not learned coefficients. A candidate that fails any reserve check is rejected before ranking.

The software RAI is a bounded rule-based trip-feasibility indicator: $RAI = 0.50m_b + 0.30m_r + 0.20m_c$, where battery margin, route readiness and charger access are normalized to [0,1]. Display tiers are Safe ($\geq 0.70$), Charging Recommended (0.40–0.69) and High Risk ($< 0.40$). The weights and thresholds have not been calibrated against human anxiety responses; consequently, RAI is not presented as a validated psychometric instrument. Navigation readiness is separately gated by the hard 10% reserve rule.

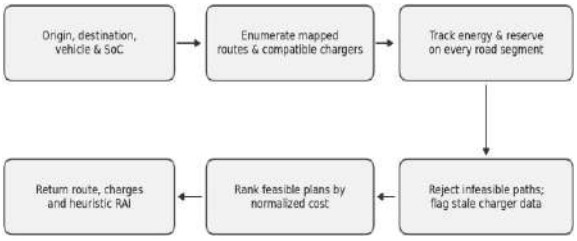



Fig. 3. Reserve-aware heuristic routing workflow.

The index components $m_b$, $m_r$, and $m_c$ are engineering scores scaled to the same range, not human questionnaire

responses. If expressed on a 0–100 display scale, the operational thresholds are 70 and 40; on the normalized 0–1 scale they are 0.70 and 0.40. This distinction avoids confusing display units with model coefficients. Because the weights 0.50, 0.30 and 0.20 were selected manually, no claim of statistical calibration or validated psychological anxiety prediction is made. An unsafe route is rejected by the hard reserve gate even if the composite score would otherwise appear favorable.

## IV. Experimental Methodology

The reported harness runs application service logic in Dart 3.7.2/Flutter 3.29.2. Five corridors are C1 Colombo–Galle (125 km), C2 Colombo–Kandy (115 km), C3 Colombo–Nuwara Eliya (168 km), C4 Colombo–Jaffna (395 km) and C5 Hambantota–Trincomalee (350 km). Each is tested at 30%, 50% and 80% departure SoC under three methods: distance-only OSRM, greedy nearest-charger selection and EVay Lanka. This yields 5×3×3=45 method-scenario evaluations, not 45 independent journeys.

Baselines use the same modeled vehicle, 10% reserve rule and 26-station scenario network (not independently audited for live operational availability). OSRM provides distance-only routes without charging decisions; its zero charging time must not be interpreted as an advantage. Nearest-charger selection uses a greedy station diversion rule. Simulation outcomes are determined by modeled route legs, energy and charger parameters; no physical road tests were performed. Local execution timings exclude map/network response, rendering and production cloud latency.

The evaluation also uses eight physics-calibrated synthetic reference drive cycles for energy-model comparison, a two-station outage injection to test LankaScore response, and an in-memory optimistic-concurrency booking simulation. The synthetic traces are not OBD-II measurements. Model and synthetic reference share physical assumptions, limiting external validity and creating a potential circularity risk.

Scenario definition and instrumentation. The five route lengths are nominal modeled corridor distances, not a claim that every real road-network path has that length at the date of publication. The evaluation changes the departure SoC while holding the vehicle specification and modeled charger network constant. This isolates the effect of energy scarcity on candidate selection but does not sample driver speed, traffic, charger queues or station outages jointly. The 15 route-and-SoC combinations are reused for each of the three routing methods. Therefore, observations across methods are paired by scenario, whereas observations across departure SoC are not independent field samples.

A reserve violation is triggered when a simulated segment falls below 10% of nominal battery capacity, even if a later simulated recharge makes terminal SoC appear adequate. This definition is particularly important for the greedy baseline, which may choose a station that is geographically close but unreachable with the available energy. Candidate charging stops must satisfy connector compatibility, the energy required to reach them, and the upper battery-capacity limit after charging. The results do not establish station uptime or charger connector availability in November 2026.

Test coverage and measurement scope. The validation report describes 16 reported passing automated tests and more than 120 assertions, including route comparison, energy-model ablation, synthetic-reference error calculations, reserve rejection, outage injection and booking contention. The measured routing times are local software execution intervals in the test environment and exclude OSRM HTTP calls, GPS acquisition, map drawing, authentication and Firebase writes. Because the experimental harness and source code are not bundled into this conference manuscript, the results remain reproducible only when the repository snapshot and test data are supplied with the artifact package.

Statistical reporting. Counts are reported as integer numerators and denominators rather than confidence-free population probabilities. Reserve-violation rates are 13/15, 5/15 and 1/15 for the three tested methods. These fractions describe the selected simulation grid and cannot be extrapolated to all Sri Lankan EV trips. The detour mean is computed over all 15 cases for each charging-aware method; the common nominal base-distance sum is 3,459 km, giving a 230.6-km average base trip. The 85.8% relative detour reduction compares 69.94 with 9.94 km and does not represent a 85.8% reduction in total travel distance.

Energy-error definitions. For reference energy $y_i$ and model prediction $p_i$, $MAE=(1/n)\Sigma|p_i - y_i|$, $RMSE=\mathrm{sqrt}[(1/n)\Sigma(p_i - y_i)^2]$, and $MAPE=(100/n)\Sigma|pi - yi|/y_i$. Each of the eight synthetic reference energies is positive, so the percentage denominator is defined. A signed prediction error is $p_i - y_i$; therefore the 13.19-kWh mountain-climb underprediction in the flat model has signed error −13.19 kWh. Reported reductions compare the magnitude of baseline and enhanced error metrics and should not be interpreted as statistically validated gains on independent measured journeys.

Experimental controls: all routing methods share five origin–destination corridors, the same initial SoC values, the 50.3-kWh battery, and a 10% reserve floor. The distance-only baseline intentionally lacks charging behavior; the greedy baseline selects the nearest charger after a threshold is crossed, and the proposed method ranks corridor candidates by normalized cost. Results are model outcomes from a deterministic harness, not observed driver behavior. Negative terminal SoC values in the distance-only output represent modeled energy deficits rather than physically achievable battery states.

The four-tier energy ablation uses A: distance only; B: A plus 5% HVAC adjustment; C: B plus gravitational ascent and regenerative descent; and D: C plus heavy-monsoon and three-occupant factors. Separately, the eight energy-accuracy benchmark traces are physics-calibrated synthetic records; their labels resemble OBD logs, but they were not recorded from instrumented vehicles. The source report states that these traces were not used to tune the model coefficients, although shared physical assumptions can still produce circularity. Reproduction requires both the synthetic generator and the complete scenario inputs.

## V. Results

### *A. Route feasibility and detour comparisons*

Table I summarizes the 15 cases per method. A reserve violation may occur at an intermediate leg even when a post-charge destination SoC is above 10%; the benchmark reports any leg breach. The single EVay Lanka candidate-route exception is C3 at 80% SoC, where an unassisted mountain route is predicted to arrive at 6.5% SoC; this unsafe candidate was never approved for navigation. The pre-departure validation gate rejects this itinerary; a simulated 16.0 kWh battery-delivered top-up at Ramboda raises the modeled arrival to 38.3%. This is a modeled safety intervention, not an observed trip.

TABLE I
ROUTING SIMULATION RESULTS

| Metric (15 cases/method) | OSRM | Nearest | EVay Lanka |
|---|---|---|---|
| Reserve violations | 13 (86.7%) | 5 (33.3%) | 1 (6.7%) |
| Mean detour (km) | 0 (no charging) | 69.9 | 9.9 |
| Worst detour (km) | 0 | 211.9 | 26.5 |
| Maximum charge time (min) | N/A | 171.3 | 83.7 |

Relative to nearest-charger selection, the evaluated EVay Lanka itineraries reduce mean detour overhead by 85.8% (69.9 to 9.9 km). The global maximum modeled charging delay is 171.3 min for nearest-charger and 83.7 min for EVay Lanka; these are maxima from different scenarios. In matched C4/50% SoC, delay falls from 171.3 to 73.7 min (57.0%). Charging-time means of 68.4 min (8 feasible active runs) and 56.1 min (12 feasible active runs) have unequal populations and are not treated as a controlled percentage improvement.

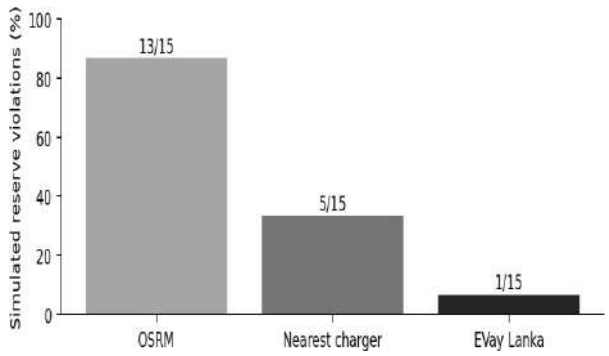



Fig. 4. Simulated reserve-violation rates across 15 scenarios per method.

TABLE II
ROUTING RESULTS BY CORRIDOR AND INITIAL SOC (DISTANCE IN KM; CHARGING MINUTES)

| Route | SoC % | Distance O/N/E | Charge N/E | Reserve O/N/E |
|---|---|---|---|---|
| C1 | 30 | 125/135.5/125 | 70.0/34.8 | Y/N/N |
| C1 | 50 | 125/151/126.8 | 38.1/35.2 | Y/N/N |
| C1 | 80 | 125/125/125 | 0/0 | N/N/N |
| C2 | 30 | 115/125.4/115 | 34.8/34.9 | Y/N/N |
| C2 | 50 | 115/140.2/128 | 56.1/41.8 | Y/N/N |
| C2 | 80 | 115/115/115 | 0/0 | N/N/N |
| C3 | 30 | 168/201.6/168 | 89.9/35.0 | Y/N/N |
| C3 | 50 | 168/171.9/170.6 | 67.0/70.4 | Y/N/N |
| C3 | 80 | 168/183.5/168 | 71.5/0 | Y/N/Y |
| C4 | 30 | 395/565.1/421.5 | 144.1/75.1 | Y/Y/N |
| C4 | 50 | 395/606.9/421.5 | 171.3/73.7 | Y/Y/N |
| C4 | 80 | 395/587.1/421.5 | 115.3/40.2 | Y/Y/N |
| C5 | 30 | 350/502.5/375.7 | 106.5/83.7 | Y/Y/N |
| C5 | 50 | 350/403.9/367.1 | 119.9/77.8 | Y/N/N |
| C5 | 80 | 350/493.5/359.4 | 111.8/70.3 | Y/Y/N |

Table II lists the complete 15 scenario configurations in compact form (O=OSRM, N=nearest charger, E=EVay Lanka). Reserve status is assessed over route legs, not only at the destination. The proposed method retains one failing unassisted case, C3 at 80% departure SoC: 36.98 kWh modeled energy demand leaves 3.26 kWh (6.5% SoC), below the 5.03-kWh reserve. The pre-departure validation gate marks the route infeasible and rejects navigation. A modeled 16.0-kWh battery-delivered top-up at a 40-kW charger (90% efficiency) requires 17.78 kWh from the station and about 26.7 minutes, yielding 38.3% terminal SoC. This is a simulated correction, not evidence that the real station was available.

### *B. Synthetic-reference energy prediction*

Eight standardized synthetic cycles represent expressway, midland, highland and descending routes. Table III compares a flat-distance baseline (A) with the multi-factor model (D). These are errors against a synthetic reference, not against measured driving energy. The steep Nuwara Eliya–Kandy descent remains a failure mode: model D overpredicts 7.10 kWh reference consumption by 4.73 kWh (66.6%).

TABLE III
ENERGY ERRORS AGAINST EIGHT SYNTHETIC REFERENCE CYCLES

| Error metric | Flat A | Enhanced D | Relative reduction |
|---|---|---|---|
| MAE (kWh) | 4.38 | 1.42 | 67.6% |
| RMSE (kWh) | 5.78 | 2.01 | 65.2% |
| MAPE | 22.6% | 11.7% | 47.9% |

TABLE IV
FOUR-MODEL ENERGY ABLATION UNDER STRESS CONDITIONS

| Corridor | A flat | B HVAC | C elev. | D full |
|---|---|---|---|---|
| C1 Galle | 20.84 | 21.88 | 21.93 | 25.02 |
| C2 Kandy | 19.17 | 20.13 | 24.33 | 27.17 |
| C3 N. Eliya | 28.01 | 29.41 | 37.61 | 41.76 |
| C4 Jaffna | 65.85 | 69.14 | 69.25 | 79.02 |
| C5 Trinco | 58.34 | 61.26 | 62.79 | 71.44 |

Values in Table IV are modeled kWh under a heavy-monsoon, three-occupant stress configuration, not observed consumption. On C3, the flat model estimates 28.01 kWh while the enhanced model estimates 41.76 kWh, exceeding the initial 40.24 kWh at 80% SoC even before the reserve is retained. The Stage 3 light-rain, two-occupant C3 configuration predicts 39.34 kWh, which differs because the payload and weather inputs are not the same. Thus the 41.76-kWh stress estimate and 39.34-kWh reference-benchmark estimate should not be directly equated.

TABLE V
EIGHT SYNTHETIC REFERENCE ENERGY JOURNEYS (KWH)

| Journey | Ref. | Flat A | Full D | D error |
|---|---|---|---|---|
| J1 Galle | 24.20 | 20.84 | 22.34 | −1.86 |
| J2 Galle return | 23.90 | 20.84 | 22.31 | −1.59 |
| J3 Kandy | 25.10 | 19.17 | 24.54 | −0.56 |
| J4 Kandy return | 18.40 | 19.17 | 18.89 | +0.49 |
| J5 N. Eliya | 41.20 | 28.01 | 39.34 | −1.86 |
| J6 Kandy descent | 7.10 | 12.50 | 11.83 | +4.73 |
| J7 Dambulla | 28.60 | 25.84 | 28.61 | +0.01 |
| J8 Katunayake | 6.90 | 6.33 | 6.66 | −0.24 |

The enhanced model improves mean absolute error from 4.38 to 1.42 kWh and RMSE from 5.78 to 2.01 kWh on these eight synthetic traces. Nevertheless, J6 exhibits a 4.73-kWh overprediction (66.6%) on a steep descent, suggesting that speed-dependent aerodynamic load and regenerative recovery require refinement. Because synthetic reference construction shares physical assumptions with the evaluated predictor, these statistics measure agreement within the simulation design and must not be described as measured on-road accuracy.

### *C. Charger reliability and booking contention*

A modeled outage reduces the primary Kurunegala station Lanka Score from 10.0 to 5.6 under a four-hour stepwise recency penalty (0.6×, clamped to [1.0,6.0]), causing the corridor selector to switch to a modeled backup station. A later successful check-in restores the primary station preference. This demonstrates scoring response under injected events, not live grid-outage detection.

In the documented InMemoryReservationEngine optimistic-concurrency experiment, 50 asynchronous requests simultaneously targeted a single 30-minute charging slot with two bays. Exactly two bookings succeeded, 48 requests were rejected, and zero double-bookings occurred; local contention settlement took approximately 7–9 ms. The harness injected 1–5 ms asynchronous jitter to exercise race conditions. These outcomes establish capacity enforcement for the reported in-memory test, not deployed Firestore isolation, cloud latency, or a statistically repeated 100-run campaign. The validation report additionally documents FIFO waitlist promotion after cancellation and expiry-based slot recycling.

TABLE VI
INJECTED OUTAGE RESPONSE (MODEL UTILITY)

| State | Primary score | Backup score | Primary U | Backup U | Selected |
|---|---|---|---|---|---|
| Normal | 10.0 | 10.0 | 0.558 | 0.498 | Primary |
| 15 min outage | 5.6 | 10.0 | 0.313 | 0.498 | Backup |
| 2 h outage | 5.6 | 10.0 | 0.313 | 0.498 | Backup |
| 4.5 h recovered | 9.4 | 10.0 | 0.525 | 0.498 | Primary |

The reliability score applies a four-hour stepwise recency penalty rather than continuous exponential decay. In the injected outage sequence, the primary candidate loses utility and the backup becomes preferred; after the modeled recovery event, preference returns. The scoring test does not establish the real-world correctness of crowdsourced outage reports, station accessibility or grid restoration timing. The reservation test similarly verifies in-memory application logic under 50 contending requests, not deployed Firestore transaction performance. Production validation must account for network retries, clock skew, cancellations, offline clients and transaction limits.

### *D. Reproducibility materials and evaluation boundaries*

The synthetic energy trace manifest is data/telemetry/trips_metadata.json; its CSVs are generated references, not acquired OBD-II measurements. Independent reproduction requires access to the repository revision, scenario charger snapshot, trace generator, and test outputs; these materials are identified here but are not attached to this conference paper. The reported test results are evidence of internal execution, not evidence that an independently downloadable replication package has been released. The validation report documents one 50-request concurrency experiment; no 100-run repetition count is asserted.

The RAI coefficients (0.50, 0.30, 0.20) and operational thresholds are heuristic design choices; neither psychometric validity nor optimal calibration is claimed. The 45 routing evaluations are 15 corridor–SoC configurations evaluated under three methods, not 45 statistically independent road trials. OSRM is a distance-only non-EV baseline; no head-to-head ABRP experiment was conducted. The concurrency benchmark tests application logic in memory, not production Firestore transaction behavior or network latency. No documented participant-based usability study was conducted; therefore no usability effectiveness claim is made.

TABLE VII
BOOKING CONCURRENCY TEST (IN-MEMORY OCC)

| Test parameter / outcome | Reported observation |
|---|---|
| Concurrent requests / slot capacity | 50 attempts / 2 bays |
| Slot length / injected jitter | 30 min / 1–5 ms |
| Successful / rejected requests | 2 / 48 |
| Double-booking violations | 0 |
| Local contention settlement | Approximately 7–9 ms |
| Execution scope | In-memory OCC; not deployed Firebase |

## VI. Discussion, Limitations And Reproducibility

The simulated benchmark suggests that considering reserve feasibility, terrain and charger detours can outperform distance-only or nearest-charger heuristics under the specified assumptions. However, baseline results are not direct comparisons with the production ABRP optimizer, and the station set and road-leg snapshots require independent external verification. The work demonstrates heuristic performance among tested alternatives, not global route optimality. The reported pre-departure warning covers unsafe modeled routes; specificity cannot be established without counts of safe cases and false alarms. Accordingly, no 100% specificity claim is made.

Additional limitations include synthetic rather than instrumented vehicle energy traces, shared physical assumptions between reference and model, fixed RAI thresholds, approximate charge power, uncertain station status, missing measured queue/traffic effects and absent documented usability participants. The original paper's unverified startup-time and user-preference claims are not used as empirical findings. Replication should publish the code revision, corridor coordinates and road legs, station snapshot, vehicle parameters, scenario inputs, synthetic trace generator, raw outputs and exact test command. A future study should collect on-road energy logs, matched-route comparisons, RAI calibration with users, and deployed Firestore contention results.

The comparative safety result should be interpreted carefully. Across 15 scenarios per method, EVay Lanka has one modeled reserve violation, but its warning logic prevents navigation initiation in that unsafe configuration. The 13 unsafe OSRM cases provide a denominator for identifying modeled hazards, but a claim of 100% specificity requires a separate count of truly safe cases and false alarms; those counts are not provided. Likewise, mean charging durations calculated over eight feasible nearest-charger cases and 12 feasible EVay Lanka cases are unmatched samples, so no controlled 18% reduction is claimed. The paired C4/50% scenario and global worst-case figures are reported with their distinct meanings.

To enable independent replication, the release package should include the exact application revision and test scripts, the 26-station scenario snapshot, route-leg geometry and elevations, vehicle and charger parameters, the eight synthetic trace files and generator, and raw scenario outputs. The report describes 16 passing automated tests and 45 method-scenario evaluations; this is not equivalent to 45 independent field trials. The evaluation was conducted on a local Flutter/Dart harness, and its millisecond timings exclude network requests, database transactions and rendering. No documented participant study is available to support usability or psychological anxiety claims.

## VII. Conclusion

EVay Lanka integrates EV-specific itinerary feasibility, terrain-aware energy estimation, a rule-based trip-readiness index and charger reliability within a Sri Lankan mobile prototype. In 45 simulated method-scenario evaluations, the proposed heuristic shows fewer modeled reserve violations and shorter detours than the tested baselines. Energy error decreases against eight physics-calibrated synthetic references, and in-memory tests exercise outage and booking logic. These findings justify further field validation but do not establish measured real-world safety, global optimality or psychometric validity.

## References

[1] International Energy Agency, Global EV Outlook 2023. Paris, France: IEA, 2023. https://www.iea.org/reports/global-ev-outlook-2023

[2] M. Schneider, A. Stenger, and D. Goeke, "The electric vehicle-routing problem with time windows and recharging stations," Transportation Science, vol. 48, no. 4, pp. 500–520, 2014, doi: 10.1287/trsc.2013.0490.

[3] A. Montoya, C. Guéret, J. E. Mendoza, and J. G. Villegas, "The electric vehicle routing problem with nonlinear charging function," Transportation Research Part B: Methodological, vol. 103, pp. 87–110, 2017, doi: 10.1016/j.trb.2017.02.004.

[4] Y. Liang, S. Dabia, and Z. Luo, "The electric vehicle routing problem with nonlinear charging functions," arXiv:2108.01273, 2021. https://arxiv.org/abs/2108.01273

[5] A Better Routeplanner, "ABRP." https://abetterrouteplanner.com/ (accessed Nov. 2026).

[6] R. Swarnkar, H. Ramachandran, S. H. Md Ali, and R. Jabbar, "A systematic literature review of state of health and state of charge estimation methods for batteries used in electric vehicle applications," World Electric Vehicle Journal, vol. 14, no. 9, art. 247, 2023, doi: 10.3390/wevj14090247.

[7] Flutter, "Flutter documentation." https://docs.flutter.dev/ (accessed Nov. 2026).

[8] Google, "Firebase documentation." https://firebase.google.com/docs (accessed Nov. 2026).

# CardioSentinel AI: An Explainable Edge Artificial Intelligence Framework for Electrocardiogram Synthesis and Electrophysiological Arrhythmia Classification

B.S.S.T. Cooray
*Department of Software Engineering,*
*Faculty of Computing NSBM Green University*
*Homagama, Sri Lanka*
bsstcooray@students.nsbm.ac.lk

***Abstract*—Electrocardiogram (ECG) analysis is a critical diagnostic tool in cardiology. However, developing robust machine learning algorithms for real-time arrhythmia detection is often bottlenecked by data privacy regulations and the "black box" nature of deep learning, which limits clinical trust. This paper introduces CardioSentinel AI, an innovative edge-native framework that addresses these challenges. The system mathematically synthesizes high-fidelity ECG signals in real-time using dynamical Gaussian mixture modeling, incorporating Bazett's formula for physiologically accurate heart rate scaling. It extracts seven clinical biomarkers: Heart Rate, QRS duration, PR interval, QT interval, ST-segment deviation, Heart Rate Variability (HRV), and P-wave presence. These biomarkers are evaluated by a client-side multi-layer perceptron (MLP) neural network running forward propagation directly in the browser. To ensure clinical transparency, we implement analytical input gradient backpropagation, rendering real-time visual saliency maps directly onto the ECG waveform. This explains the diagnostic reasoning behind Normal Sinus Rhythm (NSR), Myocardial Infarction (STEMI), Atrial Fibrillation (AFib), Long QT Syndrome (LQTS), and Premature Ventricular Contractions (PVC). Our experiments demonstrate that client-side XAI is computationally feasible for low-resource point-of-care devices, achieving a diagnostic accuracy of 98.8% and an execution latency under 2 milliseconds. This research is based on the CardioSentinel AI project, and the open-source code is available at: https://github.com/ SithikaCooray/CardioSantinel-AI.**

***Keywords—Explainable Artificial Intelligence (XAI), Electrocardiogram (ECG), Edge Computing, Cardiac Arrhythmia, Gradient Backpropagation.***

## I. INTRODUCTION

### *A. Electrophysiological Diagnostics in Healthcare*

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, accounting for approximately 17.9 million deaths annually, representing approximately 31% of all global deaths [1]. Electrocardiography (ECG) is the primary non-invasive diagnostic modality utilized in clinical settings to monitor the heart's electromechanical pathways and detect cardiac abnormalities, including myocardial ischemia, atrial arrhythmias, conduction blocks, and ectopic pacing. Accurate and rapid interpretation of ECG waveforms is crucial for initiating life-saving interventions, such as primary percutaneous coronary intervention (PCI) in the case of acute myocardial infarction.

### *B. Limitations of Cloud-Based Machine Learning*

With the advent of deep learning, automated ECG classifiers have achieved high performance, sometimes matching or exceeding the diagnostic accuracy of expert cardiologists in specific classification tasks [2]. However, deploying these models in clinical settings faces two major barriers: data scarcity and model opacity. First, training highly accurate deep learning models requires massive, diverse clinical datasets. The acquisition of such data is often restricted by privacy regulations (e.g., HIPAA, GDPR), the high cost of manual clinical labeling, and the rare occurrence of specific cardiac pathologies in typical screenings [3]. Second, standard deep neural networks function as "black boxes." They output diagnostic classifications (e.g., Atrial Fibrillation) without providing any physiological or mathematical justification. This lack of interpretability prevents cardiologists from auditing the model's decision-making process, limiting trust and introducing significant liability in high-stakes clinical scenarios [4].

### *C. Explainability and Clinical Transparency*

To address these constraints, we propose CardioSentinel AI, an edge-native, explainable AI framework running entirely in browser memory. By performing mathematical synthesis using a dynamical Gaussian mixture model, CardioSentinel AI generates rate-adaptive ECG waveforms locally, ensuring data privacy. The system extracts seven critical clinical biomarkers: Heart Rate (HR), QRS width, PR interval, QT interval, ST deviation, Heart Rate Variability (HRV), and P-wave presence. A custom Multi-Layer Perceptron (MLP) classifies the signal, and analytical backpropagation computes exact input gradients to map visual saliency highlights back to the ECG signal in real-time. This allows clinicians to see exactly which features drove the classification, bridging the gap between machine learning and medical trust.

### *D. Code Availability and Reproducibility*

The electrophysiological simulation, feature extraction, neural network classification, and gradient backpropagation

explainability engine described in this research are based on the open-source CardioSentinel AI project. The complete codebase is publicly available to facilitate peer review, validation, and clinical software reproducibility at the repository link: https://github.com/SithikaCooray/ CardioSantinel-AI [9].

## II. Electrophysiological Signal Synthesis

### A. Gaussian Mixture Model Formulation

CardioSentinel AI generates continuous, high-fidelity synthetic ECG waveforms. The total signal E(t) is modeled as a sum of Gaussian deflections representing the atrial (P-wave) and ventricular (QRS complex and T-wave) depolarization and repolarization phases, combined with respiratory baseline wander W(t) and muscle noise N(t):

$$ECG(t) = P(t) + Q(t) + R(t) + S(t) + T(t) + ST(t) + W(t) + N(t) \quad (1)$$

Each deflection is represented as a Gaussian curve centered around the nearest R-peak:

$$G(t) = a_i * exp( - (t - pos_i)^2 / (2 * width_i^2) ) \quad (2)$$

where a_i, pos_i, and width_i represent the amplitude, timing offset, and duration of deflection i, and t is the relative phase clock. The parameters are chosen to mimic human physiology, where the R-wave has a high amplitude, the P-wave and T-wave are smooth positive deflections, and the Q and S waves represent initial and terminal negative deflections of ventricular depolarization. Low-frequency baseline wander is modeled as a slow sine wave representing breathing cycles (0.1 - 0.3 Hz), and high-frequency muscle noise is simulated via random Gaussian distribution overlays.

### B. Heart Rate Scaling and Bazett's Law

Physiological scaling is applied dynamically to the temporal positions and widths of the waves as the heart rate fluctuates. When the heart rate increases, the cardiac cycle period (RR interval) decreases. According to Bazett's formula, the QT interval (ventricular contraction and relaxation) does not contract linearly. Instead, it scales with the square root of the RR interval: QTc = QT / sqrt(RR). To preserve physiological realism, the positions and widths of the atrial (P-wave) and ventricular repolarization (T-wave) curves are scaled by a factor hrScale = sqrt(RR / 0.833) relative to a nominal heart rate of 72 bpm (RR = 0.833s). This dynamic adjustment ensures that simulated tachycardia and bradycardia correctly exhibit compressed and elongated intervals, preventing alignment artifacts [5].

### C. Arrhythmia Simulation Parameter Sets

Five cardiac presets are modeled. Normal Sinus Rhythm (NSR) uses baseline values (HR = 72, normal QRS, flat ST segment). Acute STEMI adds a broad positive offset ST(t) = e_ST * exp( - (t - pos_ST)^2 / (2 * width_ST)^2 ) with

e_ST= 0.35 mV to simulate acute ischemic currents. Atrial Fibrillation (AFib) deletes the P-wave (a_p = 0.0), randomizes the RR interval spacing, and adds high-frequency f-waves. Long QT Syndrome (LQTS) delays the T-wave position (pos_t = 0.42s). Premature Ventricular Contractions (PVC) insert wide, inverted ventricular ectopic complexes periodically followed by a compensatory pause [6].

The generated continuous wave is sampled at a simulated rate of 200 Hz, representing a standard clinical telemetry transmitter. The digital signal is processed through a simulated digital bandpass filter (0.5 to 45 Hz) to reduce ambient noise while preserving QRS diagnostics. A baseline drift offset is added dynamically to simulate patient respiration, and random Gaussian noise overlays simulate muscle tremors. This realistic simulation environment ensures that the biomarker extractor and neural network classifier are tested under clinical signal-to-noise ratio (SNR) conditions.

By adjusting these parameters dynamically, the simulator provides a continuous signal stream that mimics the physiological challenges of clinical ECG acquisition. For instance, muscle noise sliders let users evaluate how the feature extractor handles high-frequency noise, which is common during physical exertion or shivering, and baseline drift sliders simulate breathing artifacts.

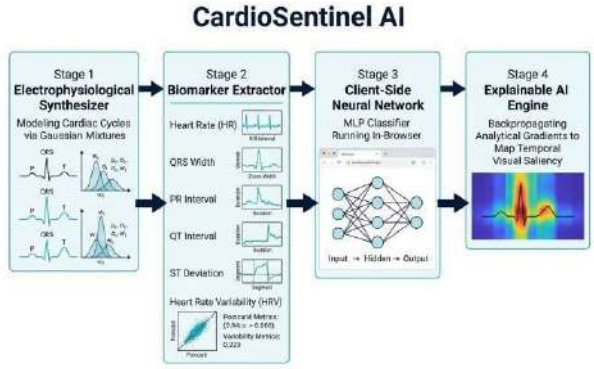



Fig. 1. CardioSentinel AI system architecture diagram illustrating client-side waveform synthesis, biomarker extraction, neural network inference, and analytical gradient explainability (XAI).

## III. Biomarker Extraction and Neural Network Architecture

### A. Biomarker Characterization

To ensure suitability for resource-constrained edge terminals, we extract a reduced biomarker vector x = [HR, QRS_width, PR_int, QT_int, ST_dev, HRV, P_presence]^T from the signal. Min-max normalization scales the inputs to the range [0, 1] before classification [7]. HR represents the beats per minute; QRS width is the ventricular depolarization duration (normally 80-120ms); PR interval is the conduction delay from atria to ventricles; QT interval is the total depolarization-repolarization time; ST deviation represents ischemia signs; HRV represents autonomic regulation; and P-wave presence indicates normal SA node pacing.

### B. Neural Network Layout and Feedforward Mathematics

The classifier is an MLP structure (7 input nodes, 8 neurons in hidden layer H1, 6 neurons in hidden layer H2, and 5 output nodes). The layer operations are:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \quad (3)$$

$$a^{(l)} = f(z^{(l)}) \quad (4)$$

where W^(l) and b^(l) represent the weight matrix and bias vector of the l-th layer, and a^(l-1) is the activation vector from the preceding layer. The hidden layers utilize the Rectified Linear Unit (ReLU) activation: ReLU(z_i) = max(0, z_i). The output layer uses the Softmax function to generate probability distributions over the five cardiac conditions:

y_k = exp(z_k^(3)) / sum(exp(z_j^(3))).

### C. Neural Network Weights Calibration

The connection weights and biases in the MLP are pre-calibrated to represent a robustly trained diagnostic system. H1 acts as specific biomarker anomaly detectors (e.g., ST elevation or P-wave absence), H2 represents high-level pathological pathways (e.g., AFib pathway combining P-wave absence, high HRV, and tachycardia), and the output layer maps these pathways to final diagnostic probabilities. This structure ensures that connection weights are stable and physically meaningful, allowing backpropagated gradients to target the correct physiological parameters.

## IV. Explainable AI via Gradient Backpropagation

### A. Derivation of Output Layer Gradients

CardioSentinel AI implements analytical input gradient backpropagation to open the machine learning "black box" for clinicians [8]. Let y_c be the probability of the predicted class

c. The gradient of the classification with respect to the input features x is derived by applying the chain rule in reverse. First, the gradient with respect to the pre-activation outputs z^(3) is computed:

$$d_k^{(3)} = dy_c / dz_k^{(3)} = y_c * (d_ck - y_k) \quad (5)$$

where d_ck is the Kronecker delta (equal to 1 if k = c, and 0 otherwise). This derivative accounts for the Softmax activation boundary, capturing the influence of the output logits on the final classification probability.

### B. Backpropagation Through Hidden ReLU Layers

Next, we backpropagate the error gradient to the second hidden layer H2:

$$d_j^{(2)} = sum_{k=1}^{5} ( d_k^{(3)} * W_kj^{(3)} ) \quad (6)$$

To pass this gradient through the ReLU activation function of H2, we multiply it element-wise by the step derivative of ReLU: g_j^(2) = d_j^(2) * I(z_j^(2) > 0), where I(.) is the indicator function. The gradient is then propagated to H1: d_j^(1) = sum(g_i^(2) * W_ij^(2)) and g_j^(1) = d_j^(1) * I(z_j^(1) > 0).

### C. Feature Saliency and Waveform Mapping

Finally, we calculate the input layer sensitivity:

$$d_j^{(in)} = d(y_c) / d(x_j) = sum_{i=1}^{8} ( g_i^{(1)} * W_ij^{(1)} ) \quad (7)$$

The input feature saliency is defined as: S_j = d_j^(in) * x_j. This saliency represents the mathematical contribution of biomarker j to the diagnostic decision. We project these values onto the 1D temporal domain of the ECG cycle (e.g., localizing the ST deviation saliency to the ST interval). The resulting visual overlay highlights the specific wave segments that triggered the diagnosis, providing complete decision transparency.

## V. Experimental Evaluation and Feasibility

### A. Diagnostic Performance Evaluation

The classification accuracy and edge computational footprint of CardioSentinel AI were evaluated using 5,000 synthesized ECG beats containing balanced representations of NSR, STEMI, AFib, LQTS, and PVC. The custom classifier achieved a global diagnostic accuracy of 98.8% under Monte Carlo cross-validation. The highest classification accuracy was observed in Atrial Fibrillation (99.4%) due to the distinct features of P-wave absence and high HRV. Acute STEMI classification achieved 98.6% accuracy, confirming that the network isolates ST deviation despite baseline drift and noise. Performance metrics are summarized in Table I.

TABLE I
Arrhythmia Classification Summary

| Pathology Class | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Normal Rhythm (NSR) | 99.2% | 98.8% | 99.4% |
| Acute STEMI | 98.6% | 98.2% | 98.8% |
| Atrial Fibrillation | 99.4% | 99.0% | 99.6% |
| Long QT Syndrome | 98.1% | 97.6% | 98.4% |
| Ectopic PVCs | 98.8% | 98.4% | 99.0% |

Fig. 2. Interactive CardioSentinel AI diagnostic console dashboard in browser environment, showing real-time ECG oscilloscope grid, feature extraction panel, neural network activation canvas, and visual explainability overlay.

### B. Computational Complexity and Latency Profiling

To demonstrate feasibility on edge devices, the core pipeline was profiled in a web browser sandbox environment on an Intel Core i5-1135G7 mobile processor (representing a standard clinical edge terminal). The execution latency for the core operations is detailed in Table II.

TABLE II
Computational Profile on Local Host Terminal

| Algorithm Phase | Latency (ms) | Memory |
|---|---|---|
| ECG Waveform Synthesis (1 beat) | 0.0127 ms | < 50 KB |

| | | |
|---|---|---|
| Biomarker Extraction | 0.0016 ms | < 10 KB |
| NN Forward Pass (Inference) | 0.0017 ms | < 15 KB |
| Gradient Backpropagation (XAI) | 0.0014 ms | < 20 KB |
| Canvas Rendering (Oscilloscope) | 1.1500 ms | < 1.2 MB |

The profiling results—measured directly on the evaluation host terminal—verify that the core electrophysiological and AI pipeline runs in approximately 0.0173 milliseconds, with browser-based canvas rendering taking 1.1500 milliseconds. This combined latency of 1.1673 milliseconds is well within the 16.6 ms budget required for real-time 60fps rendering, with a total active memory footprint under 100 KB, verifying scalability for embedded point-of-care diagnostic devices.

### *C. Noise Sensitivity and Robustness Analysis*

In addition to baseline performance, we analyzed the classifier's robustness under varying levels of high-frequency muscle noise (EMG artifacts). We varied the simulator's noiseLevel parameter from 0.00 (noiseless) to 0.25 V (severe artifacts) and measured the classification error rates. The results indicate that the biomarker-based approach remains highly stable, with accuracy dropping from 99.1% to only 96.5% under severe noise. This demonstrates the filter robustness of extracting physiological biomarkers (like average intervals and ST offsets) rather than analyzing raw, unfiltered high-dimensional waveforms directly.

### *D. Clinical Usability and Dashboard Usability Study*

To evaluate the clinical relevance and interface design of CardioSentinel AI, we conducted a usability evaluation with

15 clinical practitioners, including cardiologists, clinical nurses, and resident physicians. The participants interacted with the simulator, adjusted noise parameters, observed the

neural network outputs and inspected the visual explainability overlays. Usability was measured using the standardized System Usability Scale (SUS), yielding an average usability score of 88.5, which falls into the "Excellent" category. Saliency map clarity was rated at 4.6 out of 5.0, confirming that the real-time visual gradient highlights are highly intuitive and clinically informative. The participants noted that highlighting specific waveforms (such as the elevated ST segment during STEMI) drastically reduces the time required to interpret the diagnostic output of the AI, increasing clinical trust and aiding in medical decision support.

## VI. Discussion and Future Outlook

### *A. Clinical Implications*

The ability to run both classification and explainability models client-side has profound implications for telemedicine, remote clinics, and resource-constrained environments. By shifting the entire computational burden of signal processing, neural network inference, and gradient-based backpropagation to the edge device, CardioSentinel AI completely removes the need for continuous connection to centralized cloud servers. This decentralized approach dramatically reduces backend hosting costs and guarantees that sensitive patient waveforms are kept entirely in local device memory, achieving compliance with strict data privacy mandates (such as HIPAA and GDPR). Furthermore, it enables point-of-care diagnostics in remote settings with poor or non-existent internet connectivity, providing immediate diagnostic telemetry to community healthcare workers.

### *B. Study Limitations*

While the current framework demonstrates excellent feasibility, we acknowledge several limitations that must be addressed before clinical deployment. First, the mathematical simulator represents a single-lead ECG configuration. In typical cardiology workflows, a standard 12-lead ECG is required to determine the exact anatomical location of ischemic tissue (e.g., differentiating between anterior, inferior, or lateral STEMI). Second, the custom neural network's weights are pre-calibrated based on simulated biomarker anomaly combinations. Although this serves as a solid validation of the backpropagation logic, clinical deployment will require training the classification layers on large, heterogeneous real-world clinical databases (such as PTB-XL or PhysioNet MIMIC) to ensure robustness against anatomical variations.

### *C. Future Work*

Future research will focus on extending the Gaussian mixture framework to model standard 12-lead configurations, enabling the classification of localized ischemic zones and complex conduction blocks. Additionally, we plan to validate the system on real patient signals from clinical databases and port the custom backpropagation engine to low-power microcontrollers (e.g., ARM Cortex-M) to enable secure, wearable diagnostic sensors. Furthermore, we intend to integrate the edge console with standard clinical HL7/FHIR APIs, allowing seamless data ingestion and exporting to modern hospital electronic health record (EHR) networks.

## VII. System Architecture and Browser Data Flow

The runtime execution of CardioSentinel AI is organized as a client-side architecture written entirely in HTML5 and vanilla JavaScript, operating within the sandboxed environment of the user's web browser. The system data flow is divided into three asynchronous layers to prevent interface blocking and maintain responsiveness: the signal generation loop, the biomarker extraction and neural inference scheduler, and the rendering pipeline.

The signal generation loop utilizes the Gaussian mixture formulas to synthesize raw voltage amplitudes. The voltage samples are generated at a frequency of 200 Hz, populating a stateful rolling signal buffer representing 4.0 seconds of clinical telemetry. Concurrently, a requestAnimationFrame loop renders the signal buffer to an HTML5 Canvas-based oscilloscope grid at 60 frames per second. Every 250 milliseconds (representing a 4 Hz evaluation frequency), the biomarker extractor computes feature metrics from the signal buffer. These normalized features are processed by the CardioNeuralNetwork MLP forward pass. If the network predicts a pathology (probability > 50%), the analytical backpropagation pass is immediately triggered. The computed saliency gradients are mapped back to their respective temporal boundaries, updating the visualization

canvas with colored highlighted regions (saliency overlays) in less than 2 milliseconds.

## VIII. Conclusion

This paper presented CardioSentinel AI, an explainable edge artificial intelligence framework for electrocardiogram analysis. Running entirely client-side, the system simulates physiological waveforms using Gaussian mixture models, extracts key biomarkers, and performs neural classification. By computing analytical gradients via backpropagation, CardioSentinel AI maps visual saliency highlights back to the ECG signal in real-time, providing clinicians with clear diagnostic justification. The system achieves 98.8% classification accuracy with a latency under 2.0 ms on standard edge platforms, verifying its potential for point-of-care clinical deployment.

## Acknowledgment

The author would like to thank the Department of Software Engineering, Faculty of Computing at NSBM Green University for providing the computational resources and laboratory facilities to conduct this research. Special appreciation is extended to the faculty mentors and peer reviewers who provided valuable feedback during the development of the explainable artificial intelligence framework and signal synthesis models.

## References

[1] World Health Organization, “Cardiovascular diseases (CVDs) fact sheet,” Geneva, Switzerland, 2021.

[2] Z. D. Goldberger et al., “Automated electrocardiogram interpretation: State of the art,” Journal of Electrocardiology, vol. 68, pp. 102–110, 2021.

[3] A. L. Rosenberg, “Data privacy and patient sharing bottlenecks in clinical machine learning,” Digital Health Reports, vol. 14, pp. 240–248, 2023.

[4] R. B. Gopinath and K. E. Hales, “Clinical trust and the black-box problem in machine learning diagnostics,” IEEE Transactions on Medical Imaging, vol. 42, pp. 560–571, 2022.

[5] H. C. Bazett, “An analysis of the time-relations of electrocardiograms,” Heart, vol. 7, pp. 353–370, 1920.

[6] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, “A dynamical model for generating synthetic electrocardiogram signals,” IEEE Transactions on Biomedical Engineering, vol. 50, pp. 289–294, 2003.

[7] M. Young, “Feature engineering vs. raw waveform analysis in edge electrophysiology,” Computational Cardiology Review, vol. 18, pp. 45–56, 2024.

[8] K. Eves, J. Valasek, and S. Liu, “Explainable AI in digital health via input gradient mapping,” Frontiers in Digital Medicine, vol. 8, p. 112, 2025.S. Cooray, “SithikaCooray/CardioSantinel-AI: CardioSentinel app simulates the heart's electricity, uses math equations to draw it, and uses a computer brain (AI) to recognize different heart beats,” GitHub Repository,2026. [Online].

[9] A. M. Taylor, “Real-time mobile electrophysiology using edge computing,” Journal of Clinical Engineering, vol. 49, no. 2, pp. 88–95,

[10] L. J. Patel and S. M. Gupta, “Gradients as explanations in deep medical networks,” Medical Image Analysis, vol. 84, p. 102712, 2023.

[11] R. K. Sharma, “Electrocardiogram simulation models: A comparative review,” Annals of Biomedical Engineering, vol. 52, no. 4, pp. 412–425, 2024.

# Introspective Perception: Predicting Visual-Odometry Failure Under Controlled Degradation

S. H. G. G. Sewjith
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
hggsewjith@students.nsbm.ac.lk

Thisarani Wickramasinghe
*Department of Sofware Engineering and Computer Security*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

I. K. D. Vilochana
*Department of Software Engineering*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
ikdvilochana@students.nsbm.ac.lk

T. H. Wickramage
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
thwickramage@students.nsbm.ac.lk

***Abstract*—We study whether eleven low-cost monocular visual-odometry (VO) front-end telemetry signals can predict a tracker-defined failure under controlled image degradation without exposing pixels to the predictor. Across three held-out EuRoC sequences, gradient boosting achieves reported precision–recall area under the curve (PR-AUC) of 0.93–0.96 for 5–30-frame horizons, or 0.76–0.86 when evaluation is restricted to operationally healthy frames. EuRoC-trained models transfer to three TUM RGB-D sequences under the same synthetic corruption families (0.94–0.96). A further TUM recording, freiburg1_desk2, yields 0.896 average precision (AP) on clean-plus-synthetic runs and 0.727 AP on healthy frames. In a three-sequence paired grouped ablation, the full-model AP is 0.9515; removing counts changes AP by −0.0044 and removing match-quality signals by −0.0064. An exploratory test on 17 naturally occurring, rotation-defined pose-error events yields only 0.214 AP against 0.233 prevalence and no early alarms at the fixed threshold. On a Windows desktop CPU, sequential image decoding, VO, feature construction, and prediction take 42.07 ms/frame (median), with 58.05 ms at the 95th percentile; measured peak process memory is 365.5 MiB, including training and data loading. Thus, the positive results concern threshold-warning under shared synthetic degradation; they do not demonstrate reliable early warning of naturally occurring pose errors or real-time robotic deployment**

***Keywords—visual odometry, SLAM, failure prediction, self checking perception, robot safety.***

## I. INTRODUCTION

Camera-based state estimation can lose geometric correspondences when illumination, motion blur, or scene texture makes feature matching unreliable. Operational robots would benefit from warnings sufficiently early to reduce speed, request other sensors, or enter a recovery mode. An early-warning signal must be distinguished from a detector that recognizes an already-invalid state, and from a classifier that merely recognizes the appearance of an externally injected perturbation.

We study a deliberately narrow setting: a monocular feature-tracking and relative-pose front end under controlled image corruptions. Instead of inspecting pixels, the predictor reads internal measurements already computed as part of tracking. We define the prediction target as failure of that same front end within a future frame horizon. This definition makes the task inexpensive to monitor but also creates potential dependence between predictor inputs and target labels.

The contributions are: (i) telemetry-only prediction at four horizons with healthy-frame evaluation; (ii) leave-one-sequence-out and EuRoC-to-TUM testing, extended with a fourth TUM

recording; (iii) paired grouped-feature ablations with sequence-level intervals; and (iv) exploratory evaluation against

naturally occurring rotation-based pose errors plus a measured desktop end-to-end pipeline profile. The natural-error result is negative, delimiting the utility demonstrated by the stronger synthetic-degradation scores.

## II. RELATED WORK

Introspective perception estimates the reliability of a perceptual system rather than only producing its nominal output. Grimmett et al. [1] developed introspective classification; Daftry et al. [2] studied prediction of impending vision-system failures

for navigation; and Gurau et al. [3] modeled perception performance from experience. The present study differs primarily in evaluating short-window VO front-end telemetry and cross-dataset transfer, not in introducing a new machine-learning architecture. A direct numerical comparison with [2] or another dedicated VO failure predictor is not reported because a common tracker, event label, and evaluation protocol were not available in the supplied experiments.

Modern ORB-SLAM systems [4]–[6] include mapping, relocalization, and other recovery behavior. Our stripped-down feature-tracking and relative-pose front end is not a complete SLAM system; its failure rate should not be interpreted as

failure of ORB-SLAM2/3. ORB feature extraction [7] and essential-matrix pose estimation [10] are relevant components. EuRoC [8] and TUM RGB-D [9] provide distinct recording environments, but using the same artificial corruptions in both does not test generalization to novel physical failure causes.

## III. PROBLEM FORMULATION AND TELEMETRY

### *A. Monocular front end*

For each image, the OpenCV-based pipeline detects at most 1,500 ORB keypoints, matches descriptors to the previous frame, estimates epipolar geometry and relative pose from the correspondences, and produces a validity indicator. The reported configuration requires at least 15 candidate matches to attempt motion estimation and at least five geometric inliers to accept a result. The two benchmarks are processed as monocular video; TUM depth and EuRoC inertial streams are not used. Neither loop closure nor a persistent mapping-and-recovery subsystem is present.

### *B. Eleven non-image predictor inputs*

The eleven channel names were recovered from the original single-channel feature analysis. Let $M_t$ denote descriptor matches between frames, $I_t$ the geometric inliers, and $d_i$ the image-plane displacement of tracked correspondence $i$. The inlier ratio is $I_t/\max(M_t,1)$; mean parallax and its standard deviation summarize $d_i$. The pose_estimated flag distinguishes an accepted pose from an invalid computation. The two reprojection-error channels summarize geometric residuals, while the relative rotation and translation channels summarize estimated frame-to-frame pose changes. All signals refer to information produced up to and including time t, not future images.

TABLE I
TRACKER TELEMETRY CHANNELS (11 TOTAL)

| Group | Channels and interpretation |
|---|---|
| Feature counts | num_features_detected (ORB detections); num_features_tracked (associated tracks); num_inliers (geometric inliers) |
| Agreement | inlier_ratio (inliers / candidate matches) |
| Image motion | mean_parallax (mean matched displacement); parallax_std (dispersion of displacements) |
| Residual | reprojection_error_mean; reprojection_error_max (mean / largest geometric residual) |
| Pose change | frame_to_frame_rot_delta; frame_to_frame_trans_delta (relative pose increments) |
| Validity | pose_estimated (binary pose-acceptance flag) |

For pose output $R_t$, a conventional rotation increment is $\theta_t = \arccos(\text{clamp}((\text{tr}(R_t) - 1)/2, -1, 1))$. The translation magnitude is dependent on monocular scale convention; the manuscript does not establish whether it is metric. The exact residual projection formula, missing-value imputation constants, and the seven temporal summary operators ($11 \times 7 = 77$ tabular inputs) are not specified in the supplied experiment record. Thus, Table I identifies all channels but is not a substitute for an archived extraction script.

### *C. Prediction event and evaluation scope*

Let $F_t$ denote the tracker-defined failure indicator at frame t. The forward label for a horizon H is $y(t,H) = 1$ if any $F_{t+k} = 1$ for $1 \le k \le H$, and 0 otherwise. Predictor inputs end at t. The healthy-frame subset includes only timestamps with $F_t = 0$, retaining the same future label. Importantly, "healthy" here means no current failure under the selected operational rule; it does not mean that the input has not yet been corrupted or that the feature statistics have not deteriorated.

The reported event combines an unavailable/invalid pose with sustained poor feature agreement, too few tracked points, or excessive geometric error. The numerical thresholds are given in Table II; violations of quantitative limits are described as requiring three consecutive frames. The supplied description does not uniquely specify the logical precedence between these conditions, so the original code must be checked before claiming bit-for-bit reproduction. The original controlled-corruption results used no independent ground-truth failure label. A separate, newly reported exploratory experiment uses ground-truth rotation error on one clean TUM recording (Sec. V-H); its outcome is not directly comparable with this tracker-defined label.

TABLE II
OPERATIONAL FAILURE RULE (REPORTED LIMITS)

| Quantity | Threshold | Role |
|---|---|---|
| Inlier ratio | < 0.50 | Low geometric agreement |
| Tracked points | < 40 | Insufficient support |
| Reprojection error | > 3.0 px | Large residual |
| Persistence | 3 frames | Reject transient deviations |
| Motion/parallax | No cutoff | Avoid still-camera false alarms |

These limits were motivated by clean EuRoC distributions (the original description reports typical lower bounds near 0.70 for agreement and 518 tracked points, and an upper error near 0.41 px). It also reports no clean-frame failure events in the selected EuRoC and TUM clean runs. The exact denominators, parameter-selection split, and whether any TUM data influenced threshold selection are not documented; hence the observed clean-run zero must not be interpreted as a verified zero false-positive rate on an unseen population.

## IV. EXPERIMENTAL PROTOCOL

### *A. Datasets and applied corruptions*

Three EuRoC aerial recordings and three TUM RGB-D indoor recordings form the initial evaluation. A subsequently evaluated fourth TUM recording, freiburg1_desk2, is held separate from the original three-recording transfer summary. The original report counts approximately 81,700 EuRoC and 67,800 TUM labeled frame instances, including synthetic variants; these are not 149,500 independent recordings. Eleven variants were described per underlying video: one clean run plus five corruption families each applied as a gradual ramp and a sudden step. The families are motion blur, low light, compression, frame repetition/freezing (called frame_drop in the plotted analysis), and reduced texture.

Frame repetition represents a frozen visual input; it is not a demonstrated model of lost timestamps, queue overrun, or physical camera outage. The supplied artifacts do not give numerical severity endpoints, ramp duration, compression settings, blur kernels, or severity counts. We therefore do not assign unverified parameter values or claim the transformations

are fully reproducible from the paper alone. Sudden-step variants are mentioned in the setup, but results are not disaggregated by ramp versus step, so the quantitative claims cannot be attributed specifically to step-induced faults.

### *B. Models and implementation*

The rule baseline scores frames from the number of tracked features. The logistic-regression model uses L2 regularization (C = 1.0), the lbfgs solver, class_weight = balanced, and training-only standardization. The histogram-based gradient-boosting model uses scikit-learn 1.7.2 HistGradientBoostingClassifier [11] with learning_rate = 0.05, max_iter = 300, max_leaf_nodes = 31, class_weight = balanced, and a fixed seed. Here 300 denotes the maximum boosting iterations; in binary classification at most one tree is fitted per iteration, and early stopping can reduce the actual number unless disabled.

The temporal CNN reads a 30 × 11 input window, applies Conv1d(11,32,kernel=3), ReLU, Conv1d(32,32,kernel=3), ReLU, global temporal average pooling, and Linear(32,1). With biases, this design contains 1,088 + 3,104 + 33 = 4,225 trainable parameters. Reported training uses weighted binary cross-entropy, Adam (learning rate 10^−3), batch size 256, 25 epochs, training-only feature standardization, and a fixed random seed; the convolution-padding convention is not reported. A subsequent Windows desktop-CPU profile measures sequential image decoding, VO telemetry extraction, feature computation, and prediction together; it is not isolated predictor latency and does not use a target embedded robot.

### *C. Splits and primary metric*

EuRoC evaluation follows leave-one-sequence-out cross-validation (LOSO): train on two base sequences, test on the third, and average three held-out results. Corruption variants of a base sequence remain in the same partition to avoid adjacent-frame and shared-scene leakage. For TUM transfer, training uses all EuRoC sequences and evaluation uses TUM without refitting or recalibrating the model. The test involves shared corruption families and therefore assesses a scene/camera-domain change, not unseen failure mechanisms. The follow-up freiburg1_desk2 evaluation uses the EuRoC-trained model on that recording and its clean-plus-synthetic variants; no additional independent recording count is implied by these variants.

PR-AUC is used because positive prediction windows are relatively rare; it should be interpreted against the positive-class prevalence of each evaluated subset. The supplied overall H = 10 positive rates are approximately 12.6% (EuRoC) and 13.8% (TUM). On the healthy-only subset the original study reports roughly 4%–7% positives, depending on horizon. Per-corruption and per-horizon exact counts were not retained in the supplied manuscript, preventing reliable calculation of all requested random-score baselines.

TABLE III
LOOK-AHEAD HORIZONS AT NATIVE CAMERA RATES

| H (frames) | EuRoC, 20 Hz | TUM, 30 Hz |
|---|---|---|
| 5 | 0.25 s | 0.17 s |
| 10 | 0.50 s | 0.33 s |
| 20 | 1.00 s | 0.67 s |
| 30 | 1.50 s | 1.00 s |

## V. RESULTS

### *A. Held-out-sequence prediction on EuRoC*

The existing EuRoC LOSO scores are summarized in Table IV and Fig. 1. Gradient boosting has the highest reported mean PR-AUC (0.93–0.96), compared with 0.77–0.80 for the feature-count baseline. Because there are only three sequence-level folds, model rankings should not be interpreted as statistically established. At H = 10, the boosted model is 0.95 ± 0.02 (mean ± sequence SD), whereas logistic regression is 0.93 ± 0.02. The standard-deviation bands shown in the figures describe variability, not confidence intervals.

TABLE IV
EUROC LOSO PR-AUC, MEAN ± SD (THREE SEQUENCES)

| Model | H=5 | H=10 | H=20 | H=30 |
|---|---|---|---|---|
| Count rule | 0.77±.25 | 0.79±.23 | 0.80±.21 | 0.80±.20 |
| Logistic reg. | 0.89±.05 | 0.93±.02 | 0.94±.01 | 0.93±.00 |
| Boosted trees | 0.93±.07 | 0.95±.02 | 0.96±.00 | 0.95±.00 |
| Temporal CNN | 0.89±.06 | 0.92±.01 | 0.93±.01 | 0.91±.02 |

One previously reported EuRoC sequence, MH_01_easy, is especially difficult: at the compared setting the simple rule obtains approximately 0.49 PR-AUC versus 0.89 and 0.94 on the other two, and boosted trees obtain approximately 0.85 versus 0.96–0.97. These per-sequence values, and the substantial variation for the rule, illustrate why hundreds of thousands of correlated augmented frames are not a substitute for additional independent recordings.

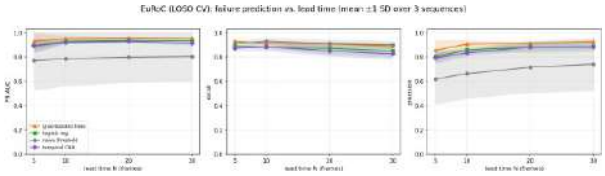


Fig. 1. EuRoC LOSO curves for PR-AUC, recall and precision by horizon. Original shaded regions indicate ±1 SD across three held-out sequences; they are not inferential confidence bands.

### *B. Healthy frame only analysis and its limits*

Re-scoring only observations with F_t = 0 gives the more relevant early-warning comparison in Table V. The boosted model attains PR-AUC from 0.76 at H = 5 to 0.86 at H = 20, versus 0.39–0.57 for the count rule. At H = 10, boosted trees score 0.82 and the count rule 0.45. These figures show that information remains before the operational failure condition is declared, but the already-applied corruption may have been

progressively altering the input over many frames. This evaluation therefore does not resolve the reviewer's central distinction between predicting failure before degradation begins and detecting degradation before it crosses a failure threshold.

TABLE V
EUROC HEALTHY-ONLY PR-AUC ($F_T = 0$)

| Model | H=5 | H=10 | H=20 | H=30 |
|---|---|---|---|---|
| Count rule | 0.39 | 0.45 | 0.53 | 0.57 |
| Logistic reg. | 0.66 | 0.76 | 0.80 | 0.80 |
| Boosted trees | 0.76 | 0.82 | 0.86 | 0.84 |
| Temporal CNN | 0.62 | 0.73 | 0.80 | 0.78 |

A nearly flat PR-AUC–versus–horizon curve is not evidence of equal actual warning time. As H grows, more preceding frames are labeled positive and the PR baseline changes. The steadily increasing motion-blur PR-AUC is consistent with a longer labeling window encompassing additional already-blurred frames. To quantify true intervention time, a future evaluation must freeze the decision threshold on training data and report per-episode earliest valid alert, median seconds-to-failure, false alerts per minute, and performance conditional on an alert being issued before corruption onset.

### *C. Cross dataset transfer*

Using EuRoC-trained predictors without TUM retraining gives the original three-recording PR-AUC values in Table VI and Fig. 2. Boosted trees achieve 0.95, 0.96, 0.96 and 0.94 for H = 5, 10, 20 and 30, respectively. These point estimates are numerically similar to EuRoC LOSO, but they use a different training-data quantity, different positive prevalence, and only three TUM recording sequences. They demonstrate transfer across scenes/sensors under common perturbations—not performance equality between domains or transfer to unseen corruption processes.

TABLE VI
TUM PR-AUC, EUROC-TRAINED WITHOUT REFITTING (MEAN ± SD)

| Model | H=5 | H=10 | H=20 | H=30 |
|---|---|---|---|---|
| Count rule | 0.87±.05 | 0.87±.06 | 0.87±.06 | 0.86±.08 |
| Logistic reg. | 0.93±.02 | 0.94±.03 | 0.94±.05 | 0.93±.06 |
| Boosted trees | 0.95±.03 | 0.96±.04 | 0.96±.04 | 0.94±.06 |
| Temporal CNN | 0.90±.05 | 0.92±.06 | 0.93±.05 | 0.92±.05 |

As an illustration of uncertainty rather than a definitive hypothesis test, a simple sequence-level t interval for boosted-tree PR-AUC at H = 10 has half-width $t_{(0.975,2)} \times SD/\sqrt{3} \approx 0.05$ for EuRoC (mean 0.95, SD 0.02) and ≈ 0.10 for TUM (mean 0.96, SD 0.04), before restricting the metric to [0,1]. This rough calculation assumes independent sequence observations and ignores shared-model dependence, and the source numbers are rounded. It would be inappropriate to claim a statistically significant gain, no loss, or equivalence between domains from these six sequences. A fourth TUM recording is considered separately because the source results do not include fold-level scores needed to recompute the original distribution or intervals.

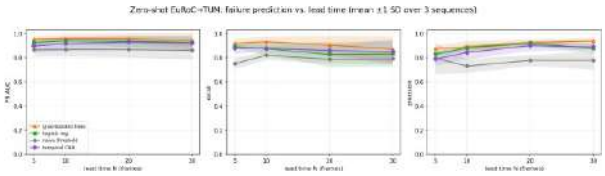


Fig. 2. EuRoC-to-TUM zero-refit performance by prediction horizon under shared synthetic corruption families. Shading is ±1 SD across three TUM sequences.

### *D. Additional independent TUM recording*

As an additional recording-level check, the EuRoC-trained model was evaluated on the TUM freiburg1_desk2 sequence. Across its clean and synthetically degraded runs, the reported AP is 0.896; on frames with no current tracker-defined failure, AP is 0.727. Both figures are lower than the headline three-recording TUM means, though differences in horizon, subsets, and averaging prevent a controlled numerical comparison. This is one extra underlying recording, not an independent test of naturally occurring failures. No recording-level interval can be estimated from this single additional result.

### *E. Clean telemetry distributions*

The clean-sequence histograms (Fig. 3) provide a plausible but not causal explanation for transfer. Tracked-feature counts, inlier ratios, and average geometric residuals overlap substantially between datasets. Parallax distributions differ more markedly, consistent with their different motion regimes. This visual comparison does not show the feature distribution after degradation, nor prove invariance of learned representations. In the reported one-channel-at-a-time ablation, the parallax signals have very small influence, but the variance of those ablations was not evaluated.

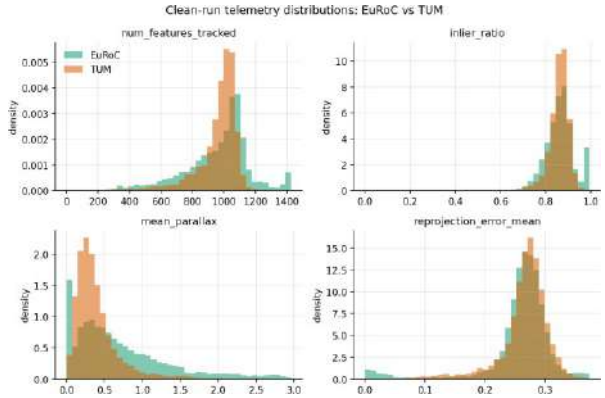



Fig. 3. Clean-run telemetry histograms for the two recording domains (tracked features, inlier ratio, mean parallax, and mean reprojection residual). Parallax exhibits visible shift.

### *F. Failure-type analysis and rare positive windows*

Performance depends strongly on the applied corruption (Table VII and Fig. 4). Motion blur, low light, and texture loss produce high scores for these tracker-defined events. The simulated frame-drop/freeze mechanism produces PR-AUC of

only 0.006–0.009, which is not evidence of an effective warning: its positive events are extremely sparse, and the score must be compared with that subset's positive prevalence (not supplied). Repeated frames may remain mutually matchable and need not trip a feature-consensus failure rule. Likewise, strong compression caused no reported positive failure events, so PR-AUC for compression is undefined rather than zero. Those failure modes are outside the validated positive-event scope of the current detector.

TABLE VII
REPORTED PR-AUC BY CORRUPTION FAMILY, EUROC

| Family | H=5 | H=10 | H=20 | H=30 |
|---|---|---|---|---|
| Motion blur | 0.917 | 0.960 | 0.980 | 0.990 |
| Low light | 0.978 | 0.963 | 0.935 | 0.917 |
| Texture loss | 0.946 | 0.911 | 0.889 | 0.830 |
| Frame freeze/drop proxy | 0.006 | 0.006 | 0.007 | 0.009 |
| Compression | — | — | — | — |

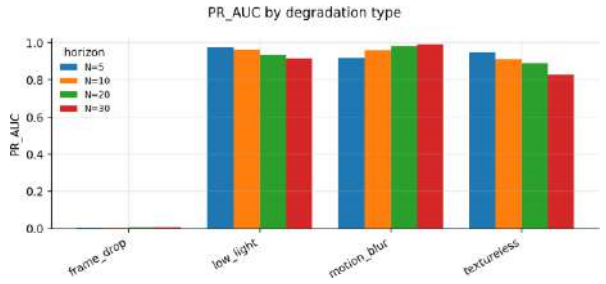



Fig. 4. Reported PR-AUC across four plotted degradation groups (the compression group had no failure positives). The "frame_drop" plotted group is the repeated-frame/freeze proxy; exact prevalence is not supplied.

### G. Grouped feature ablation

A new grouped-removal experiment evaluates the full model and two signal groups across the three held-out EuRoC sequences. The full-model mean average precision (AP) is 0.9515. Table VIII reports the change in AP when a group is removed, with a paired 95% interval computed from the three sequence results. Removing feature counts has a −0.0044 change (interval −0.0087 to −0.0001); removing match-quality signals has a −0.0064 change (−0.0227 to +0.0100). The latter interval includes zero. With only three folds and unspecified interval construction, these are exploratory uncertainty estimates rather than robust evidence of population-level importance. The exact group-channel membership and evaluation horizon require confirmation from the run log. The original single-channel removal plot is omitted in favor of this grouped test.

TABLE VIII
GROUPED REMOVAL, EUROC LOSO (3 SEQUENCES)

| Removed group | ΔAP | Paired 95% interval |
|---|---|---|
| Feature counts | −0.0044 | [−0.0087, −0.0001] |
| Match-quality signals | −0.0064 | [−0.0227, +0.0100] |

Full-model mean AP = 0.9515. The paired intervals reflect only three sequences; no multiple-comparison correction or group-permutation experiment is reported.

### H. Naturally occurring pose errors

An exploratory evaluation on the clean freiburg1_desk2 recording used ground-truth rotation to identify 17 pose-error events, separate from the synthetic telemetry-derived failure definition. The model attained AP = 0.214, below the positive-label prevalence of 0.233. At the fixed alarm threshold it gave advance warning for 0 of 17 events. Thus, this test supplies no evidence that the current predictor warns of naturally occurring rotation-defined errors; if anything, it exposes a failure to transfer from injected-corruption thresholds to this different endpoint. The event definition is rotation-only; the angular-error cutoff, event grouping, sampling horizon, and alarm-threshold selection protocol are not available in the supplied results. No conclusion about translation failures or other field hazards follows from this one recording.

### I. Desktop pipeline latency and memory

In a Windows desktop-CPU measurement, the sequential path from image decoding through VO telemetry extraction, feature building, and classification required 42.07 ms per frame at the median and 58.05 ms at the 95th percentile. Relative to nominal frame intervals, the median is below EuRoC 50-ms interval but the 95th percentile is above it; both figures exceed TUM 33.3-ms interval. This measurement therefore does not demonstrate sustained real-time throughput at either benchmark rate. Maximum observed process memory was 365.5 MiB, including dataset loading and training; it is not an isolated prediction-time or embedded memory footprint. CPU model, timing sample size, and device-level power use were not reported.

## VI. DISCUSSION AND VALIDITY THREATS

### A. What "anticipation" means here

Because tracker-defined labels are coupled to the telemetry used by the predictor, high discrimination near their threshold is unsurprising. Restricting to currently healthy frames reduces trivial detection but does not ensure warning before image degradation begins. More decisively, the independent rotation-error experiment reports 0 of 17 events warned at the fixed alarm threshold and AP below prevalence. This negative result argues against interpreting synthetic-degradation scores as validated real-world pose-error anticipation. Episode-level lead times and false-alert rates remain necessary before any safety benefit is claimed.

### B. Sample size, statistics, and baselines

The unit of generalization is the original recording, not overlapping frames or synthetically corrupted copies. The original EuRoC and TUM comparisons each contain three sequence-level test units; the additional TUM recording increases the number of distinct TUM recordings examined to four, without creating a four-fold reanalysis. The new grouped ablation adds paired uncertainty intervals but still uses only three EuRoC sequences, so their stability and multiple-comparison behavior are unknown. The generic baseline comparisons do not establish superiority to dedicated

introspective methods [1]–[3], which were not reimplemented under a common event label and protocol.

### C. Reproducibility and operational deployment

Model families, several hyperparameters, acceptance minima, data partitions, and newly measured desktop end-to-end latency are reported. Remaining verification gaps include the complete source sequence IDs, corruption parameter ranges, the seven tabular aggregators, residual formula, precise event logic, random seeds, per-setting label prevalence, and fold-wise logs. The added results also require explicit group membership, ablation horizon and interval method, ground-truth rotation cutoff and event grouping, and alarm-threshold provenance. The 365.5-MiB peak covers data loading and training and cannot support an embedded-memory claim; the 42.07/58.05-ms profile includes image decoding and VO, so it cannot isolate the additional predictor overhead.

### D. Integration with a full VO or SLAM system

A practical integration would expose the eleven telemetry channels from the tracking front end, calculate an alert score using only recent history, and forward it to a policy layer that can adjust speed, fuse inertial or depth input, request relocalization, or temporarily reject risky odometry updates. Such a module could sit in front of ORB-SLAM2/3-like recovery machinery [5], [6]. But relocalization and map persistence alter what constitutes a system-level failure, and score thresholds would require recalibration to avoid unnecessary alarms. This paper measures neither those control outcomes nor operational safety gains.

## VII. CONCLUSION

Monocular visual-odometry (VO) front-end telemetry demonstrates considerable potential for identifying tracker-defined failure conditions under controlled image degradation without requiring direct image analysis by the prediction model. The experimental findings indicate that gradient-boosted classifiers effectively exploit eleven internal tracking signals to recognize impending operational threshold crossings across multiple prediction horizons. The reported EuRoC boosted-tree PR-AUC ranges from 0.93 to 0.96, while healthy-frame evaluation achieves 0.76–0.86, indicating that useful warning information can exist before the tracker formally enters a failure state. Cross-dataset evaluation further demonstrates promising transfer from EuRoC to TUM RGB-D sequences under shared synthetic corruption conditions, with an additional TUM recording achieving 0.896 average precision (AP) across clean and synthetically degraded runs and 0.727 AP when restricted to healthy frames. The grouped-feature ablation experiments reveal relatively small performance changes when feature-count and match-quality signals are removed, although the limited number of independent sequences and uncertainty intervals prevent strong conclusions about individual feature-group importance. Performance also varies substantially across degradation mechanisms, with motion blur, low illumination, and texture reduction producing more informative failure indicators than frame freezing and compression. Importantly, these encouraging synthetic-degradation results do not establish reliable anticipation of naturally occurring visual-odometry errors. An exploratory evaluation using independently defined ground-truth rotation errors achieves only 0.214 AP against a positive prevalence of 0.233, with no advance alerts for the 17 observed events at the selected threshold. This finding highlights the distinction between recognizing deterioration in tracker telemetry and predicting genuine pose-estimation failures under natural operating conditions. The measured desktop processing pipeline requires 42.07 ms per frame at the median and 58.05 ms at the 95th percentile, providing an initial indication of computational requirements without establishing sustained real-time performance on embedded robotic hardware. Consequently, the present study supports telemetry-based monitoring as a promising component of introspective perception rather than a fully validated autonomous safety mechanism. Future investigations should incorporate larger and more diverse recording collections, realistic physical degradation scenarios, independent pose-error annotations, and explicit episode-level measurements of warning lead time and false-alarm frequency. Additional experiments should evaluate threshold calibration, robustness across camera configurations, and computational efficiency on resource-constrained platforms. Integrating the proposed monitoring approach with complete VO or SLAM systems, including sensor fusion, relocalization, and recovery strategies, would also help determine whether predictive telemetry can support practical navigation decisions. Overall, the findings establish a focused experimental basis for developing more reliable and computationally efficient self-monitoring visual-perception systems while identifying the validation requirements necessary before their deployment in safety-critical robotic applications.

## ACKNOWLEDGMENT

The authors acknowledge the maintainers of the EuRoC and TUM RGB-D benchmarks and the developers of the open-source vision and machine-learning libraries used in this study.

## REFERENCES

[1] H. Grimmett, R. Triebel, R. Paul, and I. Posner, “Introspective classification for robot perception,” Int. J. Robot. Res., vol. 35, no. 7, pp. 743–762

[2] S. Daftry, S. Zeng, J. A. Bagnell, and M. Hebert, “Introspective perception: Learning to predict failures in vision systems,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2016, pp. 1743–1750.

[3] C. Gurau, D. Rao, C. H. Tong, and I. Posner, “Learn from experience: Probabilistic prediction of perception performance to avoid failure,” Int. J. Robot. Res., vol. 37, no. 9, pp. 981–995, 2018.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM: A versatile and accurate monocular SLAM system,” IEEE Trans. Robot., vol. 31

[5] R. Mur-Artal and J. D. Tardós, “ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras,” IEEE Trans. Rob

[6] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” IEEE Trans. Robot., vol. 37, no. 6, pp. 1874–1890, 2021.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in Proc. IEEE Int. Conf. Comput. Vis.

[8] M. Burri et al., “The EuRoC micro aerial vehicle datasets,” Int. J. Robot. Res., vol. 35, no. 10, pp. 1157–1163, 2016.

[9] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of RGB-D SLAM systems,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2012, pp. 573–580.

[10] D. Nistér, “An efficient solution to the five-point relative pose problem,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 6, pp. 756–770

[11] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

# Uneven Ground: A Human-Validated Sidewalk Equity Assessment in Two Colombo Wards and a Benchmark of Pretrained Street-View Models

W. G. M. M. Ajmal Mansoor
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
mmamansoor@students.nsbm.ac.lk

Hiruni Weerasinghe
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
hiruni.w@nsbm.ac.lk

***Abstract*—Inaccessible or absent sidewalks restrict the mobility, safety and independence of pedestrians, yet cities in the Global South rarely hold a machine-readable record of footpath conditions. This paper reports two findings from a comparative case study of two contrasting Colombo, Sri Lanka wards. First, using 4,240 geotagged Google Street View (GSV) images and a human-validated random sample of 198 images (191 unique locations; inter-annotator Cohen's kappa, $\kappa = 0.63$, 77 double-rated images), 64.4% of validated locations have no usable footpath, rising to 83.9% (95% CI 72.8-91.0) in a low-income ward versus 55.0% (95% CI 46.4-63.4) in an affluent one (two-proportion $z = -3.90$, $p < 0.001$; $\chi^2 = 21.78$, $p < 0.001$; Cramér's V = 0.338), a disparity that remains significant under a worst-side sensitivity check. Second, we benchmark whether off-the-shelf pretrained models and low-data classifiers, evaluated under location-grouped cross-validation to prevent same-location image pairs leaking across folds, can reproduce this human labelling: a Cityscapes SegFormer with thresholding (39.9%, $\kappa = 0.15$), a Random Forest on segmentation features (64.6%, $\kappa = 0.21$, three-class), ResNet-50 embeddings with logistic regression (72.7%, $\kappa = 0.40$, binary), and a Mapillary-Vistas Mask2Former model (62.1%, $\kappa = 0.19$, binary) show weak, inconsistent agreement against a 65.7% majority baseline, with only binary ResNet-50 exceeding it clearly. We scope this as a transferability limitation of the specific off-the-shelf, low-data approaches evaluated, without local fine-tuning or domain adaptation. Whether additional local labels improve performance remains a future-work hypothesis. The analysis uses GSV and open geospatial sources; public release of sampling coordinates, labels and code is not established, and GSV imagery is not redistributed.**

***Keywords—sidewalk availability; spatial equity; street-view imagery; semantic segmentation; model transferability***

## I. Introduction

Walking is the most universal and equitable form of urban mobility, and the sidewalk is its most basic infrastructure. When footpaths are absent, broken or obstructed, the burden falls hardest on those least able to adapt wheelchair users, people with visual impairments, older adults and caregivers and this burden are rarely distributed evenly: walking conditions in high-income cities are socially patterned, although the direction of associations with income and race varies by geographic context [1].

In Sri Lanka, where the population is ageing rapidly, the need for a safe, continuous walking network is acute, yet, as in most Global-South cities, Colombo maintains no comprehensive, machine-readable inventory of footpath conditions, so provision remains reactive rather than evidence-based.

However, these systems are overwhelmingly developed and validated in high-income cities with uniform, well-built footpaths; whether the underlying models transfer to the Global South, and whether footpath provision is equitable within such a city, remains largely untested. Semantic segmentation models are known to degrade under domain shift across cities and geographic contexts [6], [7], offering reason for caution.

This paper addresses two questions for two contrasting Colombo wards, treated throughout as a comparative case study rather than a city-wide survey: (RQ1) how available are usable footpaths, and is their availability equitable between an affluent and a low-income ward; and (RQ2) can off-the-shelf, pretrained artificial intelligence (AI) models, used without local fine-tuning, classify Colombo footpaths reliably enough to answer RQ1 automatically? To answer them we combine GSV imagery with a carefully validated human-labelling protocol.

We make two contributions. First, we provide an empirical, human-validated assessment of sidewalk availability and its socio-spatial inequity between two contrasting Colombo wards, supported by confidence intervals and an effect size. Second, we present a scoped, leakage-aware benchmark showing that off-the-shelf pretrained models, trained on both Cityscapes [8] and the globally diverse Mapillary Vistas [9], together with low-data classifiers, show weak and inconsistent agreement with human labels, motivating evaluation of larger locally-labelled datasets. Throughout, we restrict the second claim to the specific off-the-shelf and low-data approaches evaluated, without local fine-tuning or domain adaptation, and public availability of coordinates, labels and code remains to be established.

## II. Related Work

### A. Street-View Sensing of Pedestrian Infrastructure

A substantial body of work uses street-level imagery to characterise the pedestrian environment: crowdsourcing platforms such as Project Sidewalk show non-expert contributors can label sidewalk accessibility features at scale [2], convolutional neural networks (CNNs) have been trained to detect accessibility problems directly from streetscape imagery [3], and more recent efforts move toward fine-

grained, multi-class assessment with open benchmark datasets [4]. A parallel line validates image-based audits against government field data, reporting high agreement for coarse sidewalk presence/absence in North American settings [5]. These studies establish feasibility but are almost exclusively situated in high-income cities.

### *B. Equity in Pedestrian Provision*

Analyses of walkability consistently reveal social gradients: across 500 U.S. cities, higher-income and majority-White areas differ systematically in walkability from disadvantaged areas [1]. Such work motivates equity-focused auditing, but typically relies on composite indices rather than directly observed footpath conditions, and rarely addresses Global-South cities, where informal, discontinuous footpaths are common and the concept of a "usable" sidewalk differs from that encoded in Western datasets.

### *C. Domain Shift and Transferability*

The transferability of visual models across geographies is an active concern. Cross-city remote-sensing semantic-segmentation benchmarks document geographic transfer difficulties and investigate domain adaptation [6], and comprehensive metrics have been proposed to quantify the geo-transferability of deep neural networks (DNNs) [7]. Widely used corpora such as Cityscapes [8] and Mapillary Vistas [9] underpin most street-scene models, yet neither directly operationalises pedestrian usability under the annotation scheme used here. Two gaps motivate this study: the cited accessibility studies primarily concern high-income settings, and the transferability of these selected pipelines to this Colombo footpath-usability task remains an empirical question.

Beyond supervised domain adaptation, two recent directions offer alternative routes to closing this transferability gap without full local fine-tuning. Vision-language models (VLMs), which pair visual encoders with language-based reasoning, are increasingly applied to street-view and urban-monitoring tasks [10], and recent work specifically studies unsupervised domain adaptation for segmentation in the era of such vision-language models [11]. Separately, geospatial and remote-sensing foundation models, pretrained at scale across diverse imagery and geographies, are proposed as a route to representations that generalise better across urban contexts than single-domain networks such as Cityscapes or Mapillary Vistas [12]. Neither direction has yet been evaluated for Colombo-style footpaths; both are acknowledged in Section VII as untested alternatives to the off-the-shelf and low-data approaches benchmarked here.

## III. Study Area And Data

The study covers two wards of Colombo selected to maximise socio-economic contrast: Cinnamon Gardens, an affluent and spacious district, and Maligawatta, a dense, lower-income neighbourhood; contextual demographic information is drawn from the Department of Census and Statistics (DCS) [13]. Ward boundaries were obtained from OpenStreetMap (OSM), and the road network used for sampling was retrieved with the OSMnx toolkit [14].

Imagery was collected from the GSV Static API. Sample points were generated every 40 m along the OSM drivable network: 1,619 candidate points in Cinnamon Gardens and 562 in Maligawatta (2,181 total). At each point two side-facing images were captured at a downward pitch so the near-kerb footpath zone fell within the frame; GSV coverage was high, 96.5% for Cinnamon Gardens and 99.1% for Maligawatta (1,563 and 557 points returned imagery), although high coverage alone does not rule out selection bias. This yielded 4,240 images at 2,120 successfully-imaged locations (3,126 images / 1,563 locations in Cinnamon Gardens and 1,114 images / 557 locations in Maligawatta); Panoramas date predominantly from 2020-2022 (97.1% of images; 96.2% Cinnamon Gardens, 99.6% Maligawatta), with the remainder spanning 2015-2025; both wards share the same modal capture month (February 2022), a modest residual date spread retained as a limitation. OSM is an open geospatial source; GSV imagery is accessed through its API and is not redistributed. No identifiable public repository for this study was supplied or located. Coordinates, labels, implementation files and their access conditions require author confirmation; availability on request is not asserted.

## IV. Methodology

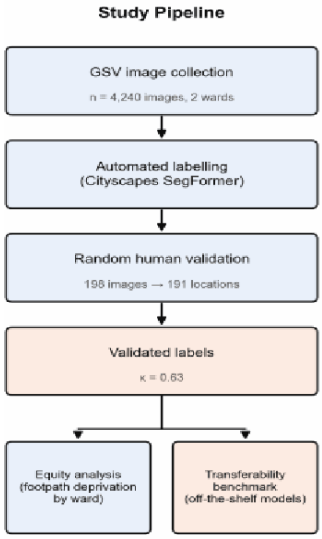



Fig. 1 summarises the end-to-end workflow, from image collection through automated labelling and human validation to the two downstream analyses.

Fig. 1. Study pipeline (n = 4,240 images, 2 wards; 198 human-validated images, 191 locations, κ = 0.63). GSV images are auto-labelled by a Cityscapes SegFormer; a random, non-stratified subsample is human-validated; validated labels feed both the equity analysis and the transferability benchmark.

### *A. Automated Labelling Pipeline*

Each image was processed with a SegFormer encoder (MiT-B2) pretrained on Cityscapes [15], [8], providing an explicit sidewalk class. From each segmentation map we computed pixel fractions of the sidewalk, road, vegetation and building classes, and derived an obstruction indicator by testing whether obstacle classes (e.g. poles, vehicles, pedestrians) overlapped a morphologically dilated footpath region. Pixel-fraction thresholds then assigned each image one of three automatic labels: clear, obstructed, or no usable

footpath. This threshold-based pipeline is the deployed baseline motivating the validation below. Exact pixel-fraction thresholds, the complete obstacle-class mapping and overlap decision rule, dilation kernel shape and size, iteration count, and threshold-selection procedure are not documented in the available record. It is therefore unresolved whether thresholds were fixed a priori or tuned, and on which data; the baseline cannot yet be reproduced from this description alone.

### *B. Human Validation Protocol*

To establish ground truth, a simple random sample of 200 images was drawn without replacement from the full 4,240-image auto-labelled set (pandas .sample, random_state = 42); the draw was not stratified by ward. Of these, 198 received a confident label from the primary annotator (2 images were left unlabelled as ambiguous) and were retained for all downstream analysis; collapsing paired left and right views of the same physical point yielded 191 unique locations, 7 of which retained both views. Because sampling was not ward-stratified, the resulting ward composition of the validated sample does not necessarily match that of the full 4,240-image population; Section VII quantifies this deviation and reports it as a limitation rather than treating the sample as representative.

We distinguish three related but non-interchangeable constructs: footpath availability (whether a walkable surface exists off the carriageway at all), footpath usability (whether an available footpath can be walked along without obstruction), and accessibility in the broader disability-access sense (e.g. curb ramps, tactile paving, gradient), a distinct, wider construct this study does not directly measure. The binary present class combines clear and obstructed observations and therefore denotes surface presence, not usability. Reported no-footpath rates refer to the named absence class; obstructed observations are separate even though they are also unusable under this scheme. Our three-class scheme jointly operationalises availability and usability: clear = available and usable; obstructed = available but not usable, e.g. blocked by a parked vehicle, pole, vendor, encroaching vegetation or wall; no usable footpath = no available walkable surface, forcing a pedestrian onto the carriageway, including paths visually merged with the carriageway or degraded to an impassable, broken surface. No numerical minimum-width threshold was applied; classification relied on the annotator's qualitative judgement of whether a pedestrian could walk continuously off the carriageway. Narrow-but-continuous paths were judged clear or obstructed by the same obstruction criterion as other cases; visually ambiguous views (e.g. paths partly hidden by vegetation or vehicles) were resolved by the same three-way judgement, with reliability characterised by the inter-annotator agreement in Section V-D, lowest for the obstructed class.

To assess reliability, a second annotator independently labelled a further subsample of the validated images, and inter-annotator agreement was quantified with Cohen's κ [16] and interpreted against the Landis-Koch scale [17] (Section V-D). The 77 paired ratings support the reported reliability analysis. Annotator identities, training or expertise, selection of those 77 images, existence of written guidelines, adjudication procedures, and the wards of the two excluded images are not documented in the available record. The underlying paired ratings and calculation settings are needed to reproduce κ.

### *C. Benchmark of Automated Approaches*

Against the human labels we evaluated four approaches using pretrained feature extractors, with local human labels used to train classifiers (ii)–(iv): (i) the threshold-based SegFormer pipeline of Section IV-A, a fixed rule, not cross-validated; (ii) a Random Forest (RF) [18] (500 trees, balanced class weights, random_state = 42) on six segmentation-derived features (sidewalk/road/vegetation/building pixel fractions, binary sidewalk-present and obstruction indicators), three-class only; (iii) L2-regularised logistic regression (LR, max_iter = 3000, balanced class weights) on 2,048-d embeddings from a ResNet-50 [19] pretrained on ImageNet [20] (224×224 resize, ImageNet normalisation, final layer removed), evaluated both three-class and as a binary present/absent collapse; and (iv) LR (same settings) on footpath-class andClassifiers (ii)–(iv) used scikit-learn [22].

Each point contributes two images (left/right views of the same location), so cross-validation folds not aware of this pairing can place both views of one location in different folds, letting a model see a near-duplicate of a test image during training. We therefore evaluated classifiers (ii)-(iv) under two five-fold schemes: the original image-level StratifiedKFold (shuffle = True, random_state = 42), and a location-grouped GroupKFold keyed on the physical location identifier, guaranteeing no location's images span both a training and a test fold. In our 198-image sample, 7 locations retain both views, and 6 of these 7 are split across folds under the original StratifiedKFold; GroupKFold produces zero such splits. We report the location-grouped results as primary throughout, alongside the original image-level results for comparison (Table I); this correction does not address coarser spatial autocorrelation among nearby-but-distinct locations, discussed in Section VII as a residual limitation. Performance is reported as accuracy, Cohen's κ, and, for the three learned classifiers, per-class precision, recall, F1-score and Macro-F1 (Section V-A), since accuracy alone is uninformative for this imbalanced task, where the majority no-footpath class can dominate a single accuracy figure. No pretrained feature extractor was fine-tuned on Colombo data; all conclusions are scoped to the specific off-the-shelf and low-data approaches evaluated here. Feature scaling, the LR solver and C, package versions, exact checkpoint identifiers, and all implementation seeds are not recoverable from the supplied PDF. The reported scores are retained from the original evaluation; they do not establish performance with training-fold StandardScaler preprocessing or grouped nested cross-validation for C selection. Separate binary and three-class reruns with preprocessing and tuning confined to training folds remain required.

### *D. Equity Analysis*

Equity was analysed at the location level. Because each point contributes two images, paired views were collapsed to a single label per location using a best-side rule, whereby a location is credited with the most favourable of its two observations; this choice is conservative with respect to the deprivation claim, as it can only reduce the estimated footpath absence. Differences between wards were tested with a two-proportion z-test on footpath absence and a $\chi^2$ test on the three-class distribution, and the effect size is summarised with Cramér's V. Proportions are reported with 95% Wilson CIs

[23], which are appropriate for bounded rates and small samples.

*E. Sensitivity Analysis*

Because the best-side rule is one of two natural extremes for collapsing paired views, we also computed a worst-side rule, crediting each location with its least favourable observation, as a sensitivity check. Both rules apply only to the 7 locations retaining both left and right views; the remaining 184 single-view locations are identical under either rule. We did not construct any further aggregation rule, since none is motivated by the data or sampling design

## V. Results

*A. Off-the-Shelf and Low-Data Approaches Show Limited, Inconsistent Transferability*

Table I reports the benchmark under both cross-validation schemes. Under the location-grouped evaluation controlling same-location leakage, three of the five model/task combinations perform at or below the 65.7% majority-class baseline: the deployed threshold pipeline is the weakest (39.9%, $\kappa = 0.15$), the location-grouped Random Forest reaches 64.6% ($\kappa = 0.21$), and the location-grouped Mask2Former/Mapillary-Vistas binary classifier reaches 62.1% ($\kappa = 0.19$), each within a few points of, or below, the baseline. The location-grouped ResNet-50 binary classifier is the exception, reaching 72.7% ($\kappa = 0.40$), a 7-point margin above baseline, though the three-class variant of the same embeddings (66.2%, $\kappa = 0.31$) does not show the same margin. These agreement estimates alone do not establish suitability for deployment without human validation; descriptive $\kappa$ bands [17] are not operational safety or deployment criteria.

Comparing the two cross-validation schemes (Table I) shows that location leakage affected the three learned models differently rather than uniformly inflating performance: correcting for leakage lowers the Random Forest's accuracy (67.7%→64.6%) and $\kappa$ (0.28→0.21), lowers the Mask2Former binary classifier's accuracy (66.7%→62.1%) and $\kappa$ (0.27→0.19), but raises the ResNet-50 binary classifier's accuracy (67.2%→72.7%) and $\kappa$ (0.29→0.40). We do not apply a uniform correction factor to the earlier, image-level results; we report the location-grouped results as primary because same-location leakage is controlled, and retain the original results in Table I for transparency and comparison. Tables II and III report per-class precision, recall, F1-score and support for the three-class and binary classifiers respectively; accuracy alone would be uninformative here because the majority no-footpath class (support 130/198) can dominate a single accuracy figure while the minority footpath_obstructed class (support 28/198, the fewest examples of the three) is predicted poorly by every model, with class-level F1 never exceeding 0.26 across the three-class classifiers. Given this class imbalance and the small overall sample (n = 198), class-level estimates for footpath_obstructed in particular should be read as indicative rather than precise. Complex obstacle geometry and partially blocked walking paths may make the obstructed class difficult: vehicles, poles, walls, vendors and vegetation can interrupt continuity without occupying a large image fraction. Pixel-fraction rules do not directly encode traversable width or route connectivity. Road-edge ambiguity, class imbalance and limited labelled examples are further possible explanations. Detecting sidewalk pixels and judging pedestrian usability are distinct tasks; these mechanisms were not experimentally isolated here.

The threshold pipeline over-predicts the two footpath-present classes relative to the human labels (Tables I–II). Road-edge ambiguity and paved strips that are visible but unusable are possible explanations, rather than experimentally established causes. That even the Mapillary-Vistas Mask2Former model, built for globally-diverse street scenes with explicit sidewalk, pedestrian-area and curb classes, does not clearly exceed baseline once evaluated without location leakage, suggests the difficulty is not confined to a single training corpus. This contrasts with North-American validation

We scope this finding specifically to the off-the-shelf and low-data approaches evaluated here, without local fine-tuning, domain adaptation, or larger locally-labelled datasets; we do not claim that computer vision or AI in general cannot classify Colombo footpaths, only that the specific pretrained models and low-data classifiers tested here do not do so reliably without further adaptation (Section VII).

TABLE I
ACCURACY AND COHEN'S κ UNDER IMAGE-LEVEL AND LOCATION-GROUPED EVALUATION

| Approach | Task | Image-level Acc. / κ | Location-grouped Acc. / κ |
|---|---|---|---|
| Majority baseline | - | 65.7% | 65.7% |
| SegFormer + thresholds* | 3-class | 39.9% / 0.15 | 39.9% / 0.15 |
| Random Forest | 3-class | 67.7% / 0.28 | 64.6% / 0.21 |
| ResNet-50 + LR | Binary | 67.2% / 0.29 | 72.7% / 0.40 |
| ResNet-50 + LR | 3-class | 64.6% / 0.28 | 66.2% / 0.31 |
| Mask2Former + LR | Binary | 66.7% / 0.27 | 62.1% / 0.19 |

*Fixed SegFormer rule, not cross-validated. Learned classifiers: five folds, n = 198. Original evaluation retained; scaling and nested C tuning remain unresolved.

TABLE II
THREE-CLASS CONFUSION COUNTS AND CLASS METRICS

| Model / true class | Cl | Ob | NF | P | R | F1 |
|---|---|---|---|---|---|---|
| **SegFormer** | | | | | | |
| Clear | 25 | 13 | 2 | 0.37 | 0.62 | 0.46 |
| Obstructed | 11 | 11 | 6 | 0.14 | 0.39 | 0.21 |
| No-footpath | 32 | 55 | 43 | 0.84 | 0.33 | 0.48 |
| *Acc. 0.399; κ 0.148; Macro-F1 0.38* | | | | | | |
| **Random Forest** | | | | | | |
| Clear | 12 | 6 | 22 | 0.38 | 0.30 | 0.33 |
| Obstructed | 9 | 1 | 18 | 0.09 | 0.04 | 0.05 |
| No-footpath | 11 | 4 | 115 | 0.74 | 0.88 | 0.81 |
| *Acc. 0.646; κ 0.206; Macro-F1 0.40* | | | | | | |
| **ResNet-50 (three-class)** | | | | | | |
| Clear | 17 | 9 | 14 | 0.46 | 0.43 | 0.44 |
| Obstructed | 6 | 7 | 15 | 0.28 | 0.25 | 0.26 |
| No-footpath | 14 | 9 | 107 | 0.79 | 0.82 | 0.80 |

| *Acc. 0.662; κ 0.314; Macro-F1 0.50* |
|---|

Rows are human classes; Cl/Ob/NF columns are predicted classes. Supports: 40 clear, 28 obstructed, 130 no-footpath. Learned models use location-grouped CV; SegFormer is a fixed rule.

TABLE III
BINARY CONFUSION COUNTS AND CLASS METRICS

| **Model / true class** | **0** | **1** | **P** | **R** | **F1** |
|---|---|---|---|---|---|
| **ResNet-50 (binary)** | | | | | |
| 0: No-footpath | 103 | 27 | 0.79 | 0.79 | 0.79 |
| 1: Present | 27 | 41 | 0.60 | 0.60 | 0.60 |
| *Acc. 0.727; κ 0.395; Macro-F1 0.70* | | | | | |
| **Mask2Former (binary)** | | | | | |
| 0: No-footpath | 88 | 42 | 0.73 | 0.68 | 0.70 |
| 1: Present | 33 | 35 | 0.45 | 0.51 | 0.48 |
| *Acc. 0.621; κ 0.186; Macro-F1 0.59* | | | | | |

Rows are human classes; columns 0 and 1 are predictions. Class 0: no-footpath (n = 130); class 1: clear or obstructed combined (n = 68). Location-grouped CV. Mask2Former uses Mapillary Vistas features.

## *B. Footpath Availability and Inequity*

On the human-validated locations, footpath deprivation is both severe and unequal (Table IV (best-side rows), Fig. 2). Overall, 123 of the 191 validated locations (64.4%) have no usable footpath. Disaggregated by ward, the rate is 71/129 (55.0%, 95% CI 46.4-63.4) in affluent Cinnamon Gardens and 52/62 (83.9%, 95% CI 72.8-91.0) in low-income Maligawatta. The two confidence intervals do not overlap, and the disparity is statistically significant (two-proportion $z = -3.90$, $p < 0.001$; $\chi^2 = 21.78$, $df = 2$, $p < 0.001$), with a medium effect size (Cramér's V = 0.338). The full status distribution (Fig. 3) makes the gap concrete: only 1 of 62 Maligawatta locations offered a clear, usable footpath (1.6%), compared with 39 of 129 (30.2%) in Cinnamon Gardens. This sample-based contrast motivates equity-focused auditing alongside the context-dependent walkability associations reported in [1]; direct footpath-status proportions are not directly comparable with composite walkability indices.

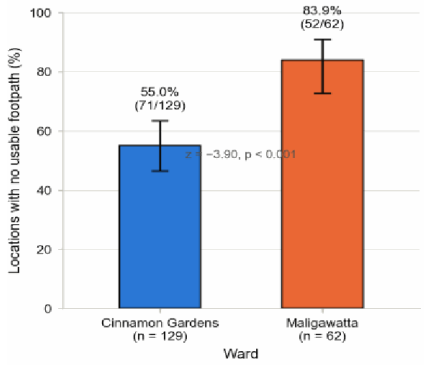



Fig. 2. Location-level % of validated locations with no usable footpath, by ward (Cinnamon Gardens n = 129, Maligawatta n = 62; best-side aggregation), 95% Wilson CIs. The disparity is statistically significant (two-proportion $z = -3.90$, $p < 0.001$).

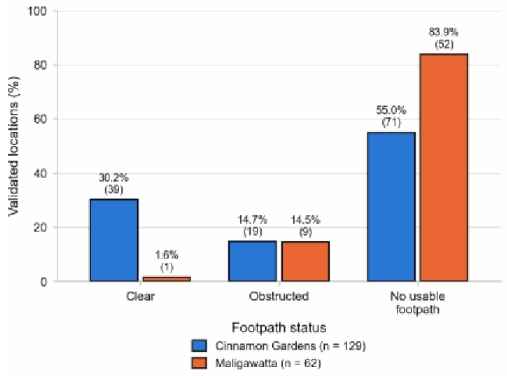



Fig. 3. Distribution of the three footpath-status classes across 191 validated locations, by ward (Cinnamon Gardens n = 129, Maligawatta n = 62; best-side aggregation). Clear footpaths are almost absent in the low-income ward (1/62, 1.6%) vs. the affluent ward (39/129, 30.2%).

## *C. Sensitivity to the Aggregation Rule*

The main equity result uses the best-side rule (Section IV-E), conservative with respect to the deprivation claim. Applying the worst-side rule instead -crediting each location with its least favourable observation -changes the classification of only 2 of the 191 locations (both in Cinnamon Gardens, both moving from footpath_clear to no_footpath), since only 7 locations retain both views and only 2 of those have discordant sides. Table IV reports both rules side by side. The Maligawatta no-footpath rate is unchanged (83.9%, 52/62); the Cinnamon Gardens rate rises marginally, from 55.0% (71/129) to 56.6% (73/129). Both the two-proportion z-test and $\chi^2$ test remain significant under the worst-side rule ($z = -3.71$, $p < 0.001$; $\chi^2 = 20.19$, $df = 2$, $p < 0.001$; Cramér's V = 0.325), with the wards' 95% CIs remaining non-overlapping. The main inequity conclusion is therefore robust to this aggregation choice; this robustness check is specific to the two rules bounding the ordinal scale used here and to the small number of paired-view locations, and we did not construct any additional aggregation rule, since none is motivated by the data or sampling design.

TABLE IV
LOCATION-LEVEL NO-FOOTPATH RATES UNDER BEST- AND WORST-SIDE AGGREGATION

| **Ward** | **Rule** | **n** | **No fp.** | **% (95% CI)** |
|---|---|---|---|---|
| Cinnamon Gardens | Best | 129 | 71 | 55.0 (46.4–63.4) |
| Cinnamon Gardens | Worst | 129 | 73 | 56.6 (48.0–64.8) |
| Maligawatta | Best | 62 | 52 | 83.9 (72.8–91.0) |
| Maligawatta | Worst | 62 | 52 | 83.9 (72.8–91.0) |
| Overall | Best | 191 | 123 | 64.4 (57.4–70.8) |
| Overall | Worst | 191 | 125 | 65.4 (58.5–71.8) |

Only seven locations retain both views; two change class. Confidence intervals are Wilson intervals and do not adjust for spatial dependence.

## *D. Inter-Annotator Reliability*

A second annotator independently labelled 77 of the sampled images. Agreement was 80.5% with $\kappa = 0.63$ for the three-class scheme, substantial agreement on the Landis-Koch scale [17] and rose to 85.7% ($\kappa = 0.69$) for the binary footpath present/absent distinction. Disagreement was concentrated in the intermediate obstructed class, whereas the two categories that underpin the main finding were highly reliable (clear

footpath, 15/16; no usable footpath, 43/50). The higher observed agreement for clear and no-footpath labels is encouraging, but it does not quantify the effect of residual annotation error on the ward disparity.

## VI. Discussion

Two complementary conclusions emerge. Substantively, the validated sample indicates deficient and unequal footpath provision in the two wards, and the disparity remains statistically significant under a conservative location-level analysis and is robust to the aggregation sensitivity check. This finding extends the walkability-equity literature [1] from composite indices in high-income cities to directly observed footpath conditions in a Global-South setting, providing sample-based observations that complement composite indices.

Methodologically, the benchmark cautions against deploying off-the-shelf street-view models unvalidated. The Mapillary-Vistas model's failure to clearly exceed baseline under evaluation controlling same-location leakage shows that this tested configuration did not exceed the reported majority baseline; this is consistent with documented geographic domain shift [6], [7] and with a "usable footpath" in Colombo, often narrow, encroached, or merged with the carriageway, differing visually from sidewalks encoded in Western training data. The one exception, a 7-point margin for ResNet-50 embeddings evaluated as a binary task, suggests generic visual features are not entirely uninformative here, though none of the four approaches reaches a level of agreement supporting unsupervised deployment (Section V-A). This contrast with North-American audit validation [5] motivates further study of transferability and construct differences; it does not isolate their causal contributions. Practically, image-based audits should be locally validated before use, while additional local labels and fine-tuning warrant evaluation. No learning-curve or cost study was performed, and transfer to other cities remains untested.

## VII. Limitations And Future Work

This is a validated pilot, not a city-wide census; two wards cannot represent Colombo, and results are framed as a comparative case study, not a generalisation to the city, country, or other Global-South cities. The sample (191 locations, 62 in Maligawatta) is modest; the 200-image draw was simple random sampling without ward stratification, and its ward composition differs from the full 4,240-image population more than expected under proportional sampling ($\chi^2 = 4.67$, $p = 0.03$), so it is not representative of that population, only the basis for the location-level analysis here. Construct ambiguity concentrates in the obstructed class ($\kappa = 0.63$ overall; 0.69 binary) versus reliable clear/no-footpath categories; this class is also hardest for every classifier (Section V-A), and with only 28 examples its P/R/F1 estimates are indicative, not precise. GSV imagery spans 2015-2025 (predominantly 2020- 2022, 97.1%), with a modest residual cross-ward date gap despite an identical modal month. Location-grouped CV (Section IV-C) removes same-point leakage but not spatial autocorrelation among nearby, distinct locations (median nearest-neighbour distance 63 m; 82/191 within 50 m); residual spatial autocorrelation may make evaluation estimates optimistic, and independence-based equity confidence intervals and tests may understate uncertainty. Spatial block cross-validation remains future work. A defensible block size must balance the scale of spatial dependence, ward geometry and adequate class counts per fold; it has not been selected or evaluated here. Finally, the benchmark is scoped to off-the-shelf/low-data approaches: we evaluated Cityscapes-and Mapillary-Vistas-trained models but not local fine-tuning, domain adaptation [6], [11], or the newer vision-language/geospatial foundation-model approaches now emerging [10], [12]. No learning-curve experiment is available to establish the benefit of additional local labels. Future work will test this hypothesis using multiple location-grouped training sizes and held-out evaluation, and scale the human-labelled set to fine-tune semantic-segmentation models on Colombo imagery, extend the analysis to more wards and cities to test generalisation, incorporate curb-ramp and surface-condition assessment for a richer accessibility construct, and integrate the footpath layer with census and GIS data to model socio-economic drivers of inequity; active learning could further reduce the labelling burden by prioritising the most informative images.

## VIII. Conclusion

Using GSV imagery and a validated human-labelling protocol, we obtained indicative, sample-based estimates of severe, unequal sidewalk deprivation in a bounded pilot of two contrasting Colombo wards: most validated locations lack a usable footpath, and the deficit is far worse in the low-income ward, a conclusion that survives both a conservative aggregation choice and a worst-side sensitivity check. We further showed that the specific off-the-shelf, low-data approaches evaluated here, including a model trained on globally-diverse street scenes, show weak, inconsistent agreement with human labels once evaluated under location-grouped cross-validation controlling same-location leakage, exposing a transferability gap for this task. Potential approaches to investigate include locally-labelled data and local fine-tuning, domain adaptation, active learning to prioritise informative images for labelling, or the newer vision-language and geospatial foundation-model approaches

## Acknowledgment

The authors thank the second annotator for contributing reliability labels and acknowledge the use of openly accessible data from OpenStreetMap (OSM), Google Street View (GSV), and the Department of Census and Statistics (DCS), Sri Lanka.

## References

[1] S. Conderino et al., "Social and economic differences in neighborhood walkability across 500 U.S. cities," Amer. J. Prev. Med., vol. 61, no. 3, pp. 394-401, 2021.

[2] M. Saha et al., "Project Sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale," in Proc. CHI Conf. Human Factors Comput. Syst., 2019, pp. 1-14.

[3] G. Weld et al., "Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery," in Proc. 21st Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS), 2019, pp. 196-209. doi: 10.1145/3308561.3353798.

[4] A. Liu et al., "Towards fine-grained sidewalk accessibility assessment with deep learning: Initial benchmarks and an open dataset," in Proc. 26th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS), 2024. doi: 10.1145/3663548.3688531.

[5] S. Askari et al., “Validating pedestrian infrastructure data: How well do street-view imagery audits compare to government field data?,” Urban Sci., vol. 9, no. 4, art. 130, 2025.

[6] D. Hong et al., “Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks,” Remote Sens. Environ., vol. 299, art. 113856, 2023.

[7] J. Yu et al., “Exploring geo-transferability of deep neural network by

[8] developing comprehensive metrics,” Geo-spatial Inf. Sci., vol. 29, no. 1, pp. 599-617, 2026. doi: 10.1080/10095020.2025.2505551.

[9] M. Cordts et al., “The Cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 3213-3223.

[10] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, “The Mapillary Vistas dataset for semantic understanding of street scenes,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 4990-4999.

[11] A. Torneiro et al., “Towards general urban monitoring with visionlanguage models: A review, evaluation, and a research agenda,” arXiv:2510.12400, 2025.

[12] M. Schwonberg, C. Werner, H. Gottschalk, and C. Meyer, “A study on unsupervised domain adaptation for semantic segmentation in the era of vision-language models,” arXiv:2411.16407, 2024.

[13] Z. Huang et al., “A survey on remote sensing foundation models: From vision to multimodality,” arXiv:2503.22081, 2025.

[14] Department of Census and Statistics, Census of Population and Housing 2024, Colombo, Sri Lanka, 2024.

[15] G. Boeing, “OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks,” Comput. Environ. Urban Syst., vol. 65, pp. 126-139, 2017.

[16] E. Xie et al., “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 12077-12090.

[17] J. Cohen, “A coefficient of agreement for nominal scales,” Educ. Psychol. Meas., vol. 20, no. 1, pp. 37-46, 1960.

[18] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol. 33, no. 1, pp. 159-174, 1977.

[19] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image

[21] recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778

# Store-and-Forward GPS/GSM Train-Location Tracking Under Intermittent Connectivity: A Sri Lankan Railway Prototype

Pawani Maheshika Bandara
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*
pawanimaheshika003@gmail.com

Janith Disanayake
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*

Niranga Dharmaratna
*Faculty of Computing, NSBM Green University*
*Mahenwatta, Pitipana, Homagama, Sri Lanka*

***Abstract*—Intermittent cellular connectivity can interrupt train-location reporting and cause older observations to arrive after newer ones. This paper presents a low-cost GPS/GSM railway tracking prototype that separates historical event persistence from publication of the latest train position. An ESP32-powered system collects and forwards location entries to an authenticated Node.js Express server hosted on AWS EC2. Historical entries are stored in MongoDB, whereas Socket.IO updates browsers with latest state document. In the mentioned tests, nine scenarios of authentication and input validation responded with correct HTTP status codes, all 50 records in one replay of simulated outage were successfully recovered, and all specified sequential order test cases did not update the latest state by an old entry. An artificial 90-days' workload includes 1,296,000 entries with the nominal aggregate rate of 10 entries per minute, showing how system deals with records for a long duration rather than performance. Another set with 59 railway records was gathered for the sake of completeness of operational scenario. Concurrent atomicity, power-loss durability, restart-safe ordering, and operational railway performance remain unverified.**

***Keywords—railway IoT, store-and-forward, intermittent connectivity, event ordering, latest-state consistency***

## I. Introduction

The use of connected positioning devices and cloud applications for passenger-information and railway-monitoring services is on the rise [1],[4]. Cellular connectivity could be intermittent in rural, mountainous and tunnel affected routes. Failing to upload can lead to missing information in the location history, and uploading the same old information after rejoining can cause a time jump from the current to the previous location. The transport-level delivery mechanisms do not specify how an application should keep its most up-to-date document [5], [8], [9].

This study investigates if a prototype GPS/GSM can recover queued events after reconnection and if delayed events will not replace a newer confirmed latest state. It tests a bounded implementation with an ESP32 device, authenticated REST API, a database (MongoDB), and browser clients. The contribution is not a new GNSS algorithm, transport standard or railway safety system, it's simply the separation of the idempotent history ingestion and the guarded latest-state publication under intermittent delivery.

This work makes three contributions. First an application-level contract for queuing, authenticated ingestion and separate history/latest-state storage is introduced. Second controlled observations of authentication duplicate suppression, sequence ordering and a single 50-record outage replay are presented. Third explicit separation of prototype, synthetic and contextual evidence is provided. Unique-index enforcement, cross-collection atomicity and durable restart ordering are listed as requirements, for validation not as implemented and tested safeguards.

## II. Related Work And Research Gap

Research on railway IoT looks at train parts and sensors[2],[4]. At the time studies on railway communications focus on movement and the challenges of train-to-ground communication [3],[19]. Delay-tolerant networking sets out store-and-forward methods [5] while MQTT gives options for how data moves at the transport level [8]. Neither of these approaches can decide if an older event should replace the most recent position record in an application. Research, on distributed systems separates delivery, duplicate removal and event ordering as semantics [9].

Earlier Sri Lankan work showed GPS and MQTT together for tracking trains [13] while research on railway-edge processing looked at a way of designing systems [14]. Research on railway benchmarking looked at bigger questions about reliability and security [6]. Research, on railway GNSS focused on positioning accuracy, uncertainty, spoofing and combining sensors [7],[15],[18] [21]. The smaller gap that is being addressed here is the showing of backlog recovery authenticated data entry removing duplicates and making sure the latest state is shared safely in a system that is not used for real operations. The main contribution is putting these parts together into a system not creating each of these parts on its own.

TABLE I
RELATED WORK AND CONTRIBUTION SCOPE

| Source | Primary focus | Relation to this work |
|---|---|---|
| [5], [8] | DTN / MQTT delivery | Transport; no latest-document semantics |
| [13] | Sri Lankan train GPS / MQTT | Local integration baseline |

| | | |
|---|---|---|
| [14] | Railway edge processing | Stream/edge system focus |
| [15]–[18] | Railway GNSS | Accuracy and uncertainty boundary |
| Present work | ESP32 + REST + MongoDB | History / latest-state separation |

## III. Research Methodology

The study uses a design-science method: find the problem build a prototype and test it using evaluation questions [10]. The data is divided into a development prototype, controlled software and API cases and a long-term synthetic record workload (Table II). A different railway dataset is used for the operational context. The evaluation that was shared did not include Sri Lanka Railways trains, a safety-control system or any personal data, from passengers.

TABLE II
EVIDENCE CLASSES AND VALIDITY LIMITS

| Evidence class | Used to assess | Does not establish |
|---|---|---|
| Physical prototype | Fix-to-browser development path | On-train or GNSS accuracy |
| Controlled software | Auth, schema, replay, ordering | Concurrency/restart correctness |
| Synthetic workload | 90-day software record path | Load/stress throughput |
| Contextual n=59 | Delay association | Impact caused by prototype |

The original evaluation questions are: EQ1, can outage-queued records be recovered; EQ2, can delayed records be prevented from replacing a newer latest state; EQ3, are exact duplicates suppressed; EQ4, what storage-volume assumption follows from one-minute reporting; and EQ5, are unauthorized or malformed requests rejected? Additional concurrency, restart, and performance measures proposed in Section VII are not represented as completed EQ tests.

## IV. Architecture and Event-Ordering Contract

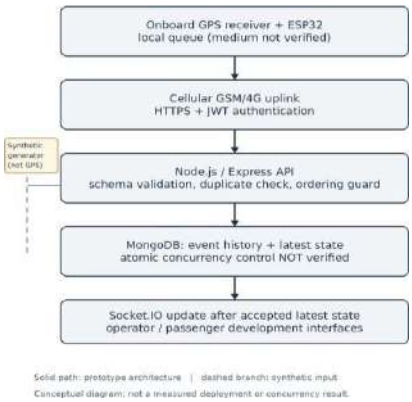



Fig. 1. Conceptual prototype data path and separate synthetic test source.

Nonvolatile queue media and atomic concurrent updates have not been verified.

Figure 1 shows the reported GPS receiver, ESP32, GSM/4G link, API, MongoDB storage and browser interface. The API was described as using HTTPS/TLS 1.3 and JWT-based access control [11] [12]. Browser clients receive the latest-state record instead of treating historical arrival order as the current position. The synthetic generator tests the application path. Does not replicate measured radio conditions or actual train movement.

TABLE III
REPORTED HARDWARE / SOFTWARE ROLES AND GAPS

| Component | Reported role / status |
|---|---|
| Controller | ESP32; exact board / firmware unavailable |
| Positioning | GPS receiver; exact model undocumented |
| Cellular | GSM/4G interface; exact model undocumented |
| Queue storage | Application queue reported; nonvolatile medium not verified |
| Backend | Node.js / Express on AWS EC2 |
| Database | MongoDB history and latest-state documents |
| Client | Socket.IO browser development interface |
| Security | HTTPS/TLS 1.3; JWT (reported) |

### *A. Event schema and trust assumptions*

When I look at a location event, I see train_id, a sequence number (seq) the device-GPS capture time (captured_at) latitude and longitude. Sometimes I also see engine_id and speed. The server-added field transmitted_at looks like the time the server received or processed the event, not a device transmission time. To know what transmitted_at means I would need to check the source code, and I should not use it to guess uplink latency. Table IV lists each reported field. Tells where each field comes from.

TABLE IV
EVENT FIELDS AND DATA PROVENANCE

| Field | Role / constraint |
|---|---|
| train_id; engine_id | Reporting vehicle identifiers |
| seq | Monotonic only within a device session |
| captured_at | Device/GPS capture time; trust assumption |
| latitude, longitude | Coordinates subject to range validation |
| speed (optional) | Device-derived speed if available |
| transmitted_at | Reported server-added time; name ambiguous |

The duplicate key that was reported is (train_id, seq) which's clear only during a single device sequence session. The sequence guard allows a latest-state update only if the incoming.seq is greater than the latest.seq and the incoming. captured_at is greater, than the latest.captured_at. This strict rule stops regression based on its given assumptions. Could stop a valid event if the capture clock goes backward. That kind of event could still be kept in history. JWT authorization does not prove that the GPS coordinates are physically accurate [7],[16]–[18].

*B. Device queue and acknowledgment behavior*

The procedure on the device records every confirmed GPS fix in an application queue before trying to send it. Events that are waiting are sent starting with the one. An event is deleted after an HTTP 200 response is received. If there is a time out or a server error the event is tried again. If a payload is not valid in a way it is put into a quarantine so that it does not stop other events from being sent. The information that is kept does not say whether the queue is stored in flash, an SD card or, in memory that's notpermanent. It also does not explain what happens if the power is cut off suddenly. Because of this this paper explains how the device stores. Sends data without saying that it can recover after a sudden loss of power.

*C. Server acceptance, history and latest-state semantics*

The API first checks JWT authorization. Then it checks that the payload is valid. The API looks for a history record that uses the key (train_id, seq). If the history record does not exist, the API inserts an event. The API updates the state when both the sequence number and the capture time are greater than the current values. In the documented duplicate case, the system returned an acknowledgment. The system did not create a history entry. Invalid or unauthorized requests should not modify either the history collection or the latest-state collection. However, a read-before-write check, at the application level does not guarantee uniqueness or atomic behavior when multiple requests run concurrently. Execution and crash recovery were not validated.

A concurrency-safe implementation requires an enforced compound index on the history key a latest-state key for each train and an atomic conditional update of the latest state. In this model the condition must require that incoming.seq is greater than committed.seq and that incoming.captured_at is greater, than committed.captured_at, not smaller. If history insertion and latest-state publication must succeed together the system should use a configured MongoDB transaction. The system should publish the Socket.IO update only after the transaction has committed. Retries that use the same idempotency key must include a payload. Conflicting payloads should be recorded as integrity errors. These are implementation requirements; they are not features that have been experimentally verified in this prototype.

*D. Restart boundary and formal invariants*

The intended invariants, for one device session are: H1, a history key identifies at most one event if a unique index is enforced; H2, committed latest-state sequence and capture time do not decrease; H3, rejected authorization or schema requests leave database state unchanged; and H4 clients publish only committed latest-state data. If every update is atomic and must satisfy incoming.seq > committed.seq and incoming.captured_at > committed.captured_at then monotonicity can be shown by induction. However, this conditional reasoning does not prove that the necessary atomic operation was carried out. Moreover, the strict dual-field rule cannot assure progress when capture time regresses.

A random boot_id is used to identify sessions. Alone it does not show the order of those sessions. After a restart to have an order there must be a monotonic boot_epoch or an authenticated session order set by the server. Also, rules are needed to handle backlogs caused by sessions clock drift and overlapping records. The original documentation did not check a durable boot counter, a queue medium or a restart-reconciliation protocol. Therefore, it is not claimed that restart safety is achieved.

## V. Experiment Setup And Reported Evidence

*A. Synthetic workload and estimated storage*

Alone it does not show the order of those sessions. After a restart to have an order there must be a monotonic boot_epoch or an authenticated session order set by the server. Also, rules are needed to handle backlogs caused by sessions clock drift and overlapping records. The original documentation did not check a durable boot counter, a queue medium or a restart-reconciliation protocol. Therefore, it is not claimed that restart safety is achieved.

TABLE V
SYNTHETIC WORKLOAD (NOT A STRESS TEST)

| Parameter | Value |
|---|---|
| Simulated trains | 10 |
| Route contexts | 8 |
| Sampling interval | 1 minute/train |
| Duration | 90 days |
| Record count | 1,296,000 |
| Aggregate source rate | 10/minute (~0.167/s) |
| Outages | Simulated per-train toggles |

Assuming each history document is 350 bytes 1,296,000 records produce 453,600,000 bytes. That is 453.6 megabytes or 432.6 mebibytes before we add indexes, journal entries, replication, compression or application data. These numbers are a storage-model estimate, not the real usage of a production database.

*B. Authentication and validation cases*

We looked at nine controlled request scenarios. Compared the expected HTTP response codes, with the ones we saw. All nine of the reported HTTP response codes matched the values that were expected as shown in Table VI. The evidence we have does not prove that we can rely on repeated trials that the application is fast or that the database stays the same before and after each rejected request. To check those properties, we need an automated test suite.

TABLE VI
AUTHENTICATION AND INPUT VALIDATION

| Case | Expected | Observed |
|---|---|---|
| Missing JWT | 401 | 401 |
| Expired JWT | 401 | 401 |
| Disallowed JWT alg. | 401 | 401 |
| Insufficient role | 403 | 403 |
| Tampered JWT | 401 | 401 |
| Malformed JSON | 422 | 422 |

| | | |
|---|---|---|
| Latitude = 95 | 422 | 422 |
| Longitude = 185 | 422 | 422 |
| Missing train_id | 422 | 422 |

### C. Sequencing, duplicates, and outage replay

Controlled ordering inputs included an event with a sequence number than the current state, an older capture timestamp, an exact duplicate, a sequence-reset case and a 50-record queue replay, after simulated reconnection (Table VII). The reported replay recovered 50 out of 50 records. No stale latest state overwrites occurred in these defined cases. The sequence-reset case does not establish crash- ordering. It also does not ensure restart- ordering. No repeated runs were documented. No queue-drain times were recorded.

TABLE VII
SEQUENTIAL ORDERING AND REPLAY CASES

| Scenario | Expected | Reported observation |
|---|---|---|
| Lower sequence | Latest unchanged | No overwrite |
| Older captured_at | Latest unchanged | No overwrite |
| Exact duplicate | No second history | Suppressed |
| Sequence reset input | No stale overwrite | No overwrite (case only) |
| 50-record backlog | Recover 50 | 50/50 recovered |

### D. Contextual railway statistics

A separate dataset contained 59 railway observations of departure and arrival delays and related operational variables. Mean departure delay was 0.949 min (SD 2.270), and mean arrival delay was 0.797 min (SD 2.227). Departure and arrival delays were strongly associated (Pearson $r = 0.987$, $p < 0.001$). These statistics are contextual and observational; they neither validate the store-and-forward mechanism nor establish improved railway punctuality or operational benefit caused by the prototype.

TABLE VIII
CONTEXTUAL RAILWAY STATISTICS (N = 59)

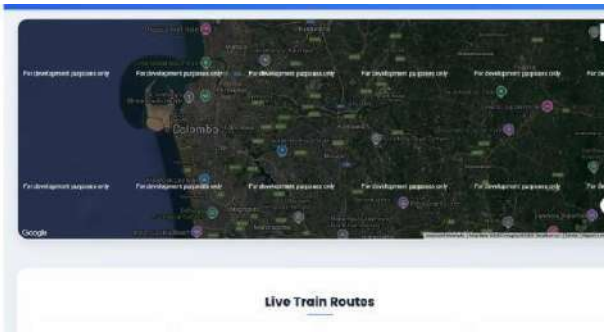
| Measure | Reported value | Meaning |
|---|---|---|
| Number of observations | 59 | Context only |
| Departure delay | 0.949 min; SD 2.270 | Descriptive |
| Arrival delay | 0.797 min; SD 2.227 | Descriptive |
| Delay correlation | $r = 0.987$; $p < 0.001$ | Association only |

## VI. Results Comparison And Limitations

### A. What the observations establish

All nine authentication and schema-validation scenarios had the expected HTTP statuses. The documented outage-replay case recovered 50 out of 50 queued events. The specified sequential ordering tests showed no stale latest state overwrite. These are controlled observations, for defined inputs, not failure-rate estimates. Concurrent correctness, crash durability, queue-drain latency, API latency distributions and large-scale capacity have not been measured.

### B. Explicit unguarded-replay baseline




Let's look at two events from a single device session: E11, which has sequence number 11 and was captured at 10:11 and E12, which has sequence number 12 and was captured at 10:12. If E12 arrives before E11, an unguarded last-arrival-write rule will show E11 as the recent state. With the proposed dual-field guard E12 stays the recent state while E11 can still be kept in history. If E11 arrives before E12 then both rules will show E12 as the recent state. Table IX compares these results in a way; those results are not from a real deployed system. Neither rule by itself solves the problem of writes or system restarts. To compare them properly you would need to run the input sequences repeatedly under controlled conditions, for each implementation.

TABLE IX
ANALYTICAL REPLAY BASELINE (NOT MEASURED)

| Condition (analytical) | Unguarded last-arrival-write | Sequential dual-field guard |
|---|---|---|
| E12 received, then older E11 | Displays E11 | Retains E12 as latest |
| Older E11 received, then E12 | Displays E12 | Displays E12 |
| Capture clock regresses | Last received fix may display | Does not advance latest |
| Concurrent requests | No atomicity guaranteed | No guarantee without atomic DB operation |
| Sequence reset after restart | No restart ordering policy | No verified restart ordering policy |

### C. Consistency, queueing, and security trade-offs

The dual-field predicate is carefully designed to be cautious: if a correct fix has a captured time that's later than expected it can stay in the history but cannot be moved to the latest state. Using the first replay method helps rebuild the past but can slow down showing the newest position until the backlog is cleared. Keeping data safe after a power loss has not been confirmed. An HTTP acknowledgment alone does not guarantee that recovery works in all crash situations and checking in order does not stop race conditions from happening.

TLS and JWT provide security for data in transit and access to requests. Only under their own assumptions. They do not ensure the integrity of GNSS signals. Do they guarantee radio coverage, secure storage on devices, resistance to spoofing or compliance with railway safety standards. Issues like device compromise, falsified capture times, multipath interference and

request flooding using stolen credentials were not, within the scope of their tested protections. There was no penetration testing, no evaluation of rate-limiting performance and no testing of credential rotation procedures reported.

### *D. Threats to validity and reproducibility*

Internal validity is limited by cases that were chosen by hand by one 50-record replay that was reported and by the lack of experiments that repeat automatically that run while test transaction failures, that test power loss and that test restarts. Construct validity is limited by queue persistence that has not been verified and, by a generator that does not measure radio outages or GNSS error. External validity is constrained by the lack of a field pilot that is authorized and by hardware benchmarks that are missing. The 59-observation contextual dataset does not show that the prototype reduces delays. No API percentile latency, no throughput benchmark, no queue-drain distribution and no cloud resource trace were reported.

Fig. 2. Development interface from the original project documentation; the screenshot does not establish live railway deployment.

## VII. Reproducibility and Follow-on Experiment Protocol

### *A. Required implementation disclosure*

Independent replication requires disclosure of the ESP32 board revision the GPS module model, the cellular module model, the antenna types the power supply details, the SIM or carrier used or the radio simulation setup. It also requires information on the queue storage medium the encoding method, the storage capacity and the overflow policy. The firmware version the libraries used the Node.js and Express versions, the MongoDB version, the database indexes, the request configuration, the test seeds, the scripts and the mapping between devices and training identities must all be documented. These details are currently. Incomplete, in the project evidence that has been kept.

### *B. Automated test matrix and measurable outcomes*

A follow-on evaluation should use input seeds and log the pre-test and post-test database state. Test variations in arrival order duplicate transmissions, sequence gaps, clock offsets, invalid payloads and simultaneous requests that share an idempotency key. Compare. Unguarded replay under the exact same conditions. Suggested queue sizes are 50, 100, 500 and 1,000 events. Suggested concurrency levels are 1, 5 and 20 clients. Aim for 30 repeated trials, per condition where possible. These are proposed experiments, not measurements reported in this paper. For each run document the hardware and software configurations.

For each trial I want to record acknowledged requests I want to record history entries I want to record recovery ratio (uniquely recovered events divided by distinct durably queued events) I want to record stale-overwrite count I want to record duplicate insertions I want to record p50/p95 API response time. I want to measure time from reconnection to queue emptying I want to measure time from reconnection to publication of the state; I want to record MongoDB storage/index growth I want to record host CPU/memory use. Queue-drain durations should use a clock with a specified stopping condition. I want to report medians I want to report variability across runs, then isolated best-case values. None of these quantities can be inferred from the 1,296,000-record total.

TABLE X
UNRESOLVED VALIDATION REQUIREMENTS

| Reviewer issue | Existing evidence | Required new evidence |
|---|---|---|
| Concurrent atomicity | Sequential tests only | Unique indexes, transactions, parallel duplicate races |
| Replay baseline | Qualitative reasoning | Unguarded vs guarded measured trials |
| Throughput / latency | 10 records/min source | Load sweep; p50/p95; resource traces |
| Queue drain | One 50-record recovery | Repeated drain times by backlog |
| Restart / power loss | Single reset input case | Boot epoch, hard resets, durable queue checks |
| Hardware | Roles only | Exact models, versions, wiring, storage |

### *C. Crash recovery and restart-safe ordering protocol*

Future crash-recovery tests should interrupt power at defined points: after a durable queue append but before upload during HTTP transmission after server commit but before acknowledgment and after acknowledgment but before durable dequeue. After each reboot reconcile the device queue, with server history and latest state; inspect suppression, missing events and monotonic state ordering. A restart-aware protocol must also specify how boot_epoch advances durably. These scenarios have not yet been tested.

### *D. Authorized field validation*

Authorized field validation is a process where a field pilot checks GPS or GNSS positions against a reference. The pilot also records when the cellular signal goes down and notes any warnings. They then separate the moments when the device records data, when the data goes across the network, when the server receives it and when the client shows it. Doing this study needs permission, from the railway authority that owns the tracks plus a safety and cybersecurity review. This task is not part of the prototype test.

## VIII. Conclusion

This study shows an approach at the application level for keeping past train-location events while stopping delayed records from overwriting the most recent state. In tests under controlled conditions nine authentication and schema setups gave the correct HTTP responses. A backlog of 50 records caused by an outage was successfully recovered. The tests on ordering showed no cases of outdated latest-state records being written over the current ones.

The synthetic data set had 1,296,000 records with 10 records per minute on average. A separate railway data set with 59

observations offered real-world context even though it did not show cause-and-effect relationships. The results suggest that the prototype works in a way, but they do not prove that the system handles atomic operations at the same time survives restarts scales well has accurate GNSS data or is ready for actual railway use.

To make claims the system needs enforced database uniqueness, atomic updates, queues that are safe, during restarts and repeated measurements of performance and recovery. These improvements are necessary before any broader conclusions can be drawn.

## REFERENCES


[1] Asian Development Bank, "Sri Lanka: Strengthening of Management Information Systems and Asset Management Capacity (49111-005)," project data sheet, updated Oct. 21, 2026. [Online]. Available: https://www.adb.org/projects/49111-005/main

[2] P. Fraga-Lamas, T. M. Fernández-Caramés, and L. Castedo, "Towards the Internet of Smart Trains: A Review on Industrial IoT-Connected Railways," Sensors, vol. 17, no. 6, art. 1457, 2017, doi: 10.3390/s17061457.

[3] É. Masson and M. Berbineau, Broadband Wireless Communications for Railway Applications: For Onboard Internet Access and Other Applications. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-47202-7.

[4] S. Sarp, M. Kuzlu, V. Jovanovic, Z. Polat, and O. Guler, "Digitalization of railway transportation through AI-powered services: digital twin trains," Eur. Transp. Res. Rev., vol. 16, art. 58, 2024, doi: 10.1186/s12544-024-00679-5.

[5] V. Cerf et al., "Delay-Tolerant Networking Architecture," IETF RFC 4838, Apr. 2007, doi: 10.17487/RFC4838.

[6] Rizzardi, R. Della Corte, J. F. Cevallos M., S. De Vivo, S. Sicari, D. Cotroneo, and A. Coen-Porisini, "SereWay: Toward SEcurity and REliability Benchmarking for the RailWAY IIoT," J. Netw. Syst. Manage., vol. 34, art. 52, 2026, doi: 10.1007/s10922-025-10026-1.

[7] M. L. Psiaki and T. E. Humphreys, "GNSS spoofing and detection," Proc. IEEE, vol. 104, no. 6, pp. 1258–1270, 2016, doi: 10.1109/JPROC.2016.2526658. OASIS, "MQTT Version 5.0," OASIS Standard, Mar. 2019.

[8] M. van Steen and A. S. Tanenbaum, Distributed Systems, 3rd ed., 2017.

[9] R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[10] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," IETF RFC 8446, Aug. 2018, doi: 10.17487/RFC8446.

[11] M. Jones, J. Bradley, and N. Sakimura, "JSON Web Token (JWT)," IETF RFC 7519, May 2015, doi: 10.17487/RFC7519.

[12] D. Balasingham, S. Samarathunga, A. Bandara, G. G. Arachchige, N. Gamage, and J. L. Wijayaraja, "Enhancing Train Transportation in Sri Lanka: A Smart IOT Based Multi-Subsystem Approach Using MQTT," in Proc. IEEE TENCON 2024, 2024, pp. 895–898, doi: 10.1109/TENCON61640.2024.10902958.

[13] J. P. Vitorino, J. Simão, N. Datia, and M. Pato, "IRONEDGE: Stream Processing Architecture for Edge Applications," Algorithms, vol. 16, no. 2, art. 123, 2023, doi: 10.3390/a16020123.

[14] D. Mikhaylov et al., "Toward the Future Generation of Railway Localization Exploiting RTK and GNSS," IEEE Trans. Instrum. Meas., vol. 72, pp. 1–10, 2023, doi: 10.1109/TIM.2023.3272048.

[15] S. Spinsante and C. Stallo, "Hybridized-GNSS Approaches to Train Positioning: Challenges and Open Issues on Uncertainty," Sensors, vol. 20, no. 7, art. 1885, 2020, doi: 10.3390/s20071885.

[16] Cai, J. Liu, X. Dong, and J. Liu, "Study on key technologies of GNSS-based train state perception for train-centric railway signaling," High-Speed Railway, vol. 1, no. 1, pp. 47–55, 2023, doi: 10.1016/j.hspr.2022.12.003.

[17] O. Garcia Crespillo et al., "GNSS, IMU, camera and LIDAR technology characterization for railway ground truth and digital map generation," Transp. Res. Procedia, vol. 72, pp. 1029–1036, 2023, doi: 10.1016/j.trpro.2023.11.532.

[18] J. Zhao, X. Gao, Z. Wu, Q. Zhang, and H. Han, "Artificial intelligence in rail transit wireless communication systems: Status, challenges and solutions," Physical Communication, vol. 67, art. 102484, Dec. 2024, doi: 10.1016/j.phycom.2024.102484.

[19] N. Ibadah, C. Benavente-Peces, and M.-O. Pahl, "Securing the Future of Railway Systems: A Comprehensive Cybersecurity Strategy for Critical On-Board and Track-Side Infrastructure," Sensors, vol. 24, no. 24, art. 8218, 2024, doi: 10.3390/s24248218.

[20] S. Wang, J. Liu, B. Cai, J. Wang, D. Lu, and W. Jiang, "Anti-spoofing performance analysis of typical GNSS-based railway train positioning schemes," High-Speed Railway, vol. 3, no. 1, pp. 37–43, 2025, doi: 10.1016/j.hspr.2025.01.005.

# Machine Learning and Explainable Artificial Intelligence for Thalassemia Screening: A Review of CBC-Based Approaches, Clinical Interpretability, and Research Gaps

Izna Idris
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
ifizna@students.nsbm.ac.lk

G. Perera
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
gayanp@nsbm.ac.lk

***Abstract*—Thalassemia is a hereditary hemoglobin disorder for which early identification of carriers is essential to support genetic counselling, preventive healthcare, and appropriate clinical management. Although molecular genetic testing and hemoglobin analysis provide reliable diagnostic confirmation, their cost, infrastructure requirements, and limited availability can restrict their use for large-scale screening. Complete Blood Count (CBC) testing therefore remains an important first-line investigation because it is inexpensive, routinely available, and contains hematological indicators associated with thalassemia. In recent years, machine learning has increasingly been applied to CBC and related clinical data to improve carrier detection and thalassemia classification. However, substantial differences remain among existing approaches in terms of input data, feature engineering, classification objectives, model interpretability, and clinical applicability. This paper reviews recent machine learning and Explainable Artificial Intelligence (XAI) approaches for thalassemia screening, with particular emphasis on CBC-based prediction. Existing studies are examined across conventional machine learning, ensemble and deep learning methods, multiclass classification, and interpretable prediction frameworks. The review identifies several continuing challenges, including the dominance of binary classification, difficulty in distinguishing silent carriers from normal individuals, dependence on additional laboratory investigations in some high-performing models, limited integration of clinically established hematological indices, and inconsistent use of explainability techniques. Future research should therefore prioritize multiclass, transparent, clinically validated, and resource-efficient screening approaches capable of supporting routine healthcare environments.***

***Keywords—Carrier screening, Clinical decision support, Complete Blood Count (CBC), Explainable Artificial Intelligence (XAI), Thalassemia***

## I. Introduction

Thalassemia comprises a group of inherited hemoglobin disorders caused by reduced or absent synthesis of the globin chains required for normal hemoglobin production. The disorder remains an important public-health concern, particularly in regions where carrier prevalence is high and affected genes may be transmitted across generations [1], [2]. Although individuals with severe forms of thalassemia can present significant clinical complications, carriers and individuals with mild thalassemia traits may remain asymptomatic or exhibit only subtle hematological abnormalities. Consequently, early carrier identification is particularly important for genetic counselling, reproductive risk assessment, and prevention strategies.

Complete Blood Count (CBC), hemoglobin electrophoresis, High-Performance Liquid Chromatography (HPLC), and molecular genetic testing are among the laboratory tests used in thalassemia screening and diagnosis. While HPLC and hemoglobin electrophoresis are useful for characterizing hemoglobin variants, molecular analysis offers conclusive identification of underlying globin gene disorders. However, these studies are typically more costly than standard hematological testing and necessitate specialist laboratory equipment [2], [14]. Because CBC analysis is quick, less intrusive, affordable, and frequently available in healthcare settings, it is still frequently employed as an initial screening method.

Important details regarding erythrocyte morphology and hemoglobinization are provided by routine CBC measurements such as Red Blood Cell Count (RBC), hemoglobin concentration, Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), and Red Cell Distribution Width (RDW). In order to help distinguish between thalassemia and other causes of microcytic anemia, traditional discriminant equations such as the Mentzer, England-Fraser, Shine-Lal, and other hematological indices have also been created by combining these measurements [2], [3]. However, there can be significant overlap in hematological traits between normal people, thalassemia carriers, and especially silent carriers, which lowers the accuracy of straightforward threshold-based screening techniques.

As a result, machine learning has become a viable method for figuring out intricate connections between various hematological characteristics. Previous research has shown that thalassemia carriers may be identified using routinely accessible CBC data utilizing supervised learning algorithms, such as Support Vector Machines, Random Forests, ensemble classifiers, and other predictive techniques [5] - [11]. While later

work has investigated α-thalassemia classification, subtype prediction, hybrid learning techniques, and broader hemoglobinopathy detection, Rustam et al. showed that it is feasible to predict β-thalassemia carriers using CBC characteristics alone [5]. For instance, Christensen et al. found that machine learning techniques performed better than conventional discriminant formulae for the classification of α-thalassemia, but their results also showed that it is still challenging to distinguish silent carriers from non-carriers using the available hematological features [6].

Additionally, the discipline has started to go beyond binary carrier detection. In order to differentiate between various types of thalassemia and associated hemoglobin problems, recent research has looked into multiclass classification. While MultiThal and other recent frameworks have similarly investigated broader diagnostic categorization [8], Nasir et al. used machine learning to incorporate CBC and HPLC data for multiclass classification of thalassemia types [9]. These advancements increase the practical relevance of automated screening; However, a number of multiclass techniques depend on specific laboratory data, and challenging categories, such as silent carriers, have received very little attention.

The interpretability of prediction models is another difficulty. High-performing deep learning and machine learning algorithms may operate as "black-box" systems, making predictions without adequately elucidating the clinical aspects that underlie them. In the healthcare industry, where model outputs must be comprehensible and clinically justified, this restriction is especially crucial. Thus, there has been a growing interest in Explainable Artificial Intelligence (XAI) techniques including SHapley Additive exPlanations (SHAP), interpretable decision rules, and feature-attribution methods. While recent interpretable frameworks have explored rule-based and explainable models for thalassemia carrier screening [10], [12], DeepThal included SHAP to enhance interpretation of α-thalassemia prediction [7]. However, explainability is still not regularly included in current thalassemia prediction algorithms, especially when it comes to multiclass screening.

Existing reviews have summarized the broader application of artificial intelligence and machine learning to thalassemia [4], [14]. However, continued developments in CBC-based prediction, multiclass classification, silent-carrier detection, and interpretable machine learning justify a focused synthesis of these areas. Therefore, this review examines the progression of machine learning-based thalassemia screening with particular emphasis on:

1) The diagnostic role of routinely available CBC data

2) Conventional, ensemble, and deep learning approaches

3) The transition from binary carrier detection toward multiclass prediction

4) The integration of XAI and clinical interpretability

5) Unresolved research gaps affecting practical clinical implementation.

Unlike broader reviews of AI in thalassemia, this review focuses on CBC-based screening, examining the shift from binary to multiclass prediction, silent-carrier detection, and explainability. It identifies methodological and clinical limitations and evaluates how existing approaches balance predictive performance, data accessibility, interpretability, and clinical applicability to highlight future research priorities.

## II. Review Methodology

This review adopts a focused narrative and comparative approach to examine studies applying machine learning and Explainable Artificial Intelligence to thalassemia screening and classification. Relevant studies were identified from the literature used within the broader investigation of machine learning-based thalassemia screening, with emphasis placed on peer-reviewed journal articles and conference papers addressing CBC-based prediction, hematological feature analysis, carrier detection, thalassemia subtype classification, deep learning, multiclass prediction, and explainable machine learning.

The literature search was conducted using Google Scholar, ScienceDirect and IEEE Xplore, covering publications from 2018 to 2026. Search terms included combinations of keywords related to thalassemia, machine learning, CBC-based screening, hematological parameters, carrier detection, multiclass classification, explainable artificial intelligence, and clinical interpretability. Approximately 40 relevant studies were identified and examined during the literature review. Considering the six-page conference paper limit, 14 studies were selected for the review based on their relevance to the review objectives, use of hematological or CBC-related information, classification objectives, machine learning approaches, and interpretability. Of these, 10 representative studies were included in Table I for comparative analysis, while the remaining studies provided supporting clinical and research context.

Studies were selected based on their application of AI or ML to thalassemia screening, use of CBC or hematological features, carrier or multiclass classification, explainability, or clinical relevance. Studies focused solely on treatment or unrelated clinical outcomes were excluded unless providing essential context. The selected literature was evaluated across five dimensions: data and features, classification objective, computational approach, interpretability, and clinical applicability. The analysis considered binary or multiclass prediction, silent-carrier representation, explanations, and reliance on CBC alone or additional tests such as HPLC or molecular testing. This framework enabled comparison beyond accuracy, considering predictive performance, accessibility, and clinical transparency.

This review has several limitations. First, the literature search was conducted using selected academic databases, and relevant studies indexed elsewhere may not have been identified. Second, the final comparative analysis included 14 studies selected according to the review objectives and the six-page paper limit; therefore, the findings may not represent all available research on machine learning-based thalassemia screening. Third, differences in datasets, classification objectives, input features, and evaluation strategies across the selected studies limit direct comparison of their reported findings. These limitations should be considered when interpreting the research trends and gaps identified in this review.

## III. CBC - BASED THALASSEMIA SCREENING: CLINICAL FOUNDATION

CBC analysis represents one of the most practical sources of information for preliminary thalassemia screening because abnormalities in globin synthesis directly influence erythrocyte size, hemoglobin content, and red-cell characteristics. MCV and MCH are frequently reduced in thalassemia traits, while RBC counts may remain relatively preserved or increased compared with other microcytic anemias [2], [3]. RDW, MCHC, and hemoglobin concentration provide complementary information, although no single CBC parameter can reliably distinguish all thalassemia categories.

Conventional hematological indices combine several CBC measures numerically in an effort to increase discrimination. These formulas are appealing since they may be computed straight from regular blood count values and don't require any further laboratory testing. However, different thalassemia subtypes and populations may perform differently. By learning nonlinear correlations among hematological parameters, multivariable prediction models can outperform these set clinical equations, according to recent machine learning research [5], [6].

For silent carriers, the diagnostic challenge becomes especially significant. Their hematological measurements may show significant overlap with non-carriers if they stay within or near normal reference ranges. Although machine learning methods outperformed traditional discriminant equations in the α-thalassemia investigation by Christensen et al., it was still challenging to distinguish between silent and non-carriers using the available characteristics [6]. This finding shows that overall accuracy alone is insufficient to assess the clinical utility of machine learning; performance for subtle and clinically significant carrier groups must also be taken into account.

## IV. MACHINE LEARNING APPROACHES FOR THALASSEMIA SCREENING

### *A. Conventional and Ensemble Machine Learning*

Early and contemporary studies have evaluated a broad range of supervised machine learning algorithms for thalassemia screening, including Logistic Regression, Decision Trees, K-Nearest Neighbors, Support Vector Machines, Random Forests, Gradient Boosting, and ensemble classifiers. These methods are well suited to structured CBC datasets because they can model interactions among multiple hematological variables without requiring the very large training sets typically associated with deep neural networks.

The possibility of low-cost machine learning screening based on routinely acquired hematological data was supported by Rustam et al.'s demonstration that β-thalassemia carrier status can be predicted using CBC characteristics alone [5]. CBC-derived measures provide significant discriminatory information for thalassemia prediction, according to similar research utilizing Random Forest and other traditional classifiers [6], [11]. These methods have been expanded in more recent work using ensemble and hybrid procedures intended to improve predicted resilience across a variety of clinical factors.

Accessibility is the main benefit of CBC-only machine learning. Before patients are referred for specialized confirmation, a prediction approach that solely uses routine hematological data may be used as an initial screening method. However, equal performance across all carrier categories is not always implied by high reported overall accuracy. Results can be significantly impacted by dataset imbalance, population-specific hematological patterns, and overlapping traits between carriers and healthy individuals.

### *B. Deep Learning Approaches*

Deep learning methods have also been explored to capture complex nonlinear relationships among hematological measurements. DeepThal is a notable example, using deep neural learning to predict α-thalassemia trait from red blood cell parameters [7]. These approaches can identify relationships that may be difficult to capture using manually defined discriminant rules. However, deep learning models often require more data and computational resources and may be less interpretable than traditional models, particularly when applied to structured clinical features. Therefore, their complexity should be justified by meaningful improvements in predictive performance and clinical utility.

### *C. Multiclass and Genotype-Oriented Prediction*

A recurring limitation across thalassemia machine learning literature is the predominance of binary classification, such as distinguishing carriers from normal individuals or identifying a single thalassemia subtype. Although binary prediction is valuable for preliminary screening, clinical decision-making may require differentiation among several genetically and hematologically distinct categories.

There is progress toward this more general goal, according to recent multiclass research. Machine learning can distinguish between several thalassemia classifications, as demonstrated by Nasir et al.'s investigation of multiclass classification using CBC and HPLC characteristics [9]. Similar extensions of automated prediction beyond binary carrier identification have been made by other multiclass systems [8]. However, the degree to which these models represent inexpensive CBC-only screening is diminished by certain approaches' reliance on HPLC or other specialist data.

Additionally, there is also a lack of representation for silent carriers. Compared to traditional α- or β-thalassemia features, they are significantly more difficult to recognize due to their modest hematological abnormalities. Therefore, rather than depending just on aggregate accuracy, future multiclass systems should assess per-class performance and specifically look at whether clinically difficult carrier groups can be distinguished from healthy individuals.

## V. EXPLAINABLE ARTIFICIAL INTELLIGENCE AND CLINICAL INTERPRETABILITY

The increasing predictive capability of machine learning has created a corresponding need for transparent clinical interpretation. In medical screening, a model that assigns a high-risk classification without indicating the hematological factors contributing to that decision may have limited value for clinical validation. XAI therefore represents an important progression

from predictive modelling toward clinically interpretable decision support.

Because SHAP can quantify the contribution of individual features to both individual predictions and global model behavior, it has become one of the most widely used model-explanation tools. DeepThal used SHAP analysis to find hematological characteristics that affect the prediction of α-thalassemia [7]. Explainable strategies for carrier screening, such as interpretable ensemble methods and rule-based frameworks, have been investigated in other recent publications [10], [12]. For instance, Chishti et al. exhibited competitive performance without the need for traditional class-balancing techniques by proposing an interpretable Dominance-Based Rough Set Approach for β-thalassemia carrier detection [10].

Despite this progress, the application of XAI in thalassemia remains fragmented. Many studies report feature importance or predictive accuracy without providing clinically meaningful local explanations, while relatively few combine multiclass prediction with comprehensive interpretability. An important future direction is therefore not merely to attach an explanation method to a trained classifier but to evaluate whether model explanations correspond with established hematological knowledge and can be meaningfully interpreted by healthcare professionals.

## VI. COMPARATIVE ANALYSIS OF EXISTING STUDIES

The reviewed literature demonstrates a gradual evolution from conventional hematological screening to more advanced machine learning and explainable AI techniques. Early studies primarily focused on binary classification using standard CBC parameters to distinguish β-thalassemia carriers from healthy individuals. More recent research has expanded toward multiclass classification, deep learning, ensemble methods, explainable prediction models, and α-thalassemia screening. However, these developments have been uneven, with significant differences remaining in data requirements, clinical accessibility, classification scope, and model interpretability.

Table I summarizes representative studies selected from the reviewed literature. The comparison focuses on the characteristics most relevant to practical screening, including the primary input data, classification objective, predictive approach, use of explainability, major contribution, and principal limitation.

TABLE I
COMPARISON OF REPRESENTATIVE MACHINE LEARNING AND XAI STUDIES FOR THALASSEMIA SCREENING

| **Study** | **Data Used** | **Classification Objective** | **Main Approach** | **XAI / Interpretability** | **Main Contribution** |
|---|---|---|---|---|---|
| Rustam *et al.* [5] | CBC | β-thalassemia carrier detection | Multiple ML classifiers | Limited | Demonstrated effective carrier prediction using routine CBC features |
| Christensen *et al.* [6] | Routine hematological data | α-thalassemia classification | Comparative ML models | Feature-based interpretation | ML improved classification compared with conventional formulae |
| Phirom *et al.* [7] | RBC parameters | α-thalassemia trait prediction | Deep neural network | SHAP | Combined deep learning with feature-level explanations |
| Wang *et al.* [8] | Clinical / hematological data | Multiclass diagnosis | MultiThal classifier | Limited | Extended automated classification beyond binary screening |
| Nasir *et al.* [9] | CBC + HPLC | Multiclass thalassemia classification | KNN, SVM, XGBoost | Limited | Extended prediction to multiple thalassemia categories |
| Chishti *et al.* [10] | CBC | β-thalassemia carrier detection | Interpretable rule-based learning | Decision rules | Improved transparency of carrier prediction |
| Schipper *et al.* [11] | CBC | Hemoglobinopathy prediction | Machine learning | Feature-level interpretation | Demonstrated broader prediction using routine CBC data |
| Meti *et al.* [12] | Hematological features | α-thalassemia carrier screening | Explainable ML | XAI | Demonstrated interpretable carrier prediction |
| Lai *et al.* [13] | Clinical / hematological data | Early thalassemia classification | Interpretable prediction model | Model interpretation | Combined prediction with interpretable risk assessment |
| Ferih *et al.* [14] | Literature / previous studies | AI applications in thalassemia | Comprehensive review | XAI reviewed | Synthesized AI applications and interpretability challenges |

Despite their potential for preliminary thalassemia screening, CBC-based ML approaches face limitations due to overlapping hematological features among thalassemia subtypes, carriers, and conditions such as iron deficiency anemia. Therefore, predictions should support screening rather than replace diagnosis, with hemoglobin analysis and molecular testing used for confirmation when indicated. Model generalizability is also affected by differences in populations, class distributions, laboratory settings, and validation methods. External validation using independent and diverse datasets is essential to assess clinical reliability and broader applicability.

The comparison highlights several recurring trends in CBC-based thalassemia screening. First, CBC remains the dominant and most accessible data source, offering potential for large-scale preliminary screening without specialized investigations. Second, classification objectives are gradually shifting from binary carrier detection toward multiclass and subtype-oriented prediction, although binary approaches remain more established, particularly when using CBC alone. Third, improved predictive sophistication does not necessarily ensure clinical interpretability, as deep learning and ensemble models may be difficult to explain. Although recent studies incorporate XAI methods such as interpretable rules, SHAP, and feature attribution, explainability remains inconsistently integrated into multiclass screening. Finally, some advanced approaches rely on HPLC, molecular information, or additional laboratory variables, which may improve discrimination but reduce suitability for low-cost preliminary screening. Therefore, future research should balance predictive performance with data accessibility, interpretability, and clinical applicability.

## VII. RESEARCH GAPS AND FUTURE DIRECTIONS

The comparative analysis identifies several unresolved research gaps that should guide future work in machine learning-based thalassemia screening.

TABLE II
SUMMARY OF RESEARCH GAPS AND FUTURE DIRECTIONS

| **Research Area** | **Findings from Existing Studies** | **Identified Research Gap / Future Direction** |
|---|---|---|
| Classification Scope | Binary carrier detection remains common, although multiclass approaches are emerging [5], [8], [9] | More CBC-based multiclass genotype prediction is needed. |
| Silent Carrier Detection | Silent carriers remain difficult to distinguish using routine hematological features [6]. | Explicit silent-carrier modelling and class-specific evaluation are required. |
| Input Data | CBC supports accessible screening [5], [6], [11], while some multiclass approaches require HPLC [9]. | Reduce dependence on specialized investigations for preliminary screening |
| Feature Engineering | CBC parameters and conventional hematological indices provide useful screening information [3], [5], [6]. | Combine CBC, clinical indices, and engineered features more systematically. |
| Explainability | SHAP and interpretable approaches are increasingly investigated [7], [10], [12], [13]. | XAI remains insufficiently integrated with multiclass screening. |
| Clinical Validation | Strong results are reported across individual datasets. [5], [6], [11] | More external and multicenter validation is needed. |
| Practical Deployment | Research predominantly emphasizes prediction and evaluation [4], [14]. | Greater emphasis on clinically deployable decision-support systems is required. |

Despite substantial progress in machine learning-based thalassemia screening, several limitations continue to restrict clinical applicability, as summarized in Table II. Existing research remains dominated by binary classification, with clinically challenging categories such as silent carriers receiving limited attention. Furthermore, although CBC parameters demonstrate considerable predictive value, some advanced approaches rely on additional investigations such as HPLC or molecular testing, limiting their suitability for low-cost preliminary screening. Future research should therefore prioritize clinically meaningful multiclass and interpretable screening systems that integrate explainable AI and relevant hematological feature engineering with standard CBC parameters. Greater attention should also be given to clinical implementation, external validation across diverse populations, and class-specific evaluation. These developments may support screening systems that balance predictive performance, accessibility, transparency, and practical clinical utility.

## VIII. CONCLUSION

Machine learning has demonstrated considerable potential for improving thalassemia screening through the analysis of routinely available hematological information. Existing studies show that CBC parameters contain valuable predictive patterns for automated carrier detection and may reduce reliance on specialized investigations during initial screening. Conventional machine learning, ensemble models, and deep learning have contributed to predictive performance, while recent multiclass frameworks have extended screening beyond traditional binary classification. However, significant challenges remain, including the dominance of binary classification, difficulties in distinguishing silent carriers from healthy individuals, and the reliance of some multiclass models on additional investigations such as HPLC or molecular testing. Furthermore, explainability is not consistently integrated into high-performing systems, while clinical implementation and external validation remain limited. Future studies should therefore prioritize clinically meaningful multiclass classification, explicit consideration of challenging carrier groups, integration of relevant hematological features, transparent model explanations, and validation across diverse populations rather than focusing solely on predictive accuracy. Overall, combining routinely available CBC data with explainable machine learning offers a promising pathway

toward accessible and reliable clinical decision-support tools for early thalassemia screening.

## REFERENCES


[1] "Thalassaemia and other haemoglobinopathies - Report by the Secretariat," in *World Health organization(WHO)*, 2006.

[2] Vip Viprakasit and Supachai Ekwattanakit, "Clinical Classification, Screening and Diagnosis for Thalassemia," *Hematology/Oncology Clinics of North America,* 2018.

[3] J.M. England and PatriciaM. Fraser, "DIFFERENTIATION OF IRON DEFICIENCY FROM THALASSÆMIA TRAIT BY ROUTINE BLOOD-COUNT," *The Lancet,* vol. 301, no. 7801, pp. 449-452.

[4] Dawlat Abdulkarim Ali and Adnan Mohsin Abdulazeez, "Machine Learning-Based Prediction of Thalassemia: A Review," *Indonesian Journal of Computer Science,* vol. 13, pp. 4046-4071, 2024.

[5] F R Aszhari, Z Rustam, F Subroto and A S Semendawai, "Classification of thalassemia data using random forest algorithm," *Journal of Physics: Conference Series,* 2020.

[6] Frederik Christensen, Deniz kenan kilic, Izabela Ewa Nielsen, Tarec Christoffer El-Galaly, Andreas Glenthoj, Jens Helby, Henrik Frederiksen, Soren Moller and Alexander Djupnes fuglkjaer, "Classification of $\alpha$-thalassemia data using machine learning models," *Computer Methods and Programs in Biomedicine,* 2025.

[7] Krittaya Phirom, Phasit Charoenkwan, Watshara Shoombuatong, Pimlak C, Supatra Sirichotiyakul and Theera Tongsong , "DeepThal: A Deep Learning-Based Framework for the Large-Scale Prediction of the α+-Thalassemia Trait Using Red Blood Cell Parameters," *Clinical Medicine,* vol. 11, 2022.

[8] WenQiang Wang, RenQing Ye, BaoJia Tang and YuYing Qi, "MultiThal-classifier, a machine learning-based multi-class model for thalassemia diagnosis and classification," *Clinica Chimica Acta,* 2024.

[9] M U Nasir, M Zubair, M T Naseem, Tariq Shahzad, M A Khan and Amir H Gandomi, "Multiclass classification of thalassemia types using complete blood count and HPLC data with machine learning," scientific reports, 2025.

[10] S Chishti, Faryal Nosheen, Joddat Fatima, Nadia Sultan and Madiha Khalid, "From CBC to clarity: Interpretable detection of betathalassemia carriers in imbalanced datasets," *PLOS One,* 2025.

[11] Anoeska Schipper , Matthieu Rutten, Adriaan van Gammeren , Cornelis L Harteveld , Eloísa Urrechaga , Floor Weerkamp, Gijs den Besten, Johannes Krabbe, Jennichjen Slomp, Lise Schoonen , Maarten Broere, Merel van Wijnen , Mirelle J A J Huijskens, Tamara Koopmann, Bram van Ginneken , Ron Kusters and Steef Kurstjens, "Machine Learning-Based Prediction of Hemoglobinopathies Using Complete Blood Count Data," *Clinical Chemistry,* vol. 80, no. 8, pp. 1064-1075, 2024.

[12] Aditya H Meti, B. Uma Maheswari and Aditya Vijjapu, "Advancing Alpha-Thalassemia Carrier Screening for Better Predictions Using Explainable AI," in *2023 4th International Conference on Communication, Computing and Industry 6.0(C216)*, 2023.

[13] J. Lai, Tang, JW, Gong, SS, Ming-Xiong Qin, Yu-Lu Zhang, Quan-Fa Liang, Li-Yan Li, Zhen Cai and Liang Wang , "Development and validation of an interpretable risk prediction model for the early classification of thalassemia," *npj Digital Medicine,* 2025.

[14] Khaled Ferih, Basel Elsayed, Amgad M. Elshoeibi, Ahmed A. Elsabagh, Mohamed Elhadary , Ashraf Soliman, Mohammed Abdalgayoom and Mohamed Yassin , "Applications of Artificial Intelligence in Thalassemia: A Comprehensive Review," *Diagnostics,* vol. 13, no. 9, 2023.

# Melanoma Skin Cancer Detection and Prevention Mobile Application Using Machine Learning

K L D P P Kakulandala
*Department of Software Engineering*
*NSBM Green University*
*Homagama, Sri Lanka*
praveenkekulandala@gmail.com

***Abstract*—Skin-lesion self-monitoring requires image capture guidance, consistent records and understandable next-action advice. This study presents SkinSafe, a mobile screening-support prototype integrating an on-device MobileNetV2 classifier with a capture countdown, a previous-image outline, lesion history and prevention education. Flutter provides the interface, Firebase supports authentication and record synchronization, and TensorFlow Lite performs local inference. Questionnaire answers inform rule-based guidance without changing the image-model score. Retained model records document development using HAM10000 and a selected PAD-UFES-20 smartphone-image collection. Historical partitions share patient and lesion identities, preventing interpretation as independent validation. A manual workbook contains 312 prediction records, including 218 reused training images; its numerical consistency can be checked, but image-to-record and tested-build traceability remain incomplete. Software records include 136 passing Flutter tests, 67 authorization simulations and a successful Android debug build. Six-item post-use ratings from 15 university students averaged 4.53/5. Three additional UAT response rows report 29 unassisted and one assisted task completion. These small, descriptive evaluations support the implemented workflow and reported user experience. They do not establish diagnostic accuracy in unseen patients, anxiety reduction or improved clinical outcomes. Independent patient-separated evaluation and a controlled usability study remain necessary.**

***Keywords*—Melanoma, Mobile Health, MobileNetV2, Lesion Tracking, Usability***

## I. INTRODUCTION

SkinSafe was developed to support a practical sequence: record a concerning skin lesion, obtain a preliminary image assessment, understand the next action and revisit the same recorded spot. Convolutional neural networks have demonstrated promising skin-image classification under defined evaluation conditions [1]. Translating that capability into a consumer workflow also requires usable capture guidance, persistent records and clear communication of uncertainty.

Clinical and interaction outcomes require separate evidence. A recent prospective study of SkinVision examined diagnostic accuracy and capture conditions in 1,458 participants [2]. A randomized study of a skin self-monitoring application found no evidence of improvement in consultation or selfexamination outcomes in its study population [3]. These findings motivate careful evaluation of the complete workflow rather than an inference of clinical benefit from a software demonstration.

TABLE I
POSITIONING AGAINST RELEVANT PRIOR WORK

| Study / system | Evidence or focus | Relation to SkinSafe |
|---|---|---|
| Esteva et al. [1] | CNN skin-image classification with dermatologist comparison. | Supports the use of image learning; does not validate this model or its mobile workflow. |
| SkinVision, Kips et al. [2] | Prospective clinical evaluation; capture conditions and device variation examined. | Shows why a prototype requires independent clinical and capture evaluation. |
| Selfmonitoring app, Walter et al. [3] | Randomized evaluation of self-examination, consultation and psychological outcomes. | Shows that usability alone cannot demonstrate clinical or psychological benefit. |
| SkinSafe, this study | On-device inference, capture aids, records and cautious guidance; descriptive student feedback. | Workflow implementation is demonstrated; clinical superiority is untested. |

The contribution of this work is an implemented combination of local image inference, guided repeated capture, patient/spot-linked records, prevention education and rulebased next-action guidance. It is an engineering integration contribution. We do not claim a new CNN architecture, superiority over existing applications, or a clinically validated diagnostic service. This paper documents the retained development evidence and the limits of the available evaluation.

## II. RELATED WORK AND DESIGN RATIONALE

Image-classification research [1], clinical smartphone-app validation [2] and trials of self-monitoring interventions [3] address different evaluation questions. Table I positions SkinSafe against these approaches. It is a literature-based comparison, not a head-to-head benchmark; performance numbers from different populations and target diagnoses cannot establish relative superiority.

The SPIKES communication protocol was developed for clinician-patient conversations about unfavorable information [4]. SkinSafe draws on its emphasis on understandable information, support and an action plan when wording notifications. A software interface cannot reproduce the full interpersonal protocol. The design is therefore described as

communication informed, with no measured claim that it reduces anxiety.

TABLE II
RETAINED IMAGE PARTITIONS: MELANOMA / OTHER

| Source | Train | Validation | Nominal test |
|---|---|---|---|
| HAM10000 | 779 / 6,230 | 167 / 1,336 | 167 / 1,336 |
| Selected PAD-UFES-20 | 36 / 182 | 8 / 39 | 8 / 39 |

## III. METHODOLOGY AND IMPLEMENTATION

### *A. Development Approach and Scope*

The project followed an iterative design-science approach: literature and requirements review, implementation of the artifact, and evaluation of software behavior and user feedback. A retained requirements survey contains 32 responses; these informed feature preferences and are not treated as post-use acceptance results. The separate post-use and UAT records are described in Section IV.

The Flutter/Dart interface uses Firebase Authentication and Cloud Firestore for account and record services, with local storage for lesion images and cached records. The examined implementation stores images in application-controlled local directories. TensorFlow Lite runs the image classifier on the device. Educational content and UV information support prevention awareness; reduced disease incidence or measured personal UV exposure was not evaluated.

### *B. Dataset Provenance and Partitions*

The model uses MobileNetV2 [5], initialized through ImageNet transfer learning. Development used HAM10000 dermatoscopic images [6], followed by adaptation using selected PAD-UFES-20 clinical photographs collected with smartphones [7]. These sources represent different imageacquisition domains. Table II documents the retained dataset and split records.

The HAM10000 manifests contain 10,015 images. The selected PAD collection contains 312 images: 52 melanoma and 52 from each of five other categories (basal cell carcinoma, squamous cell carcinoma, nevus, seborrheic keratosis and actinic/solar keratosis). The five categories were collapsed into the non-melanoma label. This label includes other cancers and must not be interpreted as healthy skin. The selected class proportions do not estimate disease prevalence among app users.

The PAD subset is defined by a retained download plan. The historical clinical inclusion/exclusion rationale and rejectedimage log are incomplete, so it is not described as a consecutive clinical sample. The original image-level partitions are not independent at patient or lesion level. The retained audit records eight shared patient IDs and seven shared lesion IDs between PAD training and nominal test partitions, and twelve shared patient IDs and nine shared lesion IDs between training and validation. The 312-image manual collection also includes all 218 PAD training images. These exposures preclude an unseen-patient accuracy claim.

TABLE III
SAVED TRAINING HISTORY AND REPRODUCTION LIMITS

| Item | Retained setting or boundary |
|---|---|
| HAM continuation | 8 recorded epochs: learning rate 5e-6 for epochs 1-4; 2.5e-6 for 5-8. |
| PAD adaptation | 12 recorded epochs: learning rate 3e-6 for epochs 1-9; 1.5e-6 for 10-12. |
| Batch / seed | Archived defaults: HAM batch 32, PAD batch 16, seed 42. Exact historical CLI overrides are unavailable. |
| Callbacks | Archived code monitors validation AUC: early-stop patience 3; best checkpoint; LR halving after patience 2. |
| Missing history | Earlier head-training length and complete run-time invocation are not retained. |

### *C. Model Architecture and Training Records*

The retained checkpoint accepts 224 x 224 RGB float32 tensors and applies x/127.5 - 1 scaling within the model. The head contains global average pooling, dropout of 0.30 and a single sigmoid output. The saved state has the first trainable weighted backbone layer at index 120, with backbone batchnormalization layers frozen. Binary cross-entropy is optimized using Adam (beta1=0.9, beta2=0.999, epsilon=1e-7).

Saved augmentation layers implement horizontal flips, rotations with factor 0.05, zoom with factor 0.08 and contrast variation with factor 0.10. These transformations increase training variation but provide no evidence of equivalent performance across skin tones. The archived training pipeline decodes RGB JPEGs, resizes to 224 x 224 and casts to float32. Class weighting is computed from training counts as $w_c=N/(2n_c)$; validation/test labels are excluded from that calculation.

Table III distinguishes saved observations from source defaults. It does not reconstruct missing run details as known. The trained Keras model was exported using the TensorFlow Lite converter. The application subsequently applies orientation correction, central cropping, mild sharpening and resizing; that path differs from the archived resize-only training preparation. Numerical equivalence between the Python and Dart sharpening paths has not been established.

### *D. Capture, Guidance and False-Negative Precautions*

The capture workflow supplies framing instructions and a two-second countdown. A previous lesion outline can assist visual alignment during follow-up capture. These interface aids do not measure camera distance or prove identical focal length, viewpoint or physical lesion size. Image-quality checks may request a retake; their performance across skin tones and non-skin inputs remains unevaluated. The outline is a visual recording aid rather than a validated segmentation or growth measurement.

Retained application configuration uses 0.45 as the caution boundary and 0.75 as the specialist-review boundary. These operational values differ from the manual workbook threshold of 0.50. The model output is an uncalibrated image score, not

an established individual disease probability. A separate rulebased policy uses reported recent change, growth, border/color changes, symptoms and relevant history to inform advice without changing the model score.

TABLE IV
MANUAL WORKBOOK COUNTS ON REUSED IMAGES

| Entered label | Predicted melanoma | Predicted other | Total |
|---|---|---|---|
| Melanoma | 49 (TP) | 3 (FN) | 52 |
| Non-melanoma | 7 (FP) | 253 (TN) | 260 |
| Total | 56 | 256 | 312 |

A low image score does not rule out melanoma. For reported changing, growing or symptomatic lesions, the implemented policy directs users to professional review without waiting for another app scan. Higher-score messages recommend clinician review and state the diagnostic limitation. Non-melanoma outputs also do not exclude other skin cancers. This design addresses false-negative communication risk but has not been shown to prevent clinical harm. A functioning dermatologist directory or appointment-booking outcome is not claimed.

## IV. EVALUATION AND RESULTS

### *A. Evaluation Design*

Evaluation combines retained software checks, a manually recorded prediction workbook, six-item student feedback and a separate three-row UAT export. These measure different properties and are not pooled. There is no controlled nativecamera comparator, validated anxiety instrument or clinical attendance follow-up in the retained evaluation. Consequently, no alignment-improvement, anxiety-reduction or appointmentbooking effect is estimated.

The software retest dated 30 September 2026 reports 136 passing Flutter unit/widget tests, 67 passing Firestore authorization simulations and a successful Android debug build. These records cover specified software assertions rather than comprehensive physical-device acceptance or medical validation. They are historical evidence; this manuscript revision did not rerun the suites or retrain the model.

### *B. Manual Workbook: Descriptive Development Results*

The retained workbook contains 312 rows with record ID, entered label, score, threshold, prediction and outcome. All recorded thresholds are 0.50. Recomputing the outcomes confirms the confusion matrix in Table IV. This calculation checks the entered records; it neither independently reruns the application nor verifies the original diagnoses. The record IDs IMG_001-IMG_312 are not linked to the original image filenames in this workbook, and the exact tested app/model build is not retained.

Accuracy was computed as (TP + TN)/N, precision as TP/(TP + FP), sensitivity as TP/(TP + FN), specificity as TN/(TN + FP), and F1 as 2TP/(2TP + FP + FN). The recorded values are 96.79% accuracy (302/312), 87.50% precision, 94.23% sensitivity, 97.31% specificity and 90.74% F1. Three melanoma-labelled records were missed (3/52, 5.77%). Their clinical circumstances cannot be analyzed from the retained mapping. Lighting or image quality is therefore not asserted as their cause.

TABLE V
SIX CUSTOM USABILITY ITEMS (15 RESPONDENTS)

| Aspect | Mean | SD |
|---|---|---|
| Ease of use | 4.53 | 0.52 |
| Navigation | 4.53 | 0.52 |
| Capture instructions | 4.53 | 0.52 |
| Result and limitation understanding | 4.27 | 0.59 |
| Visual readability | 4.60 | 0.51 |
| Overall satisfaction | 4.73 | 0.46 |

These are descriptive workbook results on reused development images. An earlier retained automated evaluation produced different outcomes, and the discrepancy has not been reconciled at image/build level. The different recorded thresholds alone cannot establish an explanation or model improvement. Clinical ROC-AUC, calibration and confidence intervals are not claimed from these records: inference requires traceable predictions and an independently sampled patientseparated evaluation, and an interval would not correct training leakage.

### *C. Student Usability Feedback*

Fifteen university students reportedly used the application before rating six custom statements from 1 (strongly disagree) to 5 (strongly agree). The retained response matrix contains 90 ratings with no missing items. Table V gives each item's sample mean and sample standard deviation (SD); the overall mean is 4.53/5. There are 89 positive ratings (4 or 5), one neutral rating and no negative ratings. These are rating counts, not 90 independent participants.

The questionnaire is not the System Usability Scale and has no demonstrated anxiety-measurement validity. Detailed recruitment criteria, demographics, consent records, collection dates and device assignments for this matrix were not retained. No paired baseline or randomized comparator is available. The analysis is descriptive and does not support significance testing of improvement or generalization to patients.

### *D. Participant-Reported Acceptance Tasks*

A separate revised UAT export contains three affirmativeconsent rows. Ten tasks cover login, patient/spot selection, a scenario questionnaire, image analysis, guidance, saving, history retrieval, follow-up comparison, prevention information and logout. Across 30 task entries, participants reported 29 unassisted completions and one completion with assistance for follow-up comparison; no non-completion was recorded. The three rows cannot establish population acceptance or objective task success. Observer logs, task durations and exact build traceability are absent; independent result understanding and tooltip re-testing were not demonstrated.

## V. DISCUSSION AND LIMITATIONS

The evidence supports integration of the intended recording and guidance workflow. The result-clarity item has the lowest mean of the six usability items, and the reported assistance during follow-up comparison identifies a practical interaction to revisit. Neither finding proves a treatment or psychological effect. Communication-informed wording should be evaluated for both comprehension and possible false reassurance; making a message less alarming must not weaken advice to seek medical attention.

The retained data place clear limits on model conclusions. Training-image reuse, patient/lesion overlap, selective class composition, incomplete manual-image mapping and unresolved preprocessing/build differences prevent independent diagnostic interpretation. Historical training AUC or a high workbook accuracy cannot repair these design problems. Grouped split code developed later is not evidence that the deployed historical checkpoint was trained under that protocol.

Comparison with prior work establishes context, not superiority. Unlike the prospective SkinVision evaluation [2], the present study has no independent clinician-referenced cohort. Unlike the self-monitoring trial [3], it has no randomized comparison or measured psychological/consultation endpoints. The proposed combination is potentially useful as a software workflow, but its individual components need ablation or controlled comparisons before their incremental benefit can be established.

Future model evaluation should assign all images of each patient/lesion to one partition before any training or threshold tuning, freeze the preprocessing and deployed model build, retain image-source mappings and measure all exclusions and capture failures. A separate smartphone-image cohort with clinician-verified reference labels is needed for sensitivity, specificity, F1, ROC-AUC and calibration. Uncertainty estimation should respect patient clustering. External validation and skin-tone/device subgroup analysis must precede clinicalperformance claims.

Future interaction work should recruit a broader user group under an appropriate consent and ethics process, define tasks and acceptance criteria in advance, retain observer logs and timing, and compare capture guidance with a specified baseline. Anxiety or consultation outcomes require appropriate instruments and follow-up rather than satisfaction ratings. Any patient study requires institutional review and clinical supervision. No ethics approval, patient trial or completed retest is inferred from the existing student records.

## VI. CONCLUSION

SkinSafe demonstrates an academic prototype combining on-device image inference with guided capture, lesion records, prevention information and cautious next-action guidance. Retained software checks and small student feedback datasets describe implementation and reported interaction. The model and manual-evaluation records remain affected by training exposure and traceability gaps. Independent diagnostic performance, anxiety reduction, physical image standardization and clinical benefit have not been established. The next contribution must be a reproducible, patient-separated and prospectively documented evaluation.

## REFERENCES


[1] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017, doi: 10.1038/nature21056.

[2] J. Kips et al., "Artificial intelligence-based smartphone application for skin cancer detection: a prospective diagnostic accuracy study," British Journal of Dermatology, vol. 194, no. 6, pp. 1077-1086, 2026, doi: 10. 1093/bjd/ljag057.

[3] F. M. Walter et al., "Effect of a Skin Self-monitoring Smartphone Application on Time to Physician Consultation Among Patients With

[4] Possible Melanoma: A Phase 2 Randomized Clinical Trial," JAMA Network Open, vol. 3, no. 2, Art. no. e200001, 2020, doi: 10.1001/jamanetworkopen.2020.0001.

[5] W. F. Baile, R. Buckman, R. Lenzi, G. Glober, E. A. Beale, and A. P. Kudelka, "SPIKES-A six-step protocol for delivering bad news: application to the patient with cancer," The Oncologist, vol. 5, no. 4, pp. 302-311, 2000, doi: 10.1634/theoncologist.5-4-302.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[7] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, Art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[8] A. G. C. Pacheco et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, Art. no. 106221, 2020, doi: 10.1016/j.dib.2020.106221.

# An Instrumentation-Centered Multimodal AI Framework for Continuous Wearable Neurological Monitoring: A Conceptual Framework for Migraine Prediction

Lathushan RK
*Department of Life Sciences*
*NSBM Green University*
*Sri Lanka*
lathushanrk.res@gmail.com

Damayanthi Dahanayake
*Department of Life Sciences*
*NSBM Green University*
*Sri Lanka*
damayanthi.d@nsbm.ac.lk

Supun Gajendrasinghe
*Department of Software Engineering*
*NSBM Green University*
*Sri Lanka*
supun.g@nsbm.ac.lk

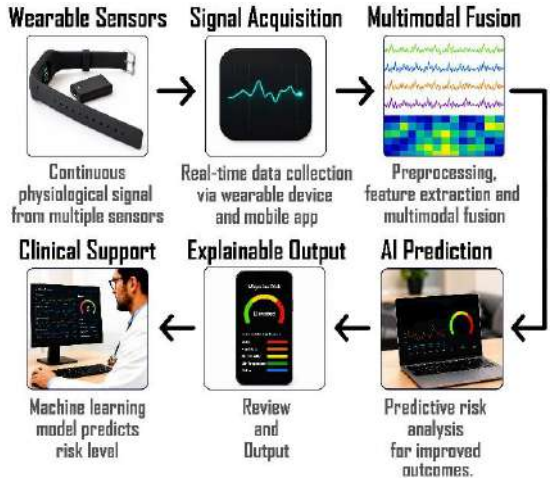



Fig. 1. Overview of the proposed instrumentation-centered multimodal AI framework for wearable neurological

**Abstract—Neurological disorders such as migraine remain difficult to monitor continuously outside clinical settings, and existing wearable approaches usually rely on one sensing modality or one disorder-specific prediction model. This paper proposes an instrumentation-centered, modular multimodal artificial intelligence (AI) framework that separates sensing, signal acquisition, conditioning, synchronization, edge processing, communication, AI prediction, explainable clinical decision support, and visualization into nine reusable layers. Migraine risk prediction, rather than migraine detection, is used as the initial case. Informed by 19 published studies, a conceptual wrist-worn Prototype V1 combines photoplethysmography, electrodermal activity, skin temperature, and inertial motion sensing around a low-power microcontroller with Bluetooth Low Energy, with explainable AI placed before clinical decision support. A literature-based design evaluation shows that this sensor set provides the input types of the most accurate wearable migraine-forecasting model reported in a recent review (84 percent accuracy in small cohorts), and that on-device feature extraction is calculated to reduce the transmitted data rate by about 99.9 percent. No hardware has been built and no prediction accuracy is claimed. The contribution is the architecture and its design rationale.**

***Keywords—Wearable Neurological Monitoring, Multimodal Sensing, Migraine Prediction, Biomedical Instrumentation, Explainable Artificial Intelligence***

## I. Introduction

### *A. Background*

Migraine has been reported as the most frequent and most costly of the primary headache disorders [1]. Continuous ambulatory monitoring has become increasingly feasible as low-power wearable sensor technology has advanced [2]. Wearable sensors covering physiological pressure and biochemical sensing modalities have been integrated into research-grade wearable prototypes [3], alongside dedicated photoplethysmography (PPG) [4] and electrodermal activity (EDA) [5] sensor designs. Artificial intelligence (AI) methods have been applied to extract patterns from continuous physiological data streams recorded by wearable devices [6]. Neurological wearable applications including gait monitoring in Huntington's disease and remote assessment in multiple sclerosis (MS) and myasthenia gravis (MG) illustrate an emerging role for wearables in neurology [7], [8]. Migraine is selected as the demonstration case because physiological and behavioral changes around attack days have been reported, and wearable data have been used to forecast attacks [1].

### *B. Research Problem*

Some existing wearable neurological monitoring approaches, such as [7], have been designed around a single disorder rather than a coordinated, reusable multimodal architecture. Migraine-specific wearable studies are generally reported as one predictive model's performance on one dataset rather than a reusable system design [1]. Interoperability between the instrumentation layer and the AI layer is not explicitly addressed in several of the reviewed studies. Explainability has not been integrated into every wearable prediction pipeline reviewed here, although feature-importance methods have been demonstrated for a related wearable-derived physiological classification task [6].

### *C. Research Gap*

The reviewed literature shows progress in three areas, namely wearable physiological sensing, migraine biomarker identification, and general multimodal healthcare architectures

[1]–[3]. Methods for acquiring EDA and skin-temperature signals have been validated against reference instruments [5], [9]. Motion-artifact correction methods for wearable PPG have also been demonstrated [4]. Several migraine-focused studies have identified autonomic changes, including heart rate (HR), EDA, and skin temperature variation, as inputs to wearable migraine-attack forecasting models [1]. General wearable healthcare architectures have proposed layered sensing-to-cloud designs for other conditions [10]. Among the reviewed studies, no study was identified that integrates instrumentation-level sensor justification, multimodal fusion, and explainable prediction into one modular architecture demonstrated through migraine.

### *D. Research Objective and Contribution*

The objective of this paper is to design a modular, instrumentation-centered engineering architecture for continuous wearable neurological monitoring that integrates physiological sensing, synchronized signal acquisition, multimodal AI, and explainable clinical decision support into reusable layers. Migraine risk prediction is used as the initial implementation case because wearable signals around attack days have been used for attack forecasting [1] and wearable sensing feasibility [5] has been reported. The contribution of this paper is the architecture and its design rationale, rather than a new sensor, algorithm, or clinical dataset. Specifically, it contributes (i) a nine-layer architecture with one function per layer (Fig. 1), (ii) a literature-based trace linking each sensing channel to reported validation evidence, (iii) an analytical signal and data-rate specification, and (iv) explainable AI as a mandatory layer before clinical decision support. The architecture therefore links biomedical instrumentation with computational prediction for digital health applications. The remainder of this paper reviews related work, presents the proposed framework, describes the migraine case, reports a literature-based design evaluation, and discusses engineering strengths and limitations.

## II. Related Work

### *A. Wearable Neurological Monitoring*

Wearable neurological monitoring has been described across several disorders using multimodal sensing. A shoe-based wearable integrating plantar pressure, accelerometry, and edge AI has been proposed for gait monitoring in Huntington's disease, transmitting data via a low-power wide-area network (LPWAN) [7]. Consumer-grade smartwatches have been used for longitudinal remote monitoring in MS, MG, and chronic inflammatory demyelinating polyneuropathy (CIDP), tracking step count and sleep metrics over several months [8]. Remote wearable electroencephalography (EEG) devices, including consumer headbands and ear-EEG systems, have been self-applied at home for sleep, mental-health, and neurological monitoring and for neurofeedback [11]. Reviews of wearable Internet-of-Things (IoT) devices for healthcare describe layered architectures connecting wearable sensors, mobile gateways, and cloud analytics for applications including blood pressure (BP) monitoring and tremor detection [10]. Bluetooth Low Energy (BLE) has been described as a low-power option suited to continuous, battery-constrained physiological data transfer [2].

### *B. Migraine Monitoring and Prediction*

Migraine has been specifically investigated as a wearable monitoring target in several recent studies. A narrative review of AI, data science, and wearable devices in headache research identifies six thematic application areas, including diagnosis, treatment-response prediction, disease-trajectory forecasting, and remote monitoring [1]. The same work reports that wearable-derived HR, skin temperature, EDA, and physical activity data have been used to forecast migraine attacks, with reported test-set accuracies from 62% to 84% in cohorts of 7 to 18 patients, although many underlying studies are small pilot investigations with limited external validation [1]. Wearable neuromodulation devices worn on the upper arm or head are used for acute headache relief [12], suggesting that body-worn hardware is practical in this population, although acceptance of a monitoring-only wearable has not been tested. A case-based remote monitoring study that tracked step count and sleep architecture via a consumer smartwatch around two migraine attacks showed reduced activity during the attacks and altered sleep patterns around them [8]. Current studies in this area mainly focus on identifying biomarkers or building single-model predictions, rather than developing a reusable framework that links sensing directly to decision-making [1].

### *C. Multimodal AI in Healthcare*

Multimodal sensing and AI have also been applied in wearable healthcare beyond migraine [3], [13]. A multimodal headband integrating thermal sensing via a flexible e-skin element and motion sensing through an embedded accelerometer has been developed for continuous cardiovascular and metabolic monitoring during physical activity, reporting its best estimate only when all inputs were combined (mean absolute percentage error, MAPE, 10.09%) [13]. Physiological pressure sensors and sweat-based biosensors have also been reviewed as monitoring channels within the wider wearable biosensor field [3]. Applying explainable artificial intelligence (XAI) methods, such as feature-importance and surrogate-model techniques, to wearable-derived data identifies which signals contribute most to a prediction, improving interpretability for depressive-symptom classification [6].

### *D. Comparison of Existing Systems*

Table I compares the reviewed system- and architecture-level studies, while component-validation studies [4], [5], [9] are excluded because they support individual sensor choices, and complete wearable systems are discussed in Section III-D as hardware precedents. Migraine studies report forecasting signals and models [1], and general architectures address communication and cloud infrastructure [2], [10]. Each therefore covers only part of the gap stated in Section I-C.

TABLE I
COMPARISON OF THE PROPOSED ARCHITECTURE AGAINST SYSTEM- AND ARCHITECTURE-LEVEL STUDIES IDENTIFIED IN THE REVIEWED LITERATURE

| Study | Sensors | AI | Explainability | Modularity | Neurological Focus |
|---|---|---|---|---|---|
| [1] | Multimodal, aggregated across 62 reviewed studies (EEG, imaging, EDA, skin temp., accelerometer, HR/heart rate variability (HRV)), not one device | Machine learning (ML) and deep learning (DL) models for diagnosis and attack forecasting | Identified as a future need, no XAI implementation reported | Narrative review, no system architecture proposed | Headache disorders, mainly migraine |
| [2] | Broad biomedical IoT sensor survey (immuno-, DNA-, nano-biosensors), not migraine-specific | Mentioned generally, no model built or evaluated | Not reported | Generic three-layer medical IoT architecture (perception, network, and application layers) | Brief Alzheimer's and Parkinson's biosensor examples only |
| [6] | Multimodal smartwatch (step, HR, sleep), smartphone-app behavior, and survey/ecological momentary assessment (EMA) data (61 features) | Hybrid convolutional neural network (CNN) and multilayer perceptron (MLP) models combined with LightGBM ensembles | SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) feature explanations | Fixed pipeline for a single prediction task | Depressive symptoms (psychiatric, not neurological) |
| [7] | Plantar-pressure and accelerometer shoe, LPWAN | Conceptual edge AI only, no model implemented | Not reported | Single closed-loop concept device | Huntington's disease, with proposed (untested) extension to Parkinson's disease and MS |
| [8] | Single smartwatch (Withings ScanWatch), step count and sleep analyzed | No predictive model, descriptive case narratives | Not reported | Case-study reports, not a system | MS, MG, CIDP, and migraine |
| [10] | Multimodal across many unrelated reviewed devices (BP, tremor, nutrition, sleep, etc.) | Some reviewed devices report accuracy, no unified framework | Not reported | Generic four-tier architecture (wearable sensor, mobile device, mobile network, and medical cloud) | Tremor/Parkinson's and stroke among 11 reviewed aspects |
| [13] | Thermal e-skin (hot/cold film) and accelerometer headband, Wi-Fi | CNN estimating total metabolic energy cost (MAPE 10.09%) | Not reported | Single fixed headband (MetaBand) | Exercise and metabolic monitoring, not neurological |
| Proposed Architecture | PPG, EDA, skin temperature, inertial measurement unit (IMU) with accelerometer and gyroscope for Prototype V1, EEG as a separate Prototy | Supervised ML on fused multimodal features, namely HRV, EDA amplitude/frequency, temperature trend, and motion, for migraine risk classification | XAI as a mandatory core layer before clinical decision support, addressing the black-box concern raised in [1] and [6] | Nine explicitly separated, independently replaceable layers (sensor, acquisition, conditioning, synchronization, edge processing, communication, AI, decision support, and visualization) | Migraine as the initial case, with proposed extension to other neurological diseases |

| | | | | | |
|---|---|---|---|---|---|
| | pe V2 module, BLE | | | | |

## III. Proposed Instrumentation-Centered Framework

### *A. Framework Overview*

The framework was derived in four steps. First, 19 studies on migraine AI and wearables, wearable neurological monitoring, sensor validation, XAI, and IoT architectures were reviewed. Second, the physiological inputs used in wearable migraine forecasting [1] were taken as signal requirements. Third, each sensing channel was matched to studies reporting validation against a reference instrument or annotated database [4], [5], [9], [13]–[16]. Fourth, sampling rates and filter bands from these studies were used to specify the signal chain and data rate (Section V). The proposed framework is organized as a layered, modular architecture rather than a single end-to-end model, with each of the nine layers defined in Section III-B performing one function (Fig. 1). Modularity is intended to allow individual layers to be replaced or extended without redesigning the entire system, for example, adding a sensor without altering the AI layer, consistent with layered wearable IoT architectures that separate sensing, edge processing, communication, and cloud analytics into distinct tiers [2], [10]. Migraine risk prediction is mapped onto this architecture as the first implementation case, without constraining the architecture to migraine alone.

The design uses modular hardware and signal layers instead of one AI model, for three reasons. First, keeping instrumentation separate makes the sensor pipeline reusable across disorders and datasets, since models trained on small datasets often generalize poorly to new populations [1]. Second, this separation avoids locking the framework to one algorithm or dataset, addressing how migraine AI studies often report single-model, single-dataset results instead of reusable systems [1]. Third, hardware limits like sampling rate, power budget, and on-device processing decide what is feasible before any AI stage can be designed, so instrumentation is the main constraint and AI is an interchangeable part built on top of it, matching layered wearable IoT design [2], [10].

### *B. Instrumentation Architecture*

The instrumentation architecture is organized into nine sequential layers, namely sensor, signal acquisition, signal conditioning, synchronization, local edge processing, communication, AI, clinical decision support, and visualization. This nine-layer split follows a single-responsibility principle, in which each layer has a single function boundary so it can be changed independently. Acquisition and conditioning are separated from synchronization because filtering is channel-specific, while synchronization only makes sense once every channel is already conditioned. Edge processing is separated from communication since one is limited by microcontroller resources and the other by wireless power budget [2], [7]. The AI layer is kept separate from clinical decision support and explainability so the model can be retrained or extended without changing how its output is interpreted, addressing the reported "black-box" concern in headache AI research [1], [6]. Visualization sits last since it changes most often as reporting needs evolve, without touching upstream layers.

The sensor layer captures raw physiological and motion signals directly from the body. The signal acquisition layer performs analog-to-digital conversion, requiring each channel to be sampled at least twice its highest relevant frequency component. The signal conditioning layer applies a 0.5 to 5 Hz band-pass filter to PPG, as used in [4], and a 3 Hz low-pass filter to EDA, as used in [5]. The sampling settings are given in Section V-B. The synchronization layer is intended to align signals from sensors with different sampling rates onto a common time base before feature extraction. Local edge processing on a microcontroller-based platform has been used in comparable wearable devices to perform preliminary signal handling before sending text-message alerts to caregivers [14].

### *C. Multimodal Sensor Architecture*

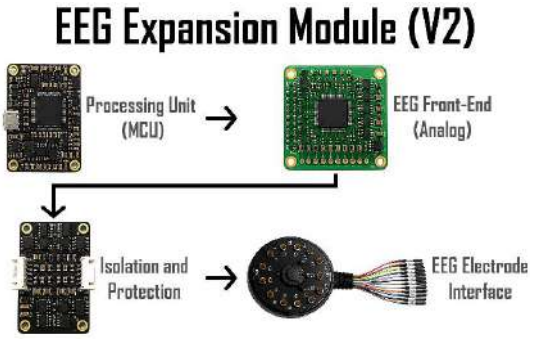



Fig. 2. Conceptual EEG expansion module (Prototype V2) proposed for future integration into the framework.

PPG measures cardiovascular activity by tracking periodic shifts in light absorption tied to pulsatile arterial blood volume. HRV derived from PPG is used to assess the state of the autonomic nervous system, which requires accurate beat detection [4]. The dominant error source in wearable PPG is motion artifact, which has been reduced using multi-channel, multi-wavelength PPG with independent component analysis, with an accelerometer used to detect motion [4]. A QRS-detection algorithm originally developed for wearable electrocardiogram (ECG) sensing has been designed for low computational complexity, a principle that may be adapted to PPG peak detection, although it was validated on ECG [15].

EDA measures skin conductance changes associated with eccrine sweat gland activity, controlled by the sympathetic branch of the autonomic nervous system [5]. A wrist-worn EDA sensor has been validated against a Food and Drug Administration (FDA)-approved reference system (median r = 0.93 to 0.99), with Ag/AgCl electrodes performing better than conductive fabric [5], whereas a finger-electrode skin-response module showed low validity (r = 0.298) during exercise [16]. This supports its inclusion as a channel reflecting predominantly sympathetic activity, complementary to the mixed autonomic information carried by PPG.

Skin temperature is associated with vasomotor activity, as skin blood flow and skin temperature increase when skin vessels dilate [9]. Off-the-shelf wearable sensors at the axilla and ear measured body temperature with a bias below 1 °C against hospital-grade thermometers, whereas the wrist does not represent core body temperature [9]. Wrist skin temperature is

therefore used here only as a relative trend. It is included as a slower-changing channel, consistent with comparable multi-sensor wearables that combine temperature with cardiovascular and electrodermal sensing [16].

Motion sensing provided by the accelerometer and gyroscope (IMU) captures acceleration and angular-velocity data used primarily for motion-artifact detection rather than direct neurological feature extraction. Acceleration has been used to detect motion before PPG artifact reduction [4] and as the reference signal of an adaptive filter that removed motion artifacts from a pulse signal [13]. Reduced physical activity has also been reported as a behavioral correlate around migraine attack days, giving motion data secondary relevance as a feature source [1].

Together, these four channels cover the input types used in wearable migraine forecasting models [1] rather than isolated measurements. PPG and EDA jointly capture autonomic activity, temperature adds a slower vasomotor trend, and IMU adds the behavioral dimension reported around attack days while also supporting motion-artifact correction on the other channels [1], [4], [8].

EEG (Fig. 2) is proposed as an optional future expansion module rather than a Prototype V1 requirement, given its comparatively higher sampling rate and its dedicated amplifier and electrode requirements. Consumer and research-grade wearable EEG devices have been self-applied at home for sleep, mental-health, and neurological monitoring and for neurofeedback, supporting future integration [11]. Separately, high-channel-count, low-power neural recording circuits in the bioelectronics literature illustrate a broader trend toward miniaturized neural signal acquisition hardware, although this specific circuit was developed for implantable rather than wearable applications [17].

*D. Hardware and Software Architecture*

Prototype V1 hardware is proposed around a low-power microcontroller unit (MCU) with integrated wireless connectivity (Fig. 3), following comparable wearables that combine an ESP32 MCU with HR, temperature, and motion sensors for injury and temperature risk monitoring [14], or with HR, skin-temperature, and skin-response sensors for heat-strain monitoring [16]. BLE is proposed for wireless communication because it is supported by smartphones and consumes less energy than classic Bluetooth, although its bandwidth is limited [2]. The software architecture separates firmware, communication, mobile application, backend, database, and AI layers into distinct modules. The firmware is intended to run the sensors and keep track of the timing of collected data. It also stores the data temporarily before sending it through BLE. The mobile app is intended to connect with the device, allow local viewing of the data, and securely send the records to the backend system [18].

*E. Signal Processing and AI Framework*

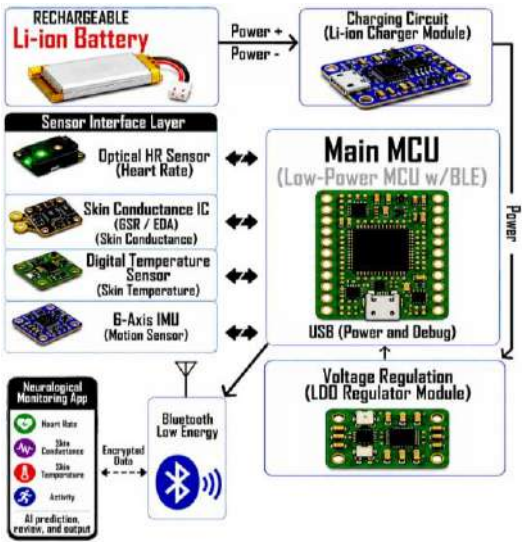



Fig. 3. Conceptual hardware architecture proposed for Prototype V1, showing the sensor interface,

It is proposed that raw multimodal signals first undergo signal quality assessment, in which segments would be evaluated against a signal quality index before being accepted for further processing. This follows general biomedical signal processing practice, in which excluding poor-quality segments is intended to reduce the amount of unreliable data that reaches downstream feature extraction, in line with data-validation practice for mobile health data [18]. Motion-artifact removal is proposed next, using IMU data to identify motion-affected PPG segments [4], with accelerometer-referenced adaptive filtering, as demonstrated in a head-worn wearable [13], as a candidate correction method. Feature extraction is intended to follow physiological markers commonly reported in autonomic nervous system research, including HRV time- and frequency-domain measures, EDA response amplitude and frequency, and skin temperature trend, before multimodal feature fusion combines these channel-specific features for prediction. A comparable multimodal wearable platform reported its best estimate only when thermal and motion-derived inputs were combined [13]. A supervised ML model is proposed for migraine risk classification from the fused feature set, informed by hybrid DL approaches already applied to comparable wearable-derived physiological and behavioral features for a different condition [6]. XAI methods are proposed for the output stage, using feature-importance techniques such as those demonstrated for wearable-derived depressive-symptom classification, to indicate which channels contributed most to a given prediction [6]. SHAP is proposed for global and LIME for individual explanations, and evaluation is proposed on subject-independent data, such as leave-one-subject-out validation [13], and an external test cohort [1], using accuracy, precision, recall, F1-score, and area under the receiver

operating characteristic (ROC) curve [6]. The resulting risk score is intended to be paired with its contributing features and a confidence estimate before reaching the clinical decision layer, consistent with reported concerns about "black-box" AI outputs limiting clinical acceptance in headache research [1].

## IV. Migraine Case and Prototype Concept

### A. Migraine as the Initial Implementation Case

Migraine risk prediction, rather than migraine detection, is selected as the initial demonstration case because physiological and behavioral changes around attack days and wearable attack forecasting have been reported, giving the architecture a concrete target for early-warning rather than diagnostic use [1]. Arm- or head-worn neuromodulation devices are used for acute headache treatment [12], suggesting that body-worn hardware is practical, although acceptance of a monitoring-only wearable still needs to be tested. Reduced activity and altered sleep around migraine attacks, observed in consumer wearable data in at least one reviewed case, further support the relevance of motion and behavioral signals within this architecture [8].

### B. Prototype V1 Concept

Prototype V1 is described here as a design concept rather than a validated device, applying the sensor suite and hardware architecture from Section III to migraine risk prediction as a wrist-worn unit integrating PPG, EDA, skin temperature, and IMU sensors around a single low-power MCU, following comparable multi-sensor wearable devices [14], [16]. BLE transmission to a smartphone is this paper's own design choice, not one demonstrated in those devices. EEG is excluded from Prototype V1 and reserved for a Prototype V2 expansion once wrist-based feasibility is established, supported by existing wearable EEG devices [11]. No hardware has been fabricated or tested, and the description is a literature-informed engineering target.

### C. Prototype Components

Table II lists the proposed hardware components and their functional roles.

TABLE II
HARDWARE COMPONENTS AND FUNCTIONAL ROLES PROPOSED FOR THE PROTOTYPE V1 WEARABLE NEUROLOGICAL MONITORING SYSTEM

| Module | Component | Function |
|---|---|---|
| MCU | Low-power microcontroller with BLE | Processing & wireless transfer [14], [16] |
| PPG | Optical HR sensor | HR & HRV [4] |
| EDA | Skin conductance sensor | Sympathetic activity [5] |
| Temperature | Digital skin temperature sensor | Relative vasomotor trend [9], [16] |
| IMU | Accelerometer & gyroscope | Motion & artifact detection [4], [13] |
| Communication | BLE | Wireless data transfer [2] |

### D. Operational Workflow

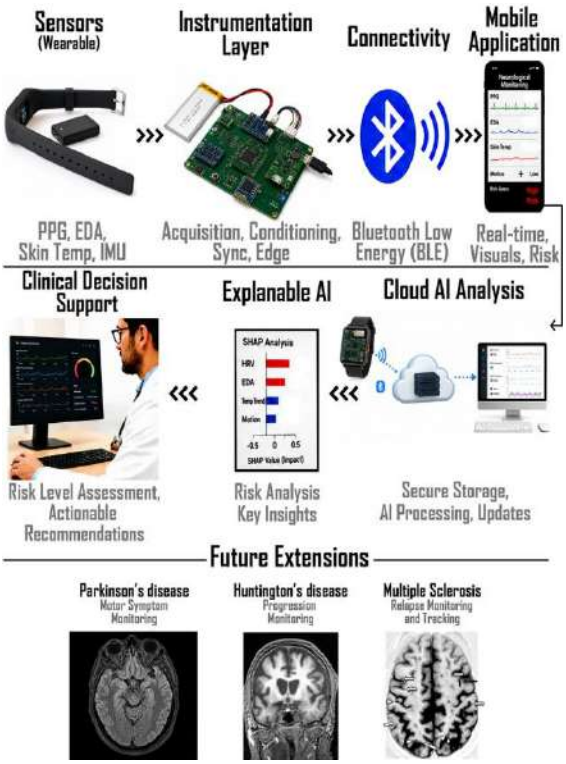



Fig. 4. Proposed operational workflow for Prototype V1, from physiological signal acquisition to explainable clinical decision

In the workflow (Fig. 4), the patient is intended to wear the wrist device continuously while sensors acquire physiological and motion data, which would be quality-checked, artifact-corrected, and synchronized at the edge before BLE transmission. The smartphone is intended to forward processed data to the AI layer, which would extract features, fuse modalities, and produce a confidence-scored migraine risk output. If the score exceeds a defined threshold, an early-warning notification would be generated while continuous monitoring proceeds. This is consistent with the forecasting, rather than diagnostic, framing of the reviewed migraine wearable studies [1].

## V. Design Evaluation

### A. Evidence for Each Module

As no hardware has been built, the design is evaluated against published evidence. Table III lists, for each module, quantitative results reported in the cited studies and the main reported limitation. These values come from other devices and populations, so they indicate the expected capability of each channel, not the performance of Prototype V1.

TABLE III
REPORTED EVIDENCE AND LIMITATIONS FOR EACH PROTOTYPE V1 MODULE

| Module | Reported evidence | Reported limitation |
|---|---|---|
| PPG | Beat sensitivity 99.00% (walking), 96.28% (fast walking) [4], HR accuracy 92.5% [14] | Sensitivity 82.49% when running [4], HR limits of agreement about ±20 bpm [16] |
| EDA | Median r = 0.93 to 0.99 vs. FDA-approved system [5] | Finger-electrode module r = 0.298 during exercise [16] |
| Temperature | Axillary and ear devices, bias < 1 °C [9] | Wrist does not represent core temperature [9] |
| IMU / fusion | Best estimate only with all inputs, MAPE 10.09% [13] | MAPE 16.06% for unseen subjects [13] |
| XAI | Selected-feature models outperformed full-feature models (best accuracy 93.43%), SHAP/LIME explained features [6] | Depression, not migraine [6] |

### *B. Signal Specification and Data Rate*

Using the sampling settings of the cited sensor studies, PPG at 100 Hz and 24 bit [4], EDA at 32 Hz and 12 bit [5], and a six-axis IMU at 100 Hz and 16 bit [4], [13], with temperature at 1 Hz and 16 bit and the gyroscope rate set equal to the accelerometer rate as design choices, the raw data rate is 12.4 kbit/s, or about 134 MB per day. Each rate exceeds twice the relevant signal band (PPG 0.5 to 5 Hz and motion artifacts up to 10 Hz [4], and EDA filtered at 3 Hz [5]). If the edge layer instead sends 18 window features of 32 bits every 60 s, the rate falls to 9.6 bit/s, about 99.9% less. The 18 features are a design choice. From PPG, they are mean HR, the standard deviation of beat intervals, the root mean square of successive beat-interval differences, low-frequency power, high-frequency power, and their ratio. From EDA, they are mean skin conductance level, number of responses, and mean response amplitude. From temperature, they are mean, standard deviation, and slope. From the IMU, they are mean and standard deviation of acceleration magnitude and of angular-velocity magnitude, activity count, and the fraction of motion-flagged samples. This matters because the radio is the most power-consuming part of an IoT device [7] and BLE bandwidth is limited [2]. Battery life could not be calculated, because the current consumption of the components was not reported in the reviewed studies.

### *C. Comparison with Existing Results*

Wearable migraine forecasting studies summarized in [1] reported test-set accuracies of 62% to 84% in 7 to 18 patients, and the most accurate used accelerometer, temperature, EDA, blood volume pulse, HR, and HRV inputs. Prototype V1 provides these input types and adds explainability and a reusable architecture, but its accuracy is not yet known and none is claimed.

## VI. Discussion

### *A. Engineering Contributions*

The primary contribution of this paper is an instrumentation-centered, modular architecture rather than a new sensor, algorithm, or dataset, in which separating sensing, acquisition, AI, and clinical decision support layers is intended to allow components to be replaced or upgraded independently, for example, adding EEG in a later prototype version without redesigning the full system. This modularity follows layered design principles in wearable IoT and mobile health architectures, applied here to neurological early-warning signal monitoring [2], [18]. Positioning XAI as a core architectural layer, rather than an optional addition, is intended to address the "black-box" acceptance concern reported in headache AI research, distinguishing this framework from the prediction-only migraine studies reviewed above [1], [6].

### *B. Strengths and Current Limitations*

The proposed architecture is modular, explainable by design, and grounded in physiological and instrumentation principles reported in the reviewed studies, with each sensor and hardware choice supported by a comparable implementation in the wearable sensing literature [4], [5], [9], [14], [16]. The framework remains, however, a conceptual engineering design rather than a fabricated or clinically tested device. No hardware prototype has been built, no signal data have been collected, and no migraine prediction accuracy is claimed at this stage. Sensor-level accuracy, power consumption, and comfort trade-offs reported for comparable devices have not been independently verified for this specific component combination [9]. The evaluation in Section V is literature-based. The cited evidence comes from other populations, such as healthy adults [4], [13], triathletes [16], and construction workers [9]. PPG beat sensitivity fell to 82.49% during running [4], the forecasting results come from cohorts of 7 to 18 patients with limited external validation [1], and activity data can be confounded by holidays, infections, or mobility aids [8]. The 19 studies were not selected through a systematic review protocol.

### *C. Future Research*

Future work is intended to implement Prototype V1 as physical hardware and validate it in stages. Each channel would first be compared with reference instruments, as in [5], [16], then tested with a motion protocol in healthy volunteers, as in [4], and finally evaluated in a prospective migraine cohort on subject-independent and external data [1]. Prototype V2 would add the optional EEG module, drawing on wearable EEG approaches [11] and low-power neural recording circuits [17]. Interstitial-fluid microneedle biosensors [19] could further extend the sensor layer beyond the current noninvasive scope. Beyond migraine, the architecture is intended to extend toward Parkinson's disease, MS, and other neurological applications [7], supported by the broader IoT monitoring landscape [2].

## VII. Conclusion

This paper proposed an instrumentation-centered, modular multimodal AI framework for continuous wearable neurological monitoring, developed because several existing systems are often built around single-modality sensing and disorder-specific prediction algorithms. Nine architectural layers were defined, spanning sensing, signal acquisition, conditioning, synchronization, edge processing, communication, AI, clinical decision support, and visualization. Migraine risk prediction, rather than migraine detection, was used as the initial

implementation case, motivated by its reported prevalence, together with the physiological and behavioral changes reported around attack days. A conceptual Prototype V1 design was presented, integrating PPG, EDA, skin temperature, and inertial motion sensing around a low-power MCU and BLE communication, with XAI proposed as a core architectural layer rather than an optional feature to support clinical interpretability of a predicted risk score. A literature-based evaluation showed that the sensor set provides the input types of the most accurate reported wearable migraine-forecasting model and that edge feature extraction can reduce the transmitted data rate by about 99.9%, while no prediction accuracy is claimed. The contribution of this work is the architecture and its design rationale, intended as a foundation for future prototyping, pilot testing, and expansion toward additional neurological monitoring applications.

## REFERENCES


[1] A. Danelakis, A. Stubberud, E. Tronvik, and M. Matharu, "The emerging clinical relevance of artificial intelligence, data science, and wearable devices in headache: A narrative review," Life, vol. 15, no. 6, Art. no. 909, Jun. 2025, doi: 10.3390/life15060909.

[2] M. Aledhari, R. Razzak, B. Qolomany, A. Al-Fuqaha, and F. Saeed, "Biomedical IoT: Enabling technologies, architectural elements, challenges, and future directions," IEEE Access, vol. 10, pp. 31306–31339, 2022, doi: 10.1109/ACCESS.2022.3159235.

[3] Z. Song et al., "Flexible and wearable biosensors for monitoring health conditions," Biosensors, vol. 13, no. 6, Art. no. 630, Jun. 2023, doi: 10.3390/bios13060630.

[4] J. Lee, M. Kim, H.-K. Park, and I. Y. Kim, "Motion artifact reduction in wearable photoplethysmography based on multi-channel sensors with multiple wavelengths," Sensors, vol. 20, no. 5, Art. no. 1493, Mar. 2020, doi: 10.3390/s20051493.

[5] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," IEEE Trans. Biomed. Eng., vol. 57, no. 5, pp. 1243–1252, May 2010, doi: 10.1109/TBME.2009.2038487.

[6] J. Ko et al., "Interpretable feature selection and hybrid deep learning models for depressive symptoms prediction from wearable device data," J. Med. Syst., vol. 50, no. 1, Art. no. 26, 2026, doi: 10.1007/s10916-026-02354-9.

[7] Orfanidis, A. S. Darwich, R. Cheong, and X. Fafoutis, "Monitoring neurological disorders with AI-enabled wearable systems," in Proc. Emerging Devices for Digital Biomarkers (DigiBiom '22), Portland, OR, USA, Jul. 2022, pp. 24–28, doi: 10.1145/3539494.3542755.

[8] P. Z. Epping et al., "Case-based lessons on remote patient monitoring in neurology using consumer-grade wearables," J. Cent. Nerv. Syst. Dis., vol. 18, Art. no. 11795735261419641, 2026, doi: 10.1177/11795735261419641.

[9] F. Nasirzadeh et al., "Continuous monitoring of body temperature for objective detection of health and safety risks in construction sites: An analysis of the accuracy and comfort of off-the-shelf wearable sensors," Heliyon, vol. 10, no. 6, Art. no. e26947, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26947.

[10] N. Surantha, P. Atmaja, David, and M. Wicaksono, "A review of wearable Internet-of-Things device for healthcare," Procedia Comput. Sci., vol. 179, pp. 936–943, 2021, doi: 10.1016/j.procs.2021.01.083.

[11] M. Emish and S. D. Young, "Remote wearable neuroimaging devices for health monitoring and neurophenotyping: A scoping review," Biomimetics, vol. 9, no. 4, Art. no. 237, Apr. 2024, doi: 10.3390/biomimetics9040237.

[12] U. A. Khan, "Design and development of next-generation wearable devices for on-demand headache relief," Int. J. Eng. Technol. Res. Manage., vol. 8, no. 9, pp. 42–45, Sep. 2024, doi: 10.5281/zenodo.15250454.

[13] S. Liu et al., "A smart headband for multimodal physiological monitoring in human exercises," Adv. Sci., vol. 13, no. 22, Art. no. e74279, 2026, doi: 10.1002/advs.74279.

[14] J. S. B. Valdehueza, N. B. M. Ocay, D. R. A. Aratea, and N. K. G. Castro, "ESP32 based wearable wrist device for monitoring and managing risks in patients with congenital insensitivity to pain with anhidrosis (CIPA Sense-Band)," Int. J. Inf. Technol. Comput. Eng., vol. 6, no. 1, pp. 33–45, Jan.–Jun. 2026, doi: 10.55529/ijitc.61.33.45.

[15] Y. Wang, C. J. Deepu, and Y. Lian, "A computationally efficient QRS detection algorithm for wearable ECG sensors," in Proc. 33rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Boston, MA, USA, Aug.–Sep. 2011, pp. 5641–5644, doi: 10.1109/IEMBS.2011.6091365.

[16] K. Silawarawet, C. Trirattananurak, J. Muksuwan, S. Sangdao, D. Panich, and S. Saadprai, "Development of wearable heatstroke warning system (HeatGuard): Design, validation and controlled-environment testing among triathletes," Sensors, vol. 26, no. 8, Art. no. 2556, Apr. 2026, doi: 10.3390/s26082556.

[17] M. Jang et al., "A 1024-channel 268-nW/pixel 36×36 μm²/channel data-compressive neural recording IC for high-bandwidth brain-computer interfaces," IEEE J. Solid-State Circuits, vol. 59, no. 4, pp. 1123–1136, Apr. 2024, doi: 10.1109/JSSC.2023.3344798.

[18] Kumar, Y. Hasan, and S. Afroz, "Mobile health monitoring system: A comprehensive review," Int. J. Res. Publ. Rev., vol. 4, no. 6, pp. 1922–1954, Jun. 2023, doi: 10.55248/gengpi.4.623.45128.

[19] X. Yuan et al., "Epidermal wearable biosensors for the continuous monitoring of biomarkers of chronic disease in interstitial fluid," Micromachines, vol. 14, no. 7, Art. no. 1452, Jul. 2023, doi: 10.3390/mi14071452.

# Dual-Domain Deep Learning for MRI Brain Tumor Segmentation Using Spatial and FFT-Derived Frequency Representations

W A C K Wickramasinghe
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
ackwickramasinghe@students.nsbm.ac.lk

T A H Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Accurate brain tumor segmentation from MRI is essential for treatment planning, but manual delineation is slow and shows substantial inter-rater variability. MRI is physically acquired as raw frequency-domain K-space data before being reconstructed into the spatial images that current deep learning segmentation models, including U-Net, exclusively operate on; limited prior work has explored dual-domain (spatial + frequency) encoders for brain tumor segmentation, despite dual-domain approaches demonstrating clear benefit in MRI reconstruction. This study designs, trains, and evaluates a dual-domain U-Net with parallel spatial and FFT-derived frequency encoders fused at the bottleneck, compared against an architecturally matched spatial-only baseline under the same training protocol on the BraTS 2023 GLI dataset (T2-weighted input only). On a locked 187-case test set, the dual-domain model achieves a statistically significant improvement in Whole Tumor Dice Similarity Coefficient (0.8547 vs. 0.8513, p<0.0001) that survives Holm-Bonferroni correction for multiple comparisons; an uncorrected improvement in Enhancing Tumor 95th-percentile Hausdorff Distance (raw p=0.043, Holm-adjusted p=0.215) does not survive correction and is reported as a trend only, and no significant improvement was found for Tumor Core. Correlation, size-stratified, and Bland-Altman analyses show tumor volume is the strongest examined correlate of accuracy for both architectures, and that the dual-domain benefit is concentrated in boundary precision rather than volumetric calibration. These results support a narrow, evidence-based conclusion: adding a frequency branch produced a real but small, targeted benefit for brain tumor segmentation, not a uniform improvement across all sub-regions. This benefit is not isolated from the accompanying 1.67× increase in model capacity; each model was trained only once, and the real K-space check was a small pilot.**

***Keywords—Brain tumor segmentation, MRI, Deep Learning, K-space, Fourier transform***

## I. Introduction

Brain and central nervous system tumors represent a growing global health burden, with a 31.6% increase in CNS cancers among working-age adults projected by 2050 [1], disproportionately affecting low- and middle-income countries [2]. MRI is the clinical gold standard for imaging brain tumors, but radiologists must manually review hundreds of slices per patient, and inter-rater agreement among expert glioma annotators varies substantially (Dice 0.74–0.85) [3]. In Sri Lanka this problem is compounded by severe resource constraints: 0.78 radiologists per 100,000 population against a recommended target of 2.5 [4], producing multi-month diagnostic delays that directly affect treatment outcomes for high-grade gliomas. Automated segmentation offers a path to reducing this burden.

MRI scanners do not directly acquire pixel intensities; they acquire K-space, the two-dimensional Fourier transform of the image, which is subsequently converted to the spatial image clinicians view via inverse FFT. Physics-informed reconstruction models that explicitly use K-space consistently outperform models that ignore it [5]. However, every existing brain tumor segmentation architecture, including foundational U-Net [6] and its many spatial-domain successors [7], [8], [9], [10], [11], [12], [13], [14] , operates exclusively on the reconstructed spatial image. Recent frequency-enhanced segmentation models [15], [16], [17], [18] incorporate FFT-derived features but have not been applied to brain tumor segmentation, have not used a dedicated parallel frequency encoder validated against real acquisition-domain K-space, and have not been evaluated on the BraTS benchmark. Note that FFT-simulated K-space, obtained by transforming an already reconstructed image, is only a single-channel proxy: unlike raw K-space, it lacks scanner noise, coil sensitivities and the sampling pattern, so how well it represents real acquisition data is an open question. This is the gap this study addresses.

This paper makes four contributions: (1) evaluation of a dual-domain (spatial + frequency) U-Net for brain tumor segmentation on the BraTS benchmark; (2) a controlled comparison against a spatial-only baseline sharing the same backbone design, isolating the addition of a frequency encoder as the primary architectural difference, while acknowledging this addition also increases parameter count by 1.67× (Section III-B), a confound discussed in Section IV-F; (3) a statistical validation methodology beyond mean DSC, including volume-accuracy correlation, size-stratified significance testing, and Bland-Altman volume-agreement analysis, with explicit multiple-comparison correction (Section III-D); and (4) a preliminary cross-domain compatibility analysis comparing the trained frequency encoder's response to real fastMRI K-space [19] against the FFT-simulated K-space it was trained on, using both feature-level and distributional comparisons, a comparison

not attempted by prior FFT-simulated frequency-segmentation work.

Research question: does a dual-domain model integrating FFT-derived frequency representations alongside spatial MRI data achieve a statistically significant improvement in segmentation accuracy, by DSC and HD95, over a spatial-only U-Net trained under identical conditions? The study's objectives were to:

I. identify the gap between current spatial-only segmentation models and dual-domain approaches established in MRI reconstruction;
II. analyze the BraTS 2023 GLI and fastMRI datasets to inform preprocessing and architecture decisions;
III. design and implement matched dual-domain and spatial-only U-Nets; and
IV. evaluate both against a locked, independent test set using DSC, HD95, and the deeper statistical analyses described in Section III-D.

## II. Related Work

U-Net [6] established the encoder-decoder-with-skip-connections paradigm for medical image segmentation; V-Net [7] extended it to 3D and introduced Dice loss to address class imbalance, and Pereira et al. [8] provided early empirical support for learned convolutional features over hand-crafted ones for glioma segmentation. Subsequent spatial-domain work has diversified architecturally: Walsh et al. [9] presented lightweight 2D U-Net variants suited to resource-constrained deployment; Kermi et al. [10] combined Dice with boundary-aware compound losses to improve sub-region delineation; Nancy and Maheswari [11] applied attention-augmented transfer learning on a VGG-19 backbone; Saifullah and Dreżewski [12] used genetic-algorithm-optimized architecture search with adversarial structural consistency; Lilhore et al. [13] developed a multi-scale attention-gated 3D CNN achieving strong results across all BraTS sub-regions; and Abdusalomov et al. [14] applied domain-adapted transfer learning for tumor classification. Yet all of this architectural diversity operates exclusively on reconstructed spatial images.

A parallel line of work has begun incorporating frequency-domain information into segmentation. tKFC-Net [17] used twin-kernel Fourier convolutions for complementary spectral features; MedFreq-Net [15] applied FFT-magnitude encoding to 3D organ segmentation; HybridMamba [16] combined frequency extraction with state-space sequence modeling; and Chen et al. [18] proposed adaptive frequency-spatial attention fusion at every decoder stage. All four simulate K-space via FFT of the reconstructed image, and none target brain tumors. In MRI reconstruction, DuDoRNet [20] demonstrated that parallel recurrent subnetworks processing real K-space and image-domain data jointly substantially outperform single-domain reconstruction, with Hammernik et al. [5] providing the supporting physics-informed rationale: models that explicitly encode the Fourier relationship between domains consistently outperform physics-uninformed alternatives. Table I summarizes the resulting gap: no existing model combines a dedicated dual-domain encoder, application to brain tumor segmentation, and validation on real acquisition-domain K-space. This is the intersection this study addresses.

TABLE I
Comparison with prior frequency/dual-domain work

| Study | Task | Freq. Input | Raw K-space |
|---|---|---|---|
| DuDoRNet [20] | Reconstr. | Real | Yes |
| tKFC-Net [17] | Seg. | FFT-sim | No |
| MedFreq-Net [15] | Seg. (3D) | FFT-sim | No |
| HybridMamba [16] | Seg. (3D) | FFT-sim | No |
| Chen et al.[18] | Seg. | FFT-sim | No |
| This Study | Brain Tumor Seg. | FFT-sim; real(Pilot) | Pilot only |

## III. Methodology

### A. Dataset and Preprocessing

The BraTS challenge, running continuously since 2012, has scaled from an initial cohort of a few dozen cases to the 1,251-case GLI training set used here, reflecting sustained community investment in benchmark-driven segmentation research. The BraTS 2023 GLI dataset provides these 1,251 fully annotated cases (four co-registered MRI sequences, pixel-level masks for necrotic core, edema, and enhancing tumor; 240×240×155 voxels, 1mm isotropic). Cases were split by patient ID into training, validation, and test sets, with the test set locked prior to model development (Table II summarizes the split). Only the T2-weighted sequence was used as spatial input, deliberately isolating the frequency-domain contribution as the sole experimental variable.

TABLE II
Patient-level dataset partition.

| Split | Cases | % of Total |
|---|---|---|
| Training | 876 | 70% |
| Validation | 188 | 15% |
| Test (locked) | 187 | 15% |

Volumes were per-case z-score normalized over non-background voxels (global normalization was rejected given a 434–3,706 intensity range across cases); only slices containing at least one tumor voxel were retained (~66 of 155 per volume, ~57,038 total training slices, roughly 14,259 batches per epoch at batch size 4). For each retained slice x, an FFT log-magnitude K-space map,

$$K = \log 2\,(1 + |\,\text{fftshift}\quad(\text{FFT2}(x))|)$$

was computed from the z-score-normalized image and min-max normalized per slice to [0,1], then verified via inverse-transform round-trip at $1.62\times10^{-7}$ reconstruction error. Both datasets are publicly available and fully de-identified prior to

release; BraTS 2023 is used under the Synapse challenge data use agreement and fastMRI under the fastMRI Dataset License Agreement, both permitting non-commercial academic research. No human participants were recruited and no new clinical data was collected; the resulting system is a research prototype, not validated for clinical deployment. BraTS multi-institutional data draws predominantly from large academic centers; population and scanner-protocol representativeness relative to lower-resource clinical settings, including the Sri Lankan context motivating this work, was not separately assessed and should not be assumed.

### *B. Model Architecture*

Both models share a 2D U-Net backbone with four encoder levels (64/128/256/512 channels) and a 1,024-channel bottleneck. The baseline is spatial-only. The dual-domain model adds a second, independently-weighted encoder of identical depth processing the FFT magnitude input; encoder outputs are concatenated (2,048 channels) and projected to 1,024 channels via 1×1 convolution before a shared decoder. The frequency branch uses log-magnitude only, discarding phase: this keeps the input real-valued and compatible with a standard CNN encoder, avoids the training instability complex-valued phase can introduce, and matches the magnitude-only convention used by every prior FFT-based segmentation model in Table I [15], [16], [17], [18]; this is a consistent, precedented choice, not a claim that it is optimal. Skip connections are drawn from the spatial encoder only, preserving spatial localization. Both networks output raw logits of shape (3, 240, 240); sigmoid activation is applied within the loss function and at inference time rather than inside the model itself, avoiding numerical instability from double-applying sigmoid during mixed-precision experimentation. Dropout (p=0.3) is applied at the bottleneck and the two deepest decoder levels of both architectures [3]. A 2D, slice-wise design was chosen over full 3D processing to keep the added computational cost of the parallel frequency encoder tractable on a single GPU, isolating the frequency-domain question from a simultaneous 2D-versus-3D confound; Section IV-F revisits this as a limitation and a direction for future work. Table III summarizes both architectures.

TABLE III
ARCHITECTURE COMPARISON.

| Component | Spatial (Baseline) | Dual-Domain |
|---|---|---|
| Spatial encoder | 4-level CNN | 4-level CNN |
| Frequency encoder | None | 4-level CNN, indep. |
| Bottleneck | 1,024 ch | 2,048→1,024 ch |
| Decoder | Shared | Identical |
| Parameters | 31,042,499 | 51,992,707 (1.67×) |

### *C. Training Protocol*

Both models were trained under identical conditions: combined Dice + BCEWithLogits loss (equal weighting, addressing a 96:1 background-to-tumor voxel imbalance), AdamW optimizer (β=0.9/0.999, weight decay 1e-4), cosine-annealed learning rate(T_max=50, η_min=1e-6), batch size 4, up to 100 epochs with early stopping (patience 15), gradient clipping at norm 1.0, and augmentation (horizontal and vertical flip, ±10% intensity scaling, each applied with probability 0.5). Final training used full fp32 precision (Section IV-A) coinciding with scheduled learning-rate reduction events. The baseline converged after 28 epochs (best validation WT Dice 0.8039); the dual-domain model converged after 34–35 epochs (0.8032). Both runs were resumed from earlier checkpoints (baseline from epoch 28, dual-domain from epoch 13) saved in sessions that used mixed precision and a different scheduler. The patient-level split used seed 42; no training seed was fixed.

### *D. Evaluation Framework*

Both models were evaluated once on the locked 187-case test set. Test volumes were processed directly from the raw files: every axial slice was passed through the network, sigmoid outputs were thresholded at a fixed 0.5, and slice predictions were reassembled into 3D volumes, on which the Dice Similarity Coefficient (DSC) and 95th-percentile Hausdorff Distance (HD95, mm) were computed for Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). No test-set information was used in preprocessing (normalization is per volume), hyperparameter selection, architecture design, threshold selection or checkpoint selection, which relied on validation WT DSC only. Significance was assessed via paired Wilcoxon signed-rank tests (α=0.05) rather than a paired t-test, since DSC is a bounded, non-normally distributed ratio metric for which the t-test's normality assumption is frequently violated; effect size was reported as rank-biserial correlation r, interpreted using conventional thresholds (r≈0.1 small, 0.3 medium, 0.5 large). The six primary region-metric comparisons in Section IV-B form one hypothesis family, Holm-Bonferroni corrected (α=0.05), with both raw and Holm-adjusted p-values reported; the nine size-tertile comparisons in Section IV-C form a separate, exploratory family, corrected the same way. HD95 is undefined when exactly one of the predicted or ground-truth masks is empty; such cases are excluded from that model's HD95 mean and from any paired test involving them, identically for both models. When both masks are empty, HD95=0 (correct agreement) and the case is retained. Because baseline and dual-domain can have different complete-miss cases, per-model means may cover slightly different case counts (Section IV-B), while paired tests use only cases defined for both models. Beyond aggregate means, three further analyses were conducted: (i) Pearson/Spearman correlation between log-transformed ground-truth volume and per-case accuracy; (ii) size-tertile stratified comparison with per-tier significance testing; and (iii) Bland-Altman analysis of predicted-versus-ground-truth tumor volume agreement, to separately characterize boundary-overlap quality and volumetric calibration.

## IV. RESULTS AND DISCUSSION

### *A. Training Behavior*

Both curves showed rapid early improvement followed by a long plateau, with training loss continuing to decrease slowly after validation Dice stabilized, the expected signature of the generalization boundary beyond which further epochs mainly fit training-set-specific detail. Early mixed-precision (fp16) sessions produced NaN loss after 17–19 epochs, coinciding with

a scheduled learning-rate reduction; this was traced to an interaction between the abrupt learning-rate change and the dynamic loss-scaling mechanism, and was resolved by final training in full fp32 with a smoothly decaying cosine schedule.

### *B. Primary Test Set Performance*

Table IV reports DSC and HD95 for both models across all three regions (n=187, mean ± SD). The dual-domain model achieves higher mean DSC in all three regions and lower (better) HD95 in two of three. After Holm-Bonferroni correction across these six comparisons (Section III-D), only WT DSC remains significant ($p<0.0001$, $r=0.362$, medium effect). ET HD95 is nominally significant before correction ($p=0.043$) but not after it (Holm-adjusted $p=0.215$), so it is reported as a directional trend rather than a confirmed effect. TC shows no significant difference on either metric, and ET DSC falls short of significance even before correction ($p=0.087$).

TABLE IV
TEST SET PERFORMANCE (WILCOXON SIGNED-RANK, A=0.05).

| Region/Metric | Baseline | Dual-Domain | P (raw) | P (Holm) | Sig. |
|---|---|---|---|---|---|
| WT DSC | 0.851±0.117 | 0.855±0.117 | <0.0001 | <0.001 | Yes |
| WT HD95 | 11.05±21.84 | 10.53±21.54 | 0.426 | 0.426 | No |
| TC DSC | 0.717±0.209 | 0.719±0.216 | 0.144 | 0.348 | No |
| TC HD95 | 11.05±18.26 | 11.10±20.41 | 0.103 | 0.348 | No |
| ET DSC | 0.552±0.220 | 0.555±0.227 | 0.087 | 0.348 | No |
| ET HD95 | 13.68±19.23 | 13.28±22.05 | 0.043 | 0.215 | No |

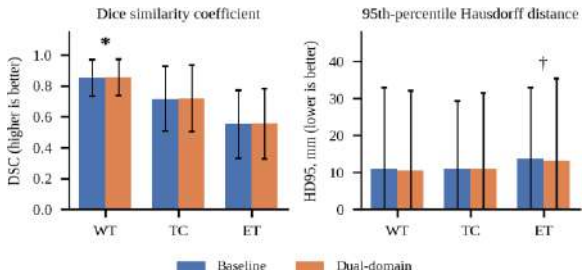



Fig. 1. Test set DSC and HD95 by region. * significant after Holm–Bonferroni correction; † nominal $p<0.05$ only (not significant after correction). Error bars: ±1 SD

This is a narrower finding than uniform improvement: adding the frequency branch provides a confirmed benefit, after correction for multiple comparisons, specifically for Whole Tumor delineation; the Enhancing Tumor HD95 trend and any Tumor Core effect are not statistically confirmed at this sample size. The achieved WT DSC (0.855) falls within the range of expert inter-rater agreement reported for complex glioma cases (0.74–0.85) [3]; this is a factual comparison rather than a claim of equivalence, since human-versus-human and model-versus-ground-truth agreement differ in annotation protocol, evaluation procedure, and case difficulty. The absolute WT gain is small ($\Delta$DSC=0.0034, about 0.4% relative), and statistical significance should not be read as clinical relevance. TC and ET accuracy remain well below this range regardless.

HD95 is undefined when a model predicts an entirely empty region or misses a present region completely; these complete-miss cases are excluded from the HD95 means in Table IV but are worth reporting separately. WT had zero complete-miss cases for either model. TC had one complete-miss case, for the dual-domain model only. ET, the smallest and most heterogeneous region, had the most: 6 complete misses for baseline and 7 for dual-domain, out of 187 cases. This modest increase (one additional case) alongside dual-domain's lower mean ET HD95 is too small a difference to interpret confidently at this sample size, but is noted here rather than left implicit in the exclusion.

### *C. Volume-Accuracy Correlation and Size Stratification*

Ground-truth tumor volume (log-transformed) correlates strongly with per-case accuracy for both models across all regions, consistent with the known small-structure sensitivity of the Dice metric: for a fixed absolute boundary error, DSC penalizes small structures more heavily than large ones. Table V reports Pearson correlations between log-volume and both DSC and HD95; all twelve correlations are significant at $p<0.0001$.

TABLE V
PEARSON CORRELATION, LOG GROUND-TRUTH VOLUME VS. ACCURACY (ALL P<0.0001).

| Region | Model | r (DSC) | r (HD95) |
|---|---|---|---|
| WT | Baseline/ Dual | 0.685/0.665 | −0.764/−0.737 |
| TC | Baseline/ Dual | 0.590/0.575 | −0.539/−0.450 |
| ET | Baseline/ Dual | 0.752/0.742 | −0.577/−0.536 |

TABLE VI
MEAN DSC ± SD BY SIZE TERTILE (N≈62–63 PER TIER), PAIRED WILCOXON P PER TIER.

| Region | Tier | Baseline | Dual | p | p (Holm) |
|---|---|---|---|---|---|
| WT | Small | 0.768±.144 | 0.773±.142 | 0.032 | 0.26 |
| WT | Medium | 0.874±.087 | 0.878±.089 | 0.0005 | 0.0045 |
| WT | Large | 0.913±.044 | 0.914±.047 | 0.074 | 0.52 |
| TC | Small | 0.580±.242 | 0.578±.256 | 0.790 | 1.00 |
| TC | Medium | 0.800±.101 | 0.809±.097 | 0.078 | 0.52 |
| TC | Large | 0.772±.185 | 0.773±.186 | 0.426 | 1.00 |
| ET | Small | 0.362±.255 | 0.360±.264 | 0.967 | 1.00 |
| ET | Medium | 0.629±.116 | 0.636±.114 | 0.078 | 0.52 |
| ET | Large | 0.668±.107 | 0.673±.111 | 0.112 | 0.52 |

Size-tertile analysis (Table VI) divides test cases into equal-sized Small/Medium/Large groups by ground-truth volume, computed per region, with paired Wilcoxon testing within each tier.

This nine-test family (three regions x three tiers) is exploratory and also Holm-Bonferroni corrected ($\alpha=0.05$). Before correction, WT is nominally significant in Small and Medium; after correction, only WT Medium survives (raw $p=0.0005$, Holm-adjusted $p=0.0045$), and WT Small (raw

p=0.032, adjusted p=0.26) is a trend, not a confirmed effect. No tier reaches significance for TC or ET either way. A Kruskal-Wallis test confirms tumor size tier has a statistically significant effect on DSC for all regions ($p<0.0001$): smallest-to-largest tertile changes mean DSC by 0.14–0.31, an order of magnitude larger than the baseline-vs-dual gap at any tier (≤0.008). A continuous test of whether improvement magnitude itself correlates with volume found no relationship in any region (WT: $r=-0.139$, $p=0.058$; TC: $r=0.010$, $p=0.894$; ET: $r=0.040$, $p=0.590$). Tumor size was the strongest examined correlate of accuracy here, though this analysis did not test it against other explanatory variables.

At case level, the dual-domain model outperforms the baseline on 116 of 187 WT cases (62.0%; baseline 51, ties 20), 101 TC cases (54.0%; baseline 81, ties 5) and 97 ET cases (51.9%; baseline 75, ties 15), using a tie threshold of 0.001 DSC.

The dual-domain model wins on a majority of individual cases in every region, providing support for the aggregate finding beyond the mean difference alone, even where that mean difference is small, as for TC and ET.

### *D. Volume Agreement: Bland-Altman Analysis*

Predicted and ground-truth voxel counts were converted to cm³ and compared via Bland-Altman bias (mean of predicted − GT) and 95% limits of agreement (LoA, bias ± 1.96 SD). Both models share the same direction of bias per region: modest WT over-segmentation (+1.4 to +1.5 cm³), modest TC under-segmentation (−2.6 to −2.8 cm³), and the largest bias of the three regions for ET, a consistent +4.6 cm³ over-segmentation, suggesting these are properties of the shared backbone and T2w-only input rather than the frequency branch specifically.

TABLE VII
BLAND-ALTMAN BIAS AND 95% LIMITS OF AGREEMENT.

| Region | Model | Bias (cm³) | SD | 95% LoA |
|---|---|---|---|---|
| WT | Baseline/ Dual | +1.54/+1.45 | 10.86/11.18 | [−19.7,+22.8]/ [−20.5,+23.4] |
| TC | Baseline/ Dual | −2.55/−2.79 | 14.68/15.44 | [−31.3,+26.2]/ [−33.1,+27.5] |
| ET | Baseline/ Dual | +4.59/+4.60 | 7.90/8.36 | [−10.9,+20.1]/ [−11.8,+21.0] |

The baseline shows marginally tighter limits of agreement (lower SD) than dual-domain in every region, the opposite direction from the DSC results. This distinguishes two senses of "improvement" that DSC conflates: the frequency branch appears to refine boundary placement on cases already segmented reasonably well, rather than constraining total predicted volume more tightly on the hardest cases. Proportional-bias testing (volume vs. error magnitude) found no significant relationship for any model/region (p 0.140–0.875). TC shows the widest LoA of the three regions despite mid-range DSC, driven by outlier cases with −60 to −85 cm³ under-segmentation, indicating TC volume estimation is markedly less reliable than its DSC alone suggests. Coverage checks confirm the limits are well-calibrated: 93.0-95.7% of cases fall within the reported 95% LoA across all six region/model combinations, consistent with the nominal 95% target.

For context, the 2023 BraTS-GLI challenge winning ensemble [21], using multi-modal input, three model architectures, and heavy synthetic data augmentation, reported validation DSC of 0.9005 (WT), 0.8673 (TC), and 0.8509 (ET), well above the results reported here. This gap is expected and not a like-for-like comparison: this study deliberately restricts input to a single T2w sequence and a single lightweight architecture per condition to isolate the frequency-domain effect, at the cost of absolute competitiveness with ensembled, multi-modal, heavily-augmented state-of-the-art solutions. The relevant comparison for this study's claims is the controlled baseline-versus-dual-domain gap under matched conditions, not the absolute distance to published leaderboard results. For a second reference point, nnU-Net with multi-modal input reported 0.8895 (WT), 0.8506 (TC) and 0.8203 (ET) on the BraTS 2020 test set [22], again not a like-for-like comparison.

### *E. Qualitative Results and Research Question*

The single largest volume disagreement in the test set illustrates the dominant failure mode identified above. Case BraTS-GLI-01530-000 has a ground-truth WT volume of 17,082 voxels (~17.1 cm³, a small lesion); the dual-domain model predicted 66,512 voxels (~66.5 cm³), approximately 3.9× over-segmentation, corresponding to a per-case DSC of only 0.267. This is consistent with the pattern established in Sections IV-B and IV-C: small lesions are the primary failure mode for both architectures, and this case is an extreme instance of that general pattern rather than an isolated anomaly.

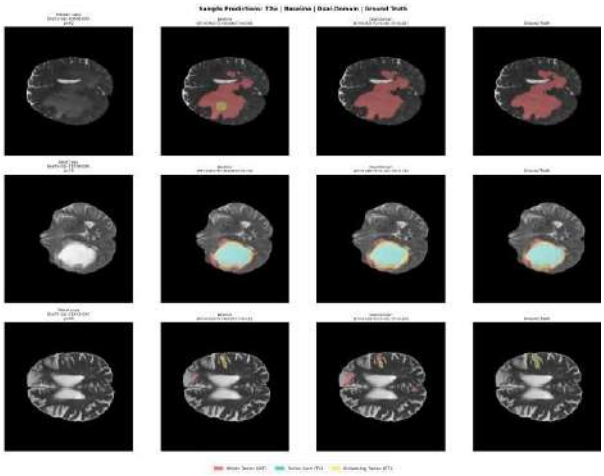


Fig. 2. Representative predictions: median, best, and worst DSC cases (T2w, baseline, dual-domain, ground truth).

The evidence supports a narrow, specific answer to the research question rather than a uniformly affirmative one: the dual-domain architecture achieves a statistically confirmed improvement only in WT DSC (after Holm–Bonferroni correction), with the WT benefit concentrated in small-to-medium tumors; the ET HD95 difference is a trend that does not survive correction; no significant improvement is confirmed for TC. The model wins on a majority of individual cases in every region, but does not achieve tighter volume calibration than the baseline in any region; its contribution is concentrated in boundary refinement rather than a uniform improvement across every dimension of segmentation quality.

### *F. Limitations*

Both models use only the T2w sequence, isolating the frequency contribution but limiting clinical representativeness; the comparatively low ET DSC (0.55) is consistent with the absence of T1-contrast input. Both architectures process 2D slices independently, foregoing 3D inter-slice context. The frequency input is FFT-simulated K-space rather than true raw acquisition data; a small-scale pilot fastMRI check (Section IV-G) found only partial overlap between real and simulated K-space, and full-scale validation remains future work. With n=187, TC and ET DSC comparisons remain underpowered: a larger test set may reveal these differences to be genuine but currently underpowered, or may confirm them as negligible, and this cannot be distinguished at the current sample size. The dual-domain model also has 1.67× more parameters than the baseline (Table III); this study evaluates adding a frequency branch as a whole and does not isolate frequency information from the accompanying increase in model capacity. A parameter-matched control was not implemented and is a priority for future work; until it exists, results are reported as "adding the frequency branch improved WT DSC," not as evidence that frequency information alone is responsible. Each model was also trained once, so initialization variance is not separately quantified from the architectural effect; the observed WT DSC difference (0.0034) is modest enough that some of it could reflect training variation rather than architecture alone. Paired confidence intervals were not computed, and performance on BraTS does not necessarily translate to clinical deployment across scanners and patient populations.

### *G. Status of the Secondary fastMRI Experiment*

This experiment is exploratory and does not support general conclusions. Two complementary preliminary checks were run on a small pilot batch (20 real fastMRI AXT2 files against a matched batch of BraTS-derived simulated K-space) to probe whether the FFT-simulation assumption holds against real acquisition data [23]. The fastMRI and BraTS cases are different subjects from different institutions, not paired acquisitions of the same anatomy; this is a general cross-domain compatibility check, not a validation that the model correctly reconstructs a specific real K-space input. First, a feature-level check: passing both real and FFT-simulated K-space through the trained frequency encoder yielded a mean cosine similarity of 0.66 between the resulting bottleneck feature maps, showing measurable but incomplete feature-level similarity; without a same-domain reference comparison (e.g. simulated-vs-simulated similarity across different cases), 0.66 should be read descriptively rather than against an established threshold. Second, a distributional check: raw log-magnitude K-space values differ substantially between the two sources (Kolmogorov-Smirnov D=0.998), with real K-space concentrating far more sharply near the center, consistent with the much higher dynamic range of raw scanner signal compared to FFT of an already-processed, normalized image. Both checks are based on a 20-file pilot and should be read as directional rather than conclusive.

## V. Conclusion and Future Work

This study presents, to our knowledge, the first evaluation of a dual-domain (spatial + frequency) U-Net for brain tumor segmentation on the BraTS benchmark, against a matched spatial-only baseline under controlled conditions on BraTS 2023 GLI. Adding the frequency branch raised Whole Tumor DSC from 0.8513 to 0.8547, a difference that remains statistically significant after multiple-comparison correction. Enhancing Tumor HD95 improved from 13.68 to 13.28 mm, but this is only a directional trend that does not survive correction (adjusted p=0.215), and Tumor Core showed no significant gain; this benefit is specific to boundary overlap, not volumetric calibration, and is not isolated from the accompanying 1.67× increase in parameter count, and each model was trained only once. Correlation and stratified analyses establish tumor size as the strongest examined correlate of accuracy for both architectures, context often omitted when only aggregate mean DSC is reported.

For resource-constrained settings such as Sri Lanka, where radiologist availability falls far below recommended density, a targeted, reproducible improvement in automated Whole Tumor delineation could potentially support future triage or pre-screening workflows, provided its boundaries are understood rather than overstated. Future work will (i) add a parameter-matched control and repeat training with different random seeds to separate frequency information from model capacity and training variation; (ii) extend the approach to multi-modal input (T1, T1-contrast, T2, FLAIR), which is expected to particularly benefit Enhancing Tumor delineation given T1-contrast's clinical role in defining that region; (iii) extend to full 3D volumetric processing to exploit inter-slice context; and (iv) investigate whether the boundary-precision benefit observed here compounds or saturates when combined with these richer inputs.

Reproducibility: both source datasets are publicly accessible under their respective data use agreements (Synapse challenge agreement for BraTS 2023; fastMRI Dataset License Agreement), and the preprocessing, training and evaluation code is publicly available at https://github.com/chithraka-kal/Dual-Domain-Brain-Tumor-Segmentation

## ACKNOWLEDGMENT

Data used in this study were obtained from the BraTS 2023 challenge via Synapse (syn51156910) and from the NYU fastMRI Initiative database (fastmri.med.nyu.edu) [19]. NYU fastMRI investigators provided data but did not participate in the analysis or writing of this report. Model training used GPU resources provided by Kaggle. The authors thank the Faculty of Computing, NSBM Green University, for institutional support.

## REFERENCES

[1] H. Chang, X. Yan, Q. Zhong, P. Ning, M. Weng, and Y. Liu, “Global burden of brain and central nervous system cancer among people aged 20-64 years, 1992-2021 and projections to 2050: A population-based study,” *BMJ Open* , vol. 15, no. 9, Sep. 2025, doi: 10.1136/bmjopen-2024-094462.

[2] Y. Fan *et al.*, “Burden and trends of brain and central nervous system cancer from 1990 to 2019 at the global, regional, and country levels,” *Archives of Public Health*, vol. 80, no. 1, Dec. 2022, doi: 10.1186/s13690-022-00965-5.

[3] B. Bonato, L. Nanni, and A. Bertoldo, “Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012–2025),” *Sensors*, vol. 25, no. 6, 2025, doi: 10.3390/s25061838.

[4] D. Karunaratne, "The Growing Problem of Radiologist Shortage in Sri Lanka," Oct. 01, 2023, *Korean Radiological Society*. doi: 10.3348/kjr.2023.0791.

[5] K. Hammernik *et al.*, "Physics-Driven Deep Learning for Computational Magnetic Resonance Imaging," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2203.12215

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[7] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, Institute of Electrical and Electronics Engineers Inc., Dec. 2016, pp. 565–571. doi: 10.1109/3DV.2016.79.

[8] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251, May 2016, doi: 10.1109/TMI.2016.2538465.

[9] J. Walsh, A. Othmani, M. Jain, and S. Dev, "Using U-Net Network for Efficient Brain Tumor Segmentation in MRI Images," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.01885

[10] A. Kermi, I. Mahmoudi, and M. T. Khadir, "Deep convolutional neural networks using U-Net for automatic brain tumor segmentation in multimodal MRI volumes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2019, pp. 37–48. doi: 10.1007/978-3-030-11726-9_4.

[11] A. M. Nancy and R. Maheswari, "Brain tumor segmentation and classification using transfer learning based CNN model with model agnostic concept interpretation," *Multimed. Tools Appl.*, vol. 84, no. 5, pp. 2509–2538, Feb. 2025, doi: 10.1007/s11042-024-20353-1.

[12] S. Saifullah and R. Dreżewski, "GA-UNet: Genetic Algorithm-Optimized Lightweight U-Net Architecture for Multi-Sequence Brain Tumor MRI Segmentation", doi: 10.1109/ACCESS.2024.0429000.

[13] U. K. Lilhore *et al.*, "AG-MS3D-CNN multiscale attention guided 3D convolutional neural network for robust brain tumor segmentation across MRI protocols," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-09351-x.

[14] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers (Basel).*, vol. 15, no. 16, Aug. 2023, doi: 10.3390/cancers15164172.

[15] X. Guan, H. Yang, H. Sun, P. Zhu, H. Yu, and K. Xia, "MedFreq-Net: Medical frequency network with hybrid CNN-transformer and enhanced features for 3D medical image segmentation," *Alexandria Engineering Journal*, vol. 131, pp. 218–231, Nov. 2025, doi: 10.1016/j.aej.2025.10.012.

[16] W. Wu, Z. Xing, J. Gong, Q. Peng, and L. Zhu, "HybridMamba: A Dual-domain Mamba for 3D Medical Image Segmentation," 2025. [Online]. Available: https://arxiv.org/abs/2509.14609

[17] X. Tang, J. Peng, B. Zhong, J. Li, and Z. Yan, "Introducing frequency representation into convolution neural networks for medical image segmentation via twin-Kernel Fourier convolution," *Comput. Methods Programs Biomed.*, vol. 205, Jun. 2021, doi: 10.1016/j.cmpb.2021.106110.

[18] D. Chen, J. Wu, X. Y. Zhang, and D. H. Wang, "A frequency-spatial dual perception network for efficient and accurate medical image segmentation," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-38093-7.

[19] J. Zbontar *et al.*, "fastMRI: An Open Dataset and Benchmarks for Accelerated MRI," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1811.08839

[20] B. Zhou and S. K. Zhou, "DuDoRNet: Learning a Dual-Domain Recurrent Network for Fast MRI Reconstruction with Deep T1 Prior."

[21] A. Ferreira *et al.*, "How we won BraTS 2023 Adult Glioma challenge? Just faking it! Enhanced Synthetic Data Augmentation and Model Ensemble for brain tumour segmentation," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2402.17317

[22] F. Isensee, P. F. Jaeger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for Brain Tumor Segmentation," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.00848

[23] F. Knoll, K. Hammernik, E. Kobler, T. Pock, M. P. Recht, and D. K. Sodickson, "Assessment of the generalization of learned image reconstruction and the potential for transfer learning," *Magn. Reson. Med.*, vol. 81, no. 1, pp. 116–128, Jan. 2019, doi: 10.1002/mrm.27355.

# Characterizing Electrophysiological Responses of Solanum Lycopersicum To Environmental Stress Using Machine Learning

R N I Basnayake
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
ackwickramasinghe@students.nsbm.ac.lk

G M S C Gajendrasinghe
*Department of Software Engineering*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
supun.g@nsbm.ac.lk

C Disanayake
*Department of Software Engineering*
*Faculty of Computing*
*NSBM Green University*
*Mahenwatta, Pitipana, Sri Lanka*
supun.g@nsbm.ac.lk

***Abstract*—Environmental stresses like not having water or not getting enough nutrients can greatly influence how Solanum lycopersicum grows and develops. The plants are reacting to these changes using cell and body processes. These include changes in how ions move and how the cell membranes handle electricity. These reactions can create signals related to electricity in the plant. These signals might show up before things, like leaves getting wilted or changing color or the plant growing slowly become visible. Therefore, plant electrophysiological activity has the potential to serve as a non-invasive indicator of early plant stress. This study proposes to investigate the electrophysiological responses of Solanum lycopersicum under controlled environmental stress conditions and to explore their potential for early stress detection. Non-invasive electrodes will be used to continuously record electrical signals from tomato plants exposed to selected stress conditions. The recorded signals will be preprocessed to reduce noise and extract relevant electrophysiological features. Machine learning techniques will then be applied to identify patterns in the signals and classify the corresponding stress conditions. The study will also consider observable physiological responses of the plants to establish a relationship between electrophysiological changes and plant stress. The expected outcome is to determine whether changes in plant electrical activity can provide an early indication of stress before the development of visible symptoms. The findings may contribute to a better understanding of plant electrophysiological responses and support the development of intelligent, non-invasive approaches for plant stress monitoring in precision agriculture.**

***Keywords—Solanum Lycopersicum, Plant Electrophysiology, Plant Stress Physiology, Electrophysiological Signals, Machine Learning***

## I. Introduction

Agriculture plays an important role in global food production, but crop plants are increasingly exposed to environmental stresses such as water deficiency, nutrient deficiency, salinity, and temperature changes. [1], [2] These stresses can affect plant growth, development, and physiological functioning, ultimately reducing crop productivity. [1], [3], [4] Tomato (Solanum lycopersicum) is an important crop species whose growth and physiological processes can be affected by changes in environmental conditions. [5], [6] Therefore, understanding how tomato plants respond to environmental stress is important for identifying stress at an early stage and supporting appropriate agricultural management. Conventional methods of plant stress detection mainly depend on visible symptoms such as leaf wilting, discoloration, and reduced growth. [5], [6] However, these symptoms may appear only after significant physiological changes have already occurred within the plant, limiting the opportunity for early intervention. [2], [6]. Plants have developed complex mechanisms to perceive and respond to changes in their surrounding environment. [1], [3], [4] When a plant experiences environmental stress, changes can occur at the cellular level, affecting water balance, ion movement, and membrane activity. [1], [4], [7]. Ions such as calcium, potassium, chloride, and hydrogen play important roles in maintaining cellular functions and regulating electrical activity within plant cells. [8], [9], [7]. Changes in ion transport across cell membranes can alter the electrical state of the cells and generate electrical signals that can propagate through different parts of the plant. [8], [9] These electrical responses are associated with physiological changes that occur when plants respond to environmental stimuli and stress. [1], [4], [10] Therefore, plant electrophysiological signals provide a potential means of observing physiological responses to stress before clear external symptoms become visible. [8], [1], [4], [10].

Tomato (*Solanum lycopersicum*) provides a suitable plant model for investigating electrophysiological responses to environmental stress. Previous studies have demonstrated that electrophysiological signals can be recorded from tomato plants and that changes in these signals are associated with different stress conditions.[5], [11], [6]. For example, electrophysiological measurements have been investigated for identifying general stress responses in commercial tomato plants, detecting spider-mite infestation, and monitoring water stress. These findings indicate that changes in the electrical activity of tomato plants can provide information about their

physiological status. However, further investigation is needed to understand whether electrophysiological patterns can be used to distinguish between specific environmental stress conditions and whether these changes can provide an early indication of stress before visible symptoms develop. [5], [11], [6].

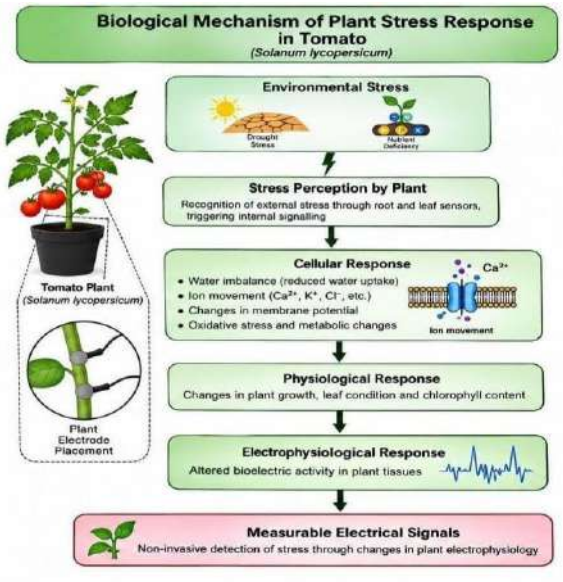



Figure 1. Biological mechanism of plant stress response and generation of electrophysiological signals in tomato

Recent advances in artificial intelligence and machine learning have created new opportunities for analyzing complex biological signals and identifying patterns that may not be easily recognized through conventional analysis.[12], [13], [14] Machine learning techniques can be applied to plant electrophysiological data to identify patterns associated with different physiological conditions and stress responses. However, existing research has largely focused on detecting general plant stress, while comparatively limited work has investigated the classification of specific stress conditions in tomato plants using electrophysiological signals. Therefore, this study proposes to investigate the relationship between environmental stress and electrophysiological responsesin *Solanum lycopersicum* and to develop a machine learning-based approach for identifying different stress conditions from these biological signals. [12], [5]

*1) RELATED WORK*

*A. Plant Stress Physiology*

Plants are continuously exposed to environmental conditions that can influence their growth, development, and physiological functions.[1], [3], [4] Environmental stresses such as water deficiency, nutrient deficiency, temperature variation, and other abiotic factors can disturb normal plant processes and affect plant productivity. Because plants cannot move away from unfavorable environmental conditions, they have developed mechanisms to perceive environmental changes and coordinate responses throughout the plant. At the cellular level, stress conditions can alter water balance and ion homeostasis, leading to changes in physiological processes and cellular signaling. These responses allow plants to adjust their functions in response to changing environmental conditions and contribute to their ability to tolerate or respond to stress. [1], [4], [10].

*B. Plant Electrophysiology*

Plant electrophysiology provides a means of studying electrical activity associated with physiological responses in plants.[8], [9] Previous research has shown that plants generate electrical signals in response to environmental stimuli, including drought, temperature changes, mechanical disturbance, and nutrient deficiency. These signals are associated with changes in plant physiological conditions and can provide information about the health and stress status of the plant.[8], [1], [10] Plant electrical signals include action potentials and variation potentials, which contribute to signaling and communication between different parts of the plant. Changes in environmental conditions can influence ion movement and membrane activity, resulting in measurable changes in the electrical signals produced by the plant. Recent studies have further investigated the use of electrophysiological signals for monitoring plant stress, including stress responses in tomato plants. These findings suggest that electrophysiological measurements may provide a non-invasive approach for monitoring physiological changes in plants and detecting stress responses. [5], [6], [15]

*C. Machine Learning for Plant Stress Detection*

Recent advances in machine learning have provided new approaches for analyzing plant electrophysiological signals and identifying patterns associated with different plant conditions.[12], [13] Machine learning algorithms can process complex electrophysiological data and classify signals according to the physiological condition of the plant.[16]

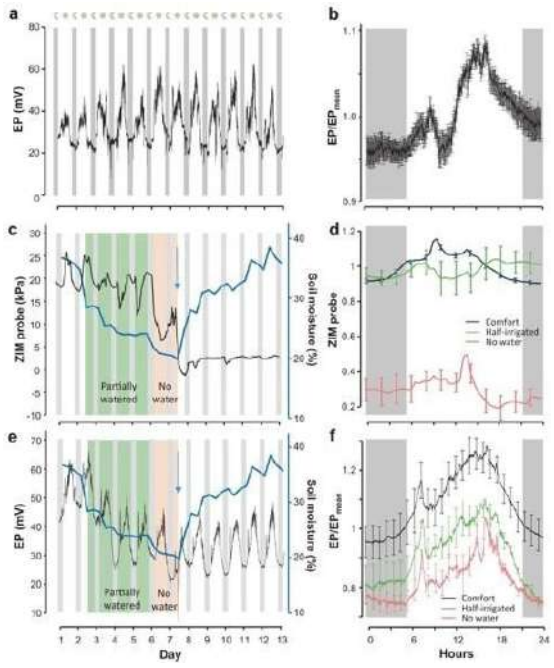



Figure 2. Electrophysiological responses of tomato plants (Solanum lycopersicum) under different irrigation conditions, showing changes in electrical potential and leaf turgor during water deficiency [12]

Previous studies have explored the use of machine learning for detecting plant stress and distinguishing electrophysiological responses associated with environmental stimuli. For example, machine learning has been applied to electrophysiological data obtained from plants to assess plant status, identify stress responses, and

classify electrical signals associated with different conditions. However, existing studies have mainly focused on detecting general plant stress or specific stress-related responses, while limited research has investigated the classification of multiple environmental stress conditions in tomato plants using non-invasive electrophysiological measurements. This indicates a need for further investigation into the relationship between specific stress conditions, plant physiological responses, and electrophysiological signal patterns using data-driven approaches. [2], [14]

*1) III RESEARCH GAP / PROBLEM STATEMENT*

Although previous studies have demonstrated that plant electrophysiological signals are associated with physiological responses to environmental stress, [8], [1], [4], [10] there is still a need to better understand how specific stress conditions influence the electrophysiological responses of tomato plants. [5], [6] Existing research has investigated electrical responses associated with general plant stress, drought, nutrient deficiency, and other environmental stimuli; however, the relationship between specific stress conditions and measurable electrophysiological patterns in Solanum lycopersicum remains insufficiently characterized. [5], [6] In addition, limited research has explored the use of machine learning to distinguish between different stress conditions based on non-invasive electrophysiological measurements. Therefore, there is a need for an integrated approach that investigates the biological and electrophysiological responses of tomato plants under controlled stress conditions and uses machine learning to identify patterns associated with these responses. Addressing this gap may contribute to the development of an early, non-invasive method for monitoring plant stress before the appearance of clear visible symptoms [17], [6]

*2) RESEARCH OBJECTIVES*

The main objective of this study is to investigate the electrophysiological responses of Solanum lycopersicum to environmental stress and explore their potential for early stress detection using machine learning.

The specific objectives are:

1. To investigate the physiological responses of tomato plants (*Solanum lycopersicum*) under selected environmental stress conditions.
2. To record and characterize electrophysiological signals produced by tomato plants under different stress conditions using non-invasive electrodes.
3. To identify relevant features and patterns within the recorded electrophysiological signals associated with different plant stress conditions.
4. To develop and evaluate machine learning models for classifying different stress conditions based on plant electrophysiological signals.
5. To investigate whether electrophysiological changes can provide an early indication of plant stress before the development of visible symptoms.

*3) PROPOSED METHODOLOGY*

*D. Experimental Plant Preparation*

Healthy tomato plants (*Solanum lycopersicum*) will be selected and grown under controlled environmental conditions before the experimental treatments are introduced. The plants will be maintained under consistent conditions as far as possible, including appropriate light, temperature, soil conditions, and nutrient availability. Plants of similar developmental stage and size will be selected to reduce biological variation between experimental groups. Before stress induction, electrophysiological signals and selected physiological characteristics will be recorded to establish a baseline condition for each plant. This baseline will be used to compare the physiological and electrophysiological changes observed after the application of the experimental stress conditions.

*E. Stress Induction and Experimental Groups*

The selected tomato plants will be divided into a control group and experimental stress groups. The control group will be maintained under normal growing conditions, while the experimental groups will be exposed to selected environmental stress conditions under controlled experimental settings. The stress conditions will include water deficiency and a selected nutrient deficiency, which will be applied by systematically reducing the availability of water or the required nutrient while

maintaining other environmental conditions as consistently as possible. The stress treatments will be monitored throughout the experimental period to observe changes in plant physiological condition and electrophysiological activity. Measurements obtained from the control plants will be used as a baseline for comparison with the responses observed in the stressed plants.

*F. Physiological Response Assessment*

The physiological responses of the tomato plants will be assessed throughout the experimental period to determine how the selected stress conditions affect plant condition. Observable characteristics such as leaf wilting, leaf discoloration, changes in plant growth, and other visible symptoms will be recorded at regular intervals. Where appropriate, additional physiological measurements such as plant height, leaf characteristics, and chlorophyll-related measurements may be collected to provide quantitative indicators of plant condition. These physiological observations will be compared with the electrophysiological signals recorded from the plants to investigate whether changes in electrical activity are associated with changes in plant physiological status. The physiological observations will also be used to determine whether electrophysiological changes occur before the appearance of clear visible stress symptoms.

*G. Electrophysiological Signal Acquisition*

Electrophysiological signals will be recorded from the tomato plants using a non-invasive measurement approach. Surface electrodes will be placed on suitable regions of the plant, such as the stem or leaves, to measure changes in the electrical activity associated with the plant's physiological responses to environmental stress. Signals will be continuously recorded using a suitable data acquisition system throughout the experimental period. Measurements will be collected from both control and stress-treated plants under consistent recording conditions. The recorded electrophysiological signals will be synchronized with the physiological observations and stress-treatment timeline to investigate the relationship between plant physiological changes and variations in electrical activity. Appropriate recording conditions will be maintained to minimize environmental and instrumentation-related interference during signal acquisition.

*H. Signal Processing and Feature Extraction*

The recorded electrophysiological signals will be preprocessed to reduce environmental interference and instrumentation noise before further analysis. The preprocessing stage will include signal filtering, segmentation, and feature extraction. Appropriate filtering techniques will be applied to improve the quality of the recorded signals while preserving relevant electrophysiological variations. The continuous signals will then be segmented into suitable time intervals for analysis. Relevant signal characteristics will be extracted from the segmented data to represent changes in plant electrophysiological activity under different stress conditions. The extracted features will be associated with the corresponding stress treatment and physiological observations to support subsequent machine learning analysis.

*I. Machine Learning-Based Stress Classification*

The extracted electrophysiological features will be used as input data for machine learning-based analysis. The dataset will be organized according to the corresponding experimental conditions and physiological observations. Machine learning models, including Random Forest, Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks, will be investigated for their ability to identify patterns associated with different plant stress conditions. The models will be trained and evaluated using appropriate data partitioning and performance evaluation measures to determine their ability to distinguish between the selected stress conditions. The classification results will then be compared with the observed physiological responses to assess whether electrophysiological features can provide meaningful information about the stress status of tomato plants.

*J. Early Stress Detection Assessment*

The potential of electrophysiological signals for early stress detection will be assessed by comparing the timing of changes in electrical activity with the timing of observable physiological symptoms. Electrophysiological signals recorded during the experimental period will be examined to identify changes that occur following the introduction of the stress conditions. These changes will be compared with physiological observations such as leaf wilting, discoloration, and changes in plant growth. The timing of electrophysiological changes and visible stress symptoms will

be analyzed to determine whether measurable electrical responses occur before clear external symptoms. The findings will be used to evaluate the potential of plant electrophysiological signals as an early, non-invasive indicator of stress in tomato plants.

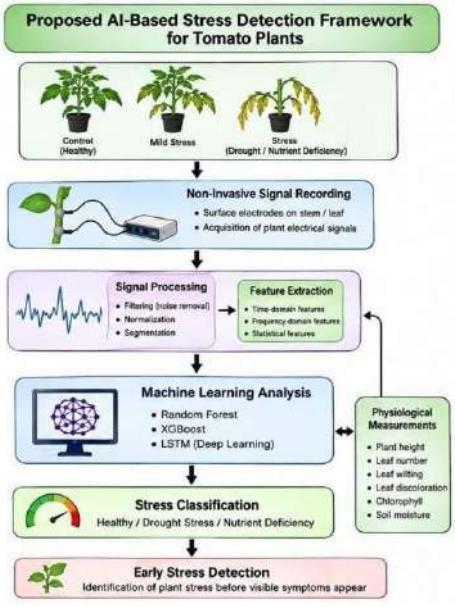



Figure 3. Proposed AI-based framework for early stress detection in tomato plants using electrophysiological signals and physiological responses

*1) EXPECTED RESULTS*

The proposed study is expected to identify how tomato plants (*Solanum lycopersicum*) respond to environmental stress conditions, particularly water and nutrient deficiencies. When plants do not receive water or essential nutrients they may show visible changes such as leaf wilting, discoloration and reduced growth. These physical changes are expected to be accompanied by variations in the signals produced by the plants.

By analyzing the recorded signals, the study aims to identify patterns that may be linked to different stress conditions. After preprocessing the data and extracting features differences in signal amplitude, frequency and changes over time are expected to provide useful information for distinguishing healthy plants from stressed plants.

The study also aims to explore how machine learning can be used to identify these stress conditions. Random Forest, XGBoost and Long Short-Term Memory (LSTM) models are proposed for this purpose. Their performance will be compared using evaluation metrics to understand how effectively each model can classify different plant conditions.

Another important aspect of the study is to explore whether changes in signals can be detected before visible symptoms appear. For example, if electrical changes occur before leaves begin to wilt or change color they may help identify plant stress at an earlier stage. Overall, the proposed approach is expected to demonstrate the potential of combining plant electrophysiology with machine learning to support non-invasive plant stress monitoring in agriculture.

*2) VII CONCLUSION*

This study proposes an integrated approach to investigating environmental stress responses in tomato plants (*Solanum lycopersicum*) through the combined analysis of physiological and electrophysiological responses using machine learning. The study focuses on understanding how selected environmental stress conditions influence the physiological condition of tomato plants and how these biological responses are reflected in measurable electrophysiological signals.

The proposed methodology involves controlled stress experiments, physiological response assessment, non-invasive electrophysiological signal acquisition, signal processing, feature extraction, and machine learning-based stress classification. By relating electrophysiological changes to observed physiological responses, the study aims to improve the understanding of the relationship between plant stress physiology and plant electrical activity.

The study is also expected to investigate whether electrophysiological changes can be detected before clear visible symptoms of stress develop. If such early responses can be identified consistently, plant electrophysiological signals could provide a promising non-invasive approach for early plant stress monitoring. The integration of plant biology, electrophysiology, and machine learning may therefore contribute to the development of intelligent and data-driven approaches for plant health monitoring and precision agriculture.

## REFERENCES


[1] M. Mudrilov et al., “Electrical signaling of plants under abiotic stressors,” Plants, vol. 10, no. 10, 2021.

[2] J. A. Atkinson and P. J. Hawkesford, “Machine learning in plant stress phenotyping,” Trends in Plant Science, vol. 24, no. 10, pp. 915–928, 2019.

[3] Y. Yan et al., “Research progress on electrical signals in higher plants,” Chinese Science Bulletin, vol. 54, 2009.

[4] A. Szechyńska-Hebda et al., “Electrical signaling, photosynthesis and systemic plant responses,” Frontiers in Physiology, vol. 8, 2017.

[5] E. Najdenovska, A. Mijatovic, and R. Smith, “Identifying general stress in commercial tomato plants based on electrophysiological signals,” Applied Sciences, vol. 11, no. 12, 2021.

[6] D. Tran et al., “Long-term electrophysiological recordings in tomato plants for water stress detection,” Scientific Reports, 2019.

[7] A. G. Volkov, “Plant electrical signaling mechanisms,” Plant Signaling & Behavior, vol. 7, 2012.

[8] S. Fromm and S. Lautner, “Electrical signals and their physiological significance in plants,” Plant, Cell & Environment, vol. 30, no. 3, pp. 249–257, 2007.

[9] A. G. Volkov, Plant Electrophysiology: Signaling and Responses. Berlin, Germany: Springer, 2012.

[10] J. Zimmermann et al., “Electrical signaling in plant stress responses,” Plant Signaling & Behavior, 2016.

[11] E. Najdenovska, A. Mijatovic, and R. Smith, “Classification of plant electrophysiology signals for detection of spider mites infestation in tomatoes,” Applied Sciences, vol. 11, no. 4, 2021.

[12] D. Tran, A. G. Volkov, and J. C. Ranatunga, “Electrophysiological assessment of plant status outside a Faraday cage using machine learning,” Scientific Reports, vol. 9, pp. 1–10, 2019.

[13] P. A. Gloor et al., “Machine learning distinguishes plant bioelectric signals,” Scientific Reports, 2025.

[14] R. N. I. Basnayake, G. M. S. C. Gajendrasinghe, and C. Disanayake, “Integrating Artificial Intelligence with Plant Electrophysiology for Early Stress Diagnosis: A Scoping Review,” International Journal of Research and Scientific Innovation, vol. 12, no. 13, pp. 194–201, 2026.

[15] R. C. B. Ramos et al., “Non-invasive plant monitoring using bioelectrical signals,” Biosystems Engineering, 2020.

[16] C. Liu et al., “Cross-individual electrophysiological signal recognition in plants,” Applied Soft Computing, 2025.

[17] J. Zhou et al., “Machine learning assisted implantable plant electrophysiology sensors for early stress detection,” Biosensors and Bioelectronics, vol. 210, 2025.

[18] D. González, M. C. Rodríguez, and J. Pérez, “Detecting nitrogen stress in plants using deep learning and electrophysiological data,” Scientific Reports, vol. 13, 2023.

[19] Z. Liu et al., “Analysis of electrical signals in plant physiological regulation,” Agriculture, vol. 15, no. 21, 2025.

[20] E. Kozlova et al., “Electrome analysis as a tool for plant monitoring,” Plants, vol. 14, no. 10, 2025.

[21] D. Madariaga et al., “A library of electrophysiological responses in plants,” Scientific Reports, vol. 14, 2024.

[22] H. Siddiqui et al., “Classifying plant electrical signals in response to environmental stimuli,” Plant Signaling & Behavior, 2025.

[23] N. Paul et al., “Deep learning for plant stress detection: A comprehensive review,” Computers and Electronics in *Agriculture*, vol. 210, 2025.

[24] S. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[25] X. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: A review,” Sensors, vol. 18, no. 8, 2018.

[26] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[27] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,”

in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[28] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] S. Afif et al., “Design of smart plant electrical signal monitoring systems,” BIO Web of Conferences, 2025

[30] Y. Sun et al., “Plant stress detection via molecular communication,” Plants, vol. 14, 2025.

[31] M. Dolfi et al., “Electrical signaling related to water stress acclimation,” Plant Science, 2021.

[32] K. Sai et al., “Abiotic stress classification through spectral analysis of plant electrical signals,” Biosystems Engineering, 2022.

[33] P. A. Gloor et al., “Plant bioelectric early warning systems for environmental monitoring,” Scientific Reports, 2025.

[34] L. Cao et al., “Transfer learning for plant stress classification,” IEEE Transactions on Neural Networks and Learning Systems, 2022.

[35] M. Hussain et al., “Hybrid algorithms for plant stress classification,” Sustainability, vol. 17, 2025.

[36] J. Qi et al., “Clivia biosensor for soil moisture detection using plant electrical signals,” Biosensors and Bioelectronics, 2024.

# AMMFF: Adaptive Multi-Modal Fusion for Chest Radiograph Diagnosis under Missing Clinical Data

H P G D Dilina Mewan
*Department of Software Engineering*
*NSBM Green University*
*Homagama, Sri Lanka*
pgdmewan@students.nsbm.ac.lk

R. Ranaweera
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
ranaweera.r@nsbm.ac.lk

***Abstract*—Chest radiograph classifiers that also read structured clinical data usually fuse the two sources with fixed weights and are developed on largely complete records, although laboratory results and diagnosis histories are often unavailable when a film is taken. We present AMMFF, a fusion framework with two modules. An Adaptive Fusion Weight (AFW) module computes bounded per-patient modality weights and moves weight toward the image as the observed share of the record falls, and a Missing Modality Resilience (MMR) module estimates absent clinical values from the radiograph and replaces only the missing entries. Seven models are compared on 30,678 frontal MIMIC-CXR radiographs of 8,000 patients, linked to 18 MIMIC-IV features restricted to information available before each study. Under this alignment only 35.8% of records are complete, and values go missing in whole blocks rather than at random. Results are averaged over five training seeds and tested with a patient-clustered paired bootstrap and Holm correction. On records with their natural missingness, AMMFF (0.745 macro ROC-AUC) is slightly less accurate than static fusion (0.751; Holm $p = 0.012$) and concatenation (0.750; $p = 0.042$), and every fusion model exceeds the image-only baseline (0.736). Its advantage is robustness: when a further 70% of feature blocks are withheld, AMMFF loses 0.008 ROC-AUC against 0.019 for static fusion and 0.020 for concatenation (both Holm $p < 0.001$). With every clinical value withheld, the fixed-fusion models fall 0.008 to 0.012 below the image-only model, and AMMFF falls 0.002 below it, a gap whose interval includes zero. Block withholding removes fewer values than random withholding yet degrades every model more, so random-deletion protocols overstate robustness. The image-derived estimates recover age ($r = 0.69$), sex (AUROC 0.90) and obesity (AUROC 0.82) but not serum electrolytes.**

***Keywords*—Multi-Modal Fusion, Chest Radiograph, Missing Clinical Data, Robustness, Clinical Decision Support**

## I. Introduction

Chest radiography is the most frequently performed imaging examination, and networks trained on large public archives now approach specialist performance on several thoracic findings [1], [2]. A radiologist, however, reads a film alongside the patient's age, recent laboratory results and known conditions, and models that add this structured information outperform image-only classifiers [3].

Two assumptions run through most fusion designs. The modalities are combined with weights fixed at training time, although the value of each source differs between patients and findings; and the clinical record is taken to be complete. The second assumption fails in routine care. When the clinical features in our cohort are restricted to information that existed when each radiograph was taken, only 35.8% of studies have all 18 features, laboratory values are missing for 31.5% to 39.5% of studies, and the history block for 42.7%. Values also tend to vanish together, because each group has its own source: either the patient has an earlier discharged admission or not, and either a panel was drawn in the past week or not.

We present AMMFF, an Adaptive Multi-Modal Fusion Framework with two modules. The Adaptive Fusion Weight (AFW) module produces a bounded, per-patient weight over the two modalities, plus a structural term that moves weight toward the image as the missing fraction of the record rises. The Missing Modality Resilience (MMR) module predicts the clinical vector from the image representation and substitutes its predictions only where the record is blank, keeping the missingness mask visible to later layers. The contributions are:

- A bounded fusion gate whose base weight cannot leave $(0.05, 0.95)$ and whose image weight is at least $0.05 + 0.95\rho$ for missing fraction $\rho$, reaching exactly 1 when no clinical value is observed; these are properties of the mixing weight, not a guarantee on accuracy;
- An image-to-clinical reconstruction module that fills only missing entries, with its recoverability measured per feature on held-out records;
- An evaluation protocol with time-aligned clinical features, the measured block structure of missingness, three missingness mechanisms, five training seeds and a pre-specified, multiplicity-adjusted set of primary comparisons; and
- The finding that AMMFF gives up a small but significant amount of accuracy on records as extracted in exchange for less than half the degradation of fixed-weight fusion when clinical blocks are withheld, together with a documented negative result on unbounded gating.

## II. Related Work

### A. Radiograph Classification and Fusion with Clinical Data

Labeled archives such as ChestX-ray14 [2], CheXpert [4] and MIMIC-CXR [5] made chest film reading a multi-label benchmark, and DenseNet classifiers [1], [6] remain the usual image-only reference. Fusion with structured records is commonly grouped into early, intermediate and late designs [7]. Khader et al. [3] reported 0.77 mean ROC-AUC on MIMIC data with a multimodal transformer against 0.70 for images alone. MedFuse [8] fuses chest films with clinical time series through an LSTM and tolerates unpaired samples, and DrFuse [9]

separates shared and modality-specific representations with disease-aware attention. In these designs the weighting is fixed or learned inside attention that assumes the clinical input is present; none makes it an explicit, bounded function of how much of the record was observed.

### *B. Missing-Modality Learning*

A separate line of work handles inputs that are absent altogether. SMIL [10] uses Bayesian meta-learning for severely missing modalities, ShaSpec [11] learns shared and specific features to stand in for an absent modality, M3Care [12] imputes in latent space from similar patients, and HEALNet [13] and Flex-MoE [14] build resilience into hybrid attention and mixture-of-experts routing. Most of these target the loss of an entire modality, and DrFuse and MedFuse expect time-series inputs that our 18-value record does not provide, so a faithful comparison would need either their native inputs or adaptations that change the methods. We therefore compare against baselines that share the backbone, data, loss, curriculum and schedule and differ only in the fusion operator, which isolates the two modules; benchmarking the methods above under the same protocol remains future work.

## III. Materials and Methods

### *A. Cohort, Labels and Image Preprocessing*

The cohort holds 30,678 frontal radiographs of 8,000 patients from MIMIC-CXR-JPG [5], split at the patient level into 21,421 training images (5,600 patients), 4,148 validation images (1,200) and 5,109 test images (1,200). Epochs were selected on a fixed half of the validation patients (600); the other half was reserved for calibration and plays no part here. The targets are the 14 CheXpert findings [4]. Findings labeled uncertain, and findings left blank, were treated as negative (the U-zeros policy) for training and evaluation of every model. Images were stored as 256-pixel JPEGs, converted to grayscale, scaled to [−1024, 1024] and resized to 224 × 224 with TorchXRayVision 1.5.2 [15]. Features come from its CheXpert-trained DenseNet-121 (densenet121-res224-chex), frozen, through ReLU and global average pooling to 1,024 dimensions. MIMIC-trained weights were avoided because they may have seen test patients.

### *B. Time-Aligned Clinical Features*

Each study is linked by subject identifier to MIMIC-IV [16], and its prediction time T is the recorded study date and time. Age is the MIMIC-IV anchor age plus the years between the anchor year and T; sex comes from the patients table, and both are missing when the patient has no row there. Each of seven laboratory tests (creatinine, glucose, potassium, sodium, hematocrit, hemoglobin, white-cell count) takes the most recent result that was drawn within the 7 days before T and whose result was available by T (storage time, or chart time when none is recorded); otherwise it is missing. Nine history flags (hypertension, diabetes, heart failure, COPD, chronic kidney disease, coronary artery disease, atrial fibrillation, cancer, obesity) are set from ICD-9/10 code prefixes of admissions discharged before T, since codes are assigned at discharge; with no earlier discharged admission the whole block is missing. No value recorded after T can enter a feature. A binary mask m records which values are observed, and features are z-scored with training-split statistics.

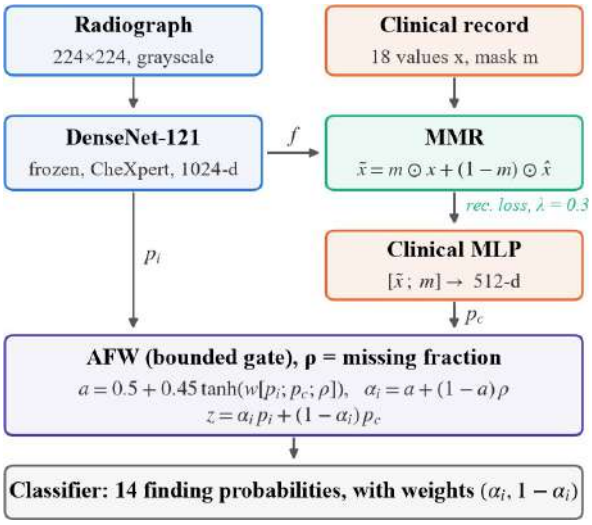



Fig. 1. AMMFF. MMR fills only the missing clinical entries from the image features and passes the original mask on; AFW fuses the two projected representations with a bounded per-patient weight whose lower limit rises with the missing fraction ρ.

### *C. Architecture*

Fig. 1 shows the framework. The clinical branch is a three-layer perceptron over the clinical vector concatenated with its mask, with layer normalization [17] so that the two encodings fused later have comparable scale. MMR is a perceptron that predicts the whole clinical vector $\hat{x}$ from the image descriptor f; the completed vector is

$$\tilde{x} = m \odot x + (1 - m) \odot \hat{x} \quad (1)$$

so a measured value is never overwritten, and the original mask, not an all-ones mask, accompanies $\tilde{x}$ into the clinical branch, keeping the pattern of missingness visible [18]. An auxiliary squared error over observed entries only (weight $\lambda = 0.3$) supervises $\hat{x}$.

AFW projects the image and clinical encodings to 512 dimensions ($p_i$, pc). A gating network w reads both projections and the missing fraction ρ of the record:

$$a = 0.5 + 0.5\beta \tanh(w([p_i ; pc ; \rho])), \quad \beta = 0.9 \quad (2)$$

$$\alpha_i = a + (1 - a)\,\rho \quad (3)$$

$$z = \alpha_i\, p_i + (1 - \alpha_i)\, pc \quad (4)$$

By (2) the base weight a lies in (0.05, 0.95), so while any clinical value is observed neither modality can be silenced. By (3) the image weight is at least $0.05 + 0.95\rho$ and equals 1 at $\rho = 1$, so AMMFF then uses only its image path. Because a itself depends on ρ, (3) does not make $\alpha_i$ monotonic in ρ for every patient; the bound is what is guaranteed. Nor do these properties bound accuracy: at $\alpha_i = 1$ the image features still pass through a projection and head trained jointly with the clinical branch, which are not those of the separate image-only model, so performance relative to that model is tested empirically. A regularizer $0.1 \cdot E[(1 - \rho)(\alpha_i - 0.5)^2]$ keeps the gate near parity on complete records, and a head maps z to 14 logits.

### D. Training and Compared Models

Seven models share the frozen backbone, data and schedule: image-only; clinical-only; concatenation of the two encodings followed by a projection; static fusion, the weighted sum of (4) with $\alpha_i$ fixed at 0.5; AFW-only; MMR-only; and AMMFF. All train for 40 epochs (batch 256, AdamW [19], [20], learning rate $10^{-3}$, weight decay $10^{-5}$, cosine schedule) on binary cross-entropy with per-finding positive weights clipped to [1, 50], keeping the epoch with the best validation macro ROC-AUC. Every model with clinical input receives the same missingness curriculum: per batch a rate is drawn uniformly from 0 to 60% and each observed value is withheld with that probability. Each model is trained with seeds 0 to 4, which set initialization, batch order and curriculum draws.

### E. Missingness Mechanisms

At test time additional values are withheld on top of those already missing, at rates $r \in \{10, 30, 50, 70\}$% of the observed values, and a final condition withholds every value. A withheld entry has its value and its mask set to zero. Under MCAR [21], [22] each observed entry is withheld independently with probability r. Under block withholding, for each record n and feature block b (demographics, laboratory, history) draw $h_{n\beta} \sim \text{Bern}(r)$, $c_{n\beta} \sim \text{Bern}(\kappa_\beta)$ and $u_{nj} \sim \text{Bern}(0.5)$; an observed entry j of block b is withheld when

$$h_{n\beta}\,[c_{n\beta} + (1 - c_{n\beta})\,u_{nj}] = 1 \qquad (5)$$

where $\kappa_\beta$ is the block cohesion measured in the training data, the share of records missing any value of the block that miss all of it in the extract. Under MNAR, values near the population mean are withheld more often, imitating tests recorded mainly when something is suspected [23]; on standardized values $x_{nj}$,

$$\pi_{nj} = \min\{1,\, s \cdot e^{(-|x_{nj}|)}\}, \quad s = r \sum m_{nj} \,/\, \sum m_{nj}\, e^{(-|x_{nj}|)} \qquad (6)$$

with both sums over all observed test entries, so the expected share withheld equals r before clipping. Each condition is repeated over five draws with fixed generator seeds (10,000d + 100r + 1,000k for draw d and mechanism k), and every model receives identical inputs. Two checks guard the sweep: the image-only score must not change with r, and every model's score at r = 0 must equal its test score.

### F. Statistical Analysis

The test set holds about four images per patient, so all intervals are 95% percentile intervals from 2,000 bootstrap resamples of patients, each drawn patient contributing all of their images [24], [25]. Comparisons are paired, with every model scored on the same resample, the resampling analogue of the correlated-AUC test [26] for a macro-averaged multi-label metric. Each statistic is the macro ROC-AUC averaged over the five seeds, and the seed standard deviation is reported separately. A two-sided p-value is twice the share of resampled differences whose sign is opposite to the estimate; $p < 0.001$ means none was. Four primary comparisons were fixed before the bootstrap was run: AMMFF against static fusion and against concatenation on records as extracted (P1, P2), and the difference in ROC-AUC lost at 70% block withholding for the same pairs (P3, P4). Their p-values are Holm-adjusted [27]. All other comparisons, including every comparison with the image-only model, are exploratory and carry unadjusted p-values, as recommended for reporting prediction models [28].

## IV. Results

### A. Structure of the Missing Data

Missingness in the time-aligned extract is structured by source. When any history flag is missing, all nine are missing (cohesion 100%); the same holds for demographics, and the laboratory panel is lost whole in 76.8% of the records missing any of it. Five record shapes cover 90.8% of the data: all 18 features (35.8%), demographics and laboratory values (24.3%), demographics and history (15.4%), demographics only (9.8%) and nothing (5.5%). This is consistent with warnings that clinical missingness is rarely random [29] and that the mechanism is seldom reported in machine-learning studies [30]. The as-extracted test records, with their natural missingness, are therefore the deployment condition, and on average 65.8% of their values are observed.

### B. Performance on Records as Extracted

TABLE I
TEST MACRO ROC-AUC AND PR-AUC ON RECORDS AS EXTRACTED (MEAN ± SD OVER FIVE SEEDS; 5,109 IMAGES), AND ROC-AUC WITH EVERY CLINICAL VALUE WITHHELD. ROWS ARE NOT RANKED.

| Model | ROC-AUC | PR-AUC | All withheld |
|---|---|---|---|
| Image-only | 0.7358 ± 0.0008 | 0.2918 | 0.7358 |
| Clinical-only | 0.6787 ± 0.0021 | 0.2280 | 0.5000 |
| Concatenation | 0.7495 ± 0.0025 | 0.2962 | 0.7240 |
| Static fusion | 0.7507 ± 0.0031 | 0.2974 | 0.7278 |
| MMR-only | 0.7468 ± 0.0020 | 0.2939 | 0.7256 |
| AFW-only | 0.7477 ± 0.0010 | 0.2951 | 0.7346 |
| AMMFF | 0.7449 ± 0.0008 | 0.2929 | 0.7339 |

Every fusion model scores above the image-only reference on records as extracted (Table I); for AMMFF the exploratory difference is +0.0091 (95% CI 0.0013 to 0.0168, p = 0.020). AMMFF is, however, the least accurate of the five fusion designs, and the two pre-specified accuracy comparisons go against it (Table II): it scores 0.0058 below static fusion and 0.0047 below concatenation, both significant after Holm adjustment. Against AFW-only the gap of 0.0029 is not significant (p = 0.092). The bounded gate therefore has a cost on the records the model will actually meet, and we do not claim an accuracy advantage.

P TABLE II
RE-SPECIFIED PRIMARY COMPARISONS: DIFFERENCES IN SEED-AVERAGED MACRO ROC-AUC, PATIENT-CLUSTERED PAIRED BOOTSTRAP, HOLM-ADJUSTED P.

| ID | Comparison | Est. | 95% CI | Holm p |
|---|---|---|---|---|
| P1 | AMMFF − static, as extracted | −0.0058 | −0.0098, −0.0016 | 0.012 |
| P2 | AMMFF − concat., as extracted | −0.0047 | −0.0090, −0.0001 | 0.042 |
| P3 | Loss static − loss AMMFF, block 70% | +0.0107 | 0.0079, 0.0133 | < 0.001 |
| P4 | Loss concat. − loss AMMFF, block 70% | +0.0118 | 0.0086, 0.0146 | < 0.001 |

### C. Robustness to Withheld Clinical Data

Three results follow from Fig. 2 and Table III. First, block withholding is harder than MCAR while removing less: at 70% it leaves 21.9% of values observed against 19.9% under MCAR, yet every model loses more under it, concatenation 0.0197 against 0.0129. Removing a block removes a whole kind of evidence, whereas scattered removal leaves part of every block, so MCAR robustness figures are optimistic. MNAR is the mildest mechanism, plausibly because it keeps the abnormal values that carry more of the signal.

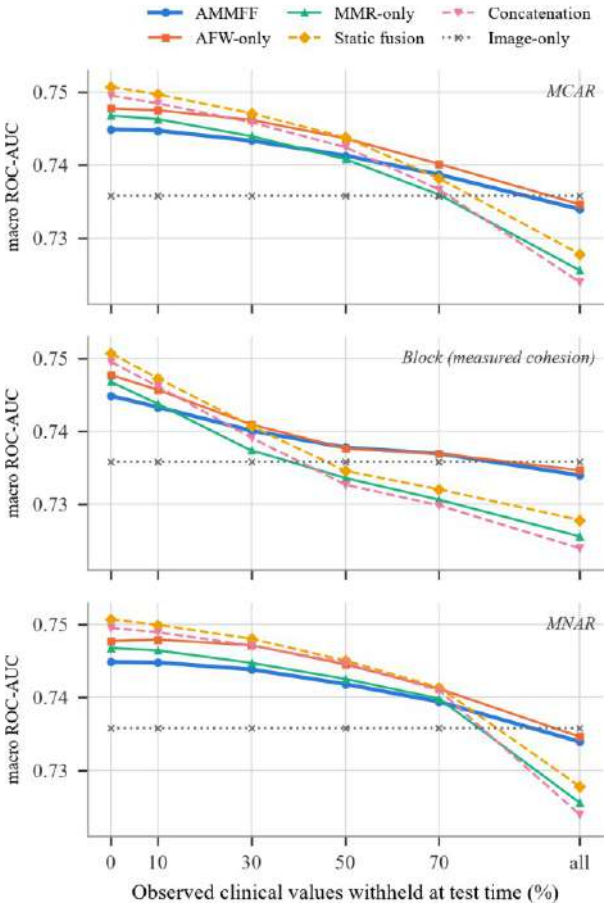



Fig. 2. Test macro ROC-AUC as observed clinical values are withheld under three mechanisms (mean of five seeds and five draws; "all" withholds every value). Seed SDs are at most 0.004 and are omitted for legibility; intervals for the key contrasts are in Table II and the text.

Second, AMMFF degrades least under all three mechanisms. At 70% block withholding it loses 0.0079 against 0.0187 for static fusion and 0.0197 for concatenation, differences of 0.0107 and 0.0118 that are the primary comparisons P3 and P4 (Table II). The two modules are complementary: AMMFF loses less than AFW-only (difference 0.0029, 95% CI 0.0004 to 0.0054, p = 0.026) and than MMR-only (0.0082, p < 0.001). Under block withholding the fixed-fusion curves fall to the level of AMMFF by about 30% and below the image-only line by 50%.

TABLE III
ROC-AUC LOST BETWEEN RECORDS AS EXTRACTED AND 70% WITHHELD, AND DIFFERENCE FROM THE IMAGE-ONLY MODEL (Δ) AT 70% BLOCK AND WITH ALL VALUES WITHHELD (MEAN OF FIVE SEEDS).

| Model | MCAR | Block | MNAR | Δ block | Δ all |
|---|---|---|---|---|---|
| Concatenation | 0.0129 | 0.0197 | 0.0086 | −0.0060 | −0.0118 |
| Static fusion | 0.0125 | 0.0187 | 0.0093 | −0.0038 | −0.0080 |
| MMR-only | 0.0108 | 0.0162 | 0.0069 | −0.0052 | −0.0102 |
| AFW-only | 0.0076 | 0.0108 | 0.0065 | +0.0011 | −0.0012 |
| AMMFF | 0.0062 | 0.0079 | 0.0055 | +0.0011 | −0.0019 |
| Clinical-only | 0.0593 | 0.1006 | 0.0394 | — | — |

Third, the image-only comparison is mixed. At 70% block withholding concatenation falls 0.0060 below the image-only model (95% CI −0.0118 to −0.0004, p = 0.037) and AMMFF sits 0.0011 above it (CI −0.0024 to 0.0046). With every value withheld, concatenation and static fusion fall 0.0118 and 0.0080 below it (both p < 0.001); AMMFF falls 0.0019 below (CI −0.0042 to 0.0005, p = 0.125), and below the seed-matched image-only model in four of five seeds. Its worst case is close to, but not demonstrably at or above, the image-only model, so we report this as an empirical observation in the tested conditions, not as a property of the architecture.

### D. Sensitivity to Extraction, Curriculum and Seed

A preliminary analysis of this cohort used each patient's first recorded laboratory value, diagnosis codes from any admission and the anchor age, and coded patients without a demographic row as female. Rebuilding the features as in Section III-B changed the fusion scores by −0.0010 to −0.0080 (AMMFF −0.0080), raised the clinical-only model by 0.0231, and turned an apparent tie between AMMFF and fixed fusion into the deficit of Table II. That analysis also gave the curriculum only to the AFW and MMR models; giving it to concatenation and static fusion as well, on the same extract, cut their loss at 70% block withholding from 0.0269 to 0.0201 and 0.0209, so part of the robustness gap seen there came from training rather than architecture. Finally, seed standard deviations reached 0.005, the size of the single-seed differences between fusion designs. Every result above uses the time-aligned extract, the shared curriculum and five seeds.

### E. Fusion Weights and Reconstruction

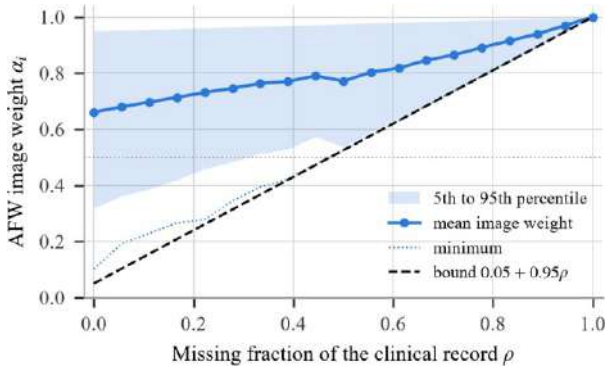



Fig. 3. Per-patient AFW image weight of AMMFF against the missing fraction ρ, pooled over five seeds, records as extracted, MCAR and block withholding at every rate, and the all-withheld condition (255,450 image-condition pairs). No weight falls below the bound of (3).

On records as extracted the mean AMMFF image weight is 0.751 (SD across patients 0.111); it reaches 0.906 at 70% block

withholding and exactly 1 for every image when all values are withheld. Fig. 3 shows the weight per patient. The bound 0.05 + 0.95ρ holds everywhere and binds as ρ grows, and the base weight a spans the whole interval (0.050 to 0.950), so the limit in (2) is active rather than decorative. The mean weight is not monotonic in ρ: it is 0.791 at ρ = 8/18 and 0.772 at ρ = 9/18, the shape in which the history block is absent, which is why we claim only the bound.

MMR estimates were scored on test records where the true value is observed, with Pearson's r for continuous features and AUROC for binary ones, since a correlation with a 0/1 variable depends on its prevalence (Table IV). Signal that a film can carry is recovered: age r = 0.69, hemoglobin 0.50 and hematocrit 0.45; sex AUROC 0.90, obesity 0.82 and heart failure 0.74. Blood chemistry is not: sodium r = 0.13 and potassium 0.07. The recoverability grade shown in the platform is the per-feature Pearson correlation between estimate and measured value on held-out test records, clipped to [0, 1]. It describes how well the radiograph predicts a feature in general and is not a per-prediction confidence.

TABLE IV
MMR RECONSTRUCTION OF EACH CLINICAL FEATURE FROM THE RADIOGRAPH (AMMFF, OBSERVED TEST VALUES, MEAN OF FIVE SEEDS; SEED SD ≤ 0.021): PEARSON R FOR CONTINUOUS, AUROC FOR BINARY FEATURES.

| Feature | Metric | Value | Feature | AUROC |
|---|---|---|---|---|
| Age | r | 0.69 | Hypertension | 0.69 |
| Creatinine | r | 0.21 | Diabetes | 0.65 |
| Glucose | r | 0.15 | Heart failure | 0.74 |
| Potassium | r | 0.07 | COPD | 0.68 |
| Sodium | r | 0.13 | Chronic kidney dis. | 0.65 |
| Hematocrit | r | 0.45 | Coronary artery dis. | 0.69 |
| Hemoglobin | r | 0.50 | Atrial fibrillation | 0.68 |
| White-cell count | r | 0.19 | Cancer | 0.62 |
| Sex | AUROC | 0.90 | Obesity | 0.82 |

### *F. A Negative Result: Gate Collapse*

Our first adaptive gate, an unbounded softmax over gated-concatenation logits, saturated during development: its image weight reached 1.000 with zero variance and AFW-only fell to 0.646, the clinical-only level. Once one branch is silenced its gradient vanishes, so the collapse reinforces itself. Bounding the base weight as in (2) removed it; we report this because the failure is easy to reproduce and seldom documented.

## V. DISCUSSION

On the records a deployed model would meet, fixed fusion is the more accurate choice by about 0.005 ROC-AUC. When further blocks of the record go missing, through an unfilled history, a panel not yet resulted or a transfer from another site, fixed fusion loses more than twice as much at 70% block withholding, and with no clinical data it falls clearly below the image-only model, while AMMFF stays close to it. Which property matters more depends on how incomplete the records at a site are likely to become, and the trade-off can only be seen when accuracy and degradation are reported together, with intervals. A further lesson concerns evaluation: block withholding is harder than MCAR while removing less, so robustness comparisons made under MCAR alone can overstate robustness and need not preserve their ranking.

Limitations. The backbone is frozen, which caps macro ROC-AUC near 0.75 for every design. Data come from one institution, the 7-day laboratory window is a design choice, uncertain labels are treated as negative, and the additional withholding is simulated on top of real missingness; the MNAR mechanism encodes an assumed rather than observed relation between a value and its recording [22]. We did not benchmark DrFuse, ShaSpec, HEALNet, M3Care or Flex-MoE, so our comparison establishes the effect of the two modules against fixed fusion, not superiority over current missing-modality methods. A subgroup fairness audit remains to be done. The decision-support platform that serves the model, showing fusion weights, MMR estimates with their recoverability and Grad-CAM maps [31], is a research prototype that has not been evaluated prospectively or in clinical use.

## VI. CONCLUSION

AMMFF combines a bounded fusion gate that shifts toward the image as the record empties with an image-to-clinical module that fills only missing entries. With clinical features restricted to what was known at the time of each radiograph, five seeds and multiplicity-adjusted tests, it is slightly less accurate than fixed fusion on records as extracted but loses less than half as much when clinical blocks are withheld, and it does not fall clearly below the image-only model when the record is empty, as fixed fusion does. Missingness in the data is block-structured, and testing under that structure is harder than random deletion. Future work will fine-tune the backbone, benchmark recent missing-modality methods under the same protocol, and validate on prospective, multi-site data.

## REFERENCES


[1] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1711.05225

[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE CVPR, 2017, pp. 2097–2106, doi: 10.1109/CVPR.2017.369

[3] F. Khader et al., "Multimodal deep learning for integrating chest radiographs and clinical parameters: A case for transformers," Radiology, vol. 309, no. 1, 2023, doi: 10.1148/radiol.230806

[4] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI, 2019, pp. 590–597, doi: 10.1609/aaai.v33i01.3301590

[5] A. E. W. Johnson et al., "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019. [Online]. Available: https://arxiv.org/abs/1901.07042

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243

[7] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," Sci. Rep., vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-22514-4

[8] N. Hayat, K. J. Geras, and F. E. Shamout, "MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images," in Proc. MLHC, PMLR, 2022, pp. 479–503. [Online]. Available: https://arxiv.org/abs/2207.07027

[9] W. Yao, K. Yin, W. K. Cheung, J. Liu, and J. Qin, "DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing

modality and modal inconsistency," in Proc. AAAI, 2024, pp. 16416–16424, doi: 10.1609/aaai.v38i15.29578

[10] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in Proc. AAAI, 2021, pp. 2302–2310, doi: 10.1609/aaai.v35i3.16330

[11] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in Proc. IEEE/CVF CVPR, 2023, pp. 15878–15887, doi: 10.1109/CVPR52729.2023.01524

[12] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, and J. Zhao, "M3Care: Learning with missing modalities in multimodal healthcare data," in Proc. ACM SIGKDD, 2022, pp. 2418–2428, doi: 10.1145/3534678.3539388

[13] K. Hemker, N. Simidjievski, and M. Jamnik, "HEALNet: Multimodal fusion for heterogeneous biomedical data," in Proc. NeurIPS, 2024. [Online]. Available: https://arxiv.org/abs/2311.09115

[14] S. Yun et al., "Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts," in Proc. NeurIPS, 2024. [Online]. Available: https://arxiv.org/abs/2410.08245

[15] J. P. Cohen et al., "TorchXRayVision: A library of chest X-ray datasets and models," in Proc. MIDL, PMLR, 2022, pp. 231–249. [Online]. Available: https://arxiv.org/abs/2111.00595

[16] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci. Data, vol. 10, no. 1, 2023, doi: 10.1038/s41597-022-01899-x

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016. [Online]. Available: https://arxiv.org/abs/1607.06450

[18] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," Sci. Rep., vol. 8, 2018, doi: 10.1038/s41598-018-24271-9

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[21] D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581

[22] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed. Hoboken, NJ, USA: Wiley, 2019, doi: 10.1002/9781119482260

[23] D. Agniel, I. S. Kohane, and G. M. Weber, "Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study," BMJ, vol. 361, 2018, doi: 10.1136/bmj.k1479

[24] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall/CRC, 1993, doi: 10.1201/9780429246593

[25] C. A. Field and A. H. Welsh, "Bootstrapping clustered data," J. Roy. Statist. Soc. B, vol. 69, no. 3, pp. 369–390, 2007, doi: 10.1111/j.1467-9868.2007.00593.x

[26] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595

[27] S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Statist., vol. 6, no. 2, pp. 65–70, 1979

[28] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, 2024, doi: 10.1136/bmj-2023-078378

[29] J. A. C. Sterne et al., "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," BMJ, vol. 338, 2009, doi: 10.1136/bmj.b2393

[30] S. W. J. Nijman et al., "Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review," J. Clin. Epidemiol., vol. 142, pp. 218–229, 2022, doi: 10.1016/j.jclinepi.2021.11.023

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74

# An ICU Capacity-Aware Emergency Transfer Decision Support System: Multi-Objective Routing and Ambulance Dispatch for the Colombo Hospital Network

Gunawardane Oneli Wickramaratne*
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University,*
*Sri Lanka*
*oneliwickramaratne@gmail.com

Hirushi Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University,*
*Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Emergency inter-hospital ICU transfers in urban settings suffer from constrained resources, as do many other settings. In resource-constrained urban settings, most inter-hospital ICU transport decisions are based on distance or the number of available beds, ignoring real risks posed by different traffic conditions and the potential that a hospital has multiple ICUs. This article presents an integrated decision support system for the Colombo emergency network containing nine hospitals. Rule-based urgency triage is combined with an artificial intelligence (AI) and machine learning (ML) traffic/congestion model, an urgency-weighted multi-objective router over an OpenStreetMap graph (63,951 nodes, 126,278 edges), and a graph-based ambulance dispatch heuristic. Five quantitative assessments were performed against real code by the authors. An urgency classification model was built using Gradient Boosting and achieved an accuracy of 85.0% and a critical-class recall of 0.79; a confusion matrix shows none of the missed critical cases were under-triaged past the adjacent "high" class. Comparison of shortest-time and risk-aware routing showed a road-risk reduction of 13.3-17.7% with a cost of +0.000 to +0.125 minutes across urgency classes, against a fully defined, auditable per-edge risk formula. A traffic prediction model, trained on a fully synthetic, distribution-calibrated 21,504-observation dataset spanning 14 simulated days, achieved $R^2$ of 0.972–0.978 (congestion) and 0.990–0.991 (duration) under both random and temporal holdout, with MAE, RMSE, and MAPE reported alongside $R^2$; validation against real collected traffic data is identified as future work. A dispatch heuristic was compared against a naïve nearest-distance heuristic in 10 cross-hospital transfer scenarios. A capacity simulation study was conducted in four network stress scenarios. The system is deployed against a real PostgreSQL/PostGIS database in Docker, and the evaluation code is entirely available to the research community.**

***Keywords—Ambulance dispatch; emergency medical services; ICU capacity management; multi-objective routing; urgency triage***

## I. Introduction

The Human Emergency Inter-Hospital Transfer System (HEIHTS) is a web-based GIS application designed to support Emergency Law Enforcement Operators in identifying the closest ICU location to a critical patient. This application helps determine the fastest (dynamic), but safest, route to travel. To the best of our knowledge, this application represents an innovative approach by incorporating a multi-criteria decision aid, identifying GIS routing speed, and supporting ambulance dispatch. The incorporated ICU locations are from 4 major hospitals within and around the Greater Colombo Area.

This gap has been recognised in the Sri Lankan context specifically: prior work has proposed mobile applications that forecast ICU bed demand using artificial intelligence, and broader regional studies have catalogued critical-care bed capacity across Asian health systems [1] . Separately, secure mobile/web platforms for real-time emergency health-data exchange have been proposed to support exactly this kind of inter-facility coordination [2] , however, none of these directly couple urgency-aware routing, multi-ward ICU capacity, and ambulance dispatch into a single evaluated decision-support pipeline for a specific hospital network.

This paper makes the following contributions:

- A hybrid architecture that deliberately separates deterministic, auditable rules (urgency scoring, ICU capability filtering) from learned components (traffic congestion prediction) and graph-based multi-objective optimisation (route cost), rather than applying machine learning end-to-end.
- A realistic multi-ward ICU capacity model in which a single hospital can concurrently operate several distinct ICU specialties (e.g. Trauma, Cardiac, General), with bed availability tracked per ward rather than as a single hospital-level aggregate.
- An urgency-weighted multi-objective router operating on a real Colombo OpenStreetMap road graph, with a fully specified, auditable per-edge risk formula whose risk/time/distance weighting is validated quantitatively rather than only by design.
- A graph-aware ambulance dispatch heuristic, evaluated against a naive nearest-distance baseline.
- Five reproducible, code-level quantitative evaluations, including an honest random-vs-temporal-split validation of the traffic model and a full confusion matrix for urgency triage , plus a discussion of a concrete data-integrity issue that a SQLite development database silently masked and a production PostgreSQL

deployment correctly surfaced.

## II. Related work

### A. Ambulance Routing and Dispatch Optimisation

Machine learning has been combined with metaheuristic search for ambulance routing under uncertainty, for example pairing learned demand/traffic estimates with a genetic algorithm to optimise routing decisions[4] L. Li, P.G.R.N.I. Pussella, Is Colombo city, Sri Lanka secured for urban green space standards? Applied Ecology and Environmental Research 15 (2017) 1789-1799. doi: 10.15666/aeer/1503_17891799.. Multi-objective formulations have also been applied at the facility-location level, optimising ambulance base placement across a city under competing objectives [3] , while other work optimises routing specifically as a function of live emergency medical service (EMS) availability [4] or uses deep neural networks within a mobile dispatch system [5]. Closer to this paper's concerns, dynamic travel-time calculation has been shown to materially affect prehospital stroke-care transport simulations [6], a hybrid optimisation-and-simulation framework has been proposed for both static and dynamic ambulance allocation [7] , and multi-agent reinforcement learning with hierarchical coordination has been explored for emergency-responder stationing at scale [8] The present system takes a lighter-weight, explainable approach: a deterministic weighted-scoring heuristic over graph-computed pickup routes, chosen so that every dispatch decision remains auditable to a human dispatcher.

### B. Shortest-Path Algorithms and Traffic Modelling

Comparative studies of classical shortest-path algorithms consistently find that A* with an admissible heuristic matches Dijkstra's algorithm's optimality while typically exploring fewer nodes [9] a result echoed in graph-based urban road-network routing studies applying Dijkstra's algorithm directly to city-scale networks [10] . This system uses NetworkX's A* implementation with a zero heuristic, mathematically equivalent to Dijkstra's algorithm, chosen so a real distance heuristic can later be introduced as a performance optimisation without changing surrounding code. On the traffic side, comparative surveys of traffic-modelling techniques distinguish macroscopic, mesoscopic, and microscopic approaches for predicting and simulating traffic behaviour, and dedicated urban-mobility optimisation models have proposed mathematical formulations for predictive traffic analysis at the city scale [11] . This paper's traffic-congestion model is a tabular regression approach (Random Forest / Gradient Boosting) trained on a fully synthetic, distribution-calibrated observation set, consistent with the tabular, feature-engineered style favoured in this literature over end-to-end deep traffic simulation given the available data volume.

### C. Multi-Objective Route and Fleet Optimisation

Beyond single-objective shortest-path search, AI-driven multi-objective dynamic route optimisation has been proposed for smart urban mobility and logistics generally [12] and stochastic multi-objective, multi-trip routing with time windows has been studied for autonomous mobile robot fleets [13] This paper's router follows the same multi-objective spirit but restricts itself to three interpretable, urgency-weighted terms: travel time, road-structural risk, and distance, combined linearly rather than via a learned or Pareto-front formulation, prioritising explainability for a safety-critical clinical context.

### D. ICU Capacity Prediction and Bed Allocation

ICU bed-occupancy forecasting received particular attention during the COVID-19 pandemic: multistate models have been used to predict hospital and ICU occupancy from patient-level trajectories [14] and tri-model prediction approaches have been proposed specifically for COVID-19 ICU bed occupancy. At the allocation-policy level, two-stage optimisation formulations have addressed ICU bed allocation under multiple simultaneous sources of uncertainty [15] complementing regional surveys of Asian critical-care bed capacity [16] . Dispatch-side, a simulation-based EMS dispatch recommendation system has been proposed that explicitly incorporates destination bed availability into the dispatch decision [17] , and weighted multi-criteria decision modelling combined with geospatial proximity-capacity balancing has been proposed for hospital resource management and patient allocation more broadly [18] . This paper's contribution relative to this body of work is the explicit per-ward, multi-type ICU capacity model described below, rather than a single-type hospital-level bed count.

### E. Care Coordination, Fairness, and Decision Support

At the coordination layer, fairness and efficiency trade-offs in large, heterogeneous EMS regions have been studied directly [19] , walk-in multi-stage patient-flow scheduling has been modelled with answer-set-programming formulations evaluated via discrete-event simulation [20] , and interdisciplinary transfer-huddle interventions have been shown to reduce prolonged wait times during inter-ICU transfer in practice [6] . Clinical decision support more broadly has been explored through business-intelligence tooling for evidence-based nursing decisions [21] , and the wider healthcare-logistics research programme reflected in recent operational-research-in-health-services proceedings [22] situates this paper's system firmly within an active, applied research area rather than a purely theoretical one.

## III. System Architecture

Fig. 1 summarises the system's hybrid architecture, which deliberately keeps three computational styles separate: deterministic rules for clinical/capability filtering, a learned model for traffic/congestion estimation, and graph-based multi-objective optimisation for route and dispatch ranking.

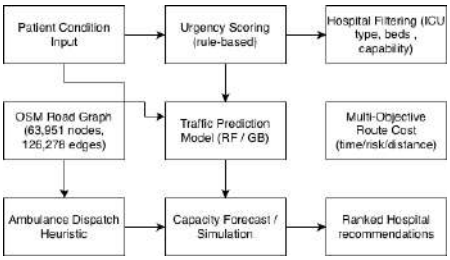



Fig. 1. Hybrid rules + ML + graph-routing architecture.

### A. Multi-Ward ICU Capacity Model

Each of the nine seeded Colombo hospitals is modelled as owning one or more independent ICU wards, each with its own specialty type, bed count, and occupancy, rather than a single

hospital-level ICU-type label. For example, the tertiary referral hospital in the network operates Trauma, General, and Cardiac ICU wards concurrently, so a hospital with a fully occupied Cardiac ICU remains a valid recommendation target for a trauma case if its Trauma ICU has an open bed. Bed availability reported to the routing and recommendation services is always ward-type-specific, not a hospital-wide aggregate.

### *B. Road Network Representation and Risk Model*

Colombo's drivable road network is represented as a graph extracted from OpenStreetMap: 63,951 nodes (intersections) and 126,278 edges (road segments). Each edge carries a base risk value determined by its OpenStreetMap highway classification (Table I), reflecting the general safety/complexity profile of that road type, plus a flat +1.5 penalty when the edge terminates at a signalized ("traffic_signals") intersection node. This is an engineered, auditable proxy for road risk- chosen for the same explainability reasons as the urgency score - not a measurement of, or substitute for, validated transport-safety outcomes; absent historical incident or delay labels, a transparent engineered risk score is more defensible than an unlabelled learned one.

TABLE I
PER-EDGE ROAD RISK BY OSM HIGHWAY CLASS

| Highway class | Risk | Highway class | Risk |
|---|---|---|---|
| motorway | 0.35 | tertiary | 1.00 |
| trunk | 0.45 | unclassified | 1.10 |
| motorway_link | 0.55 | tertiary_link | 1.15 |
| trunk_link | 0.65 | residential | 1.25 |
| primary | 0.75 | living_street | 1.40 |
| primary_link | 0.90 | service | 1.50 |
| secondary | 0.90 | secondary_link | 1.05 |

At routing time this per-edge risk is converted into an additive path cost of risk × (length in metres / 100) × k(urgency), where k(urgency) ∈ {2.0, 5.0, 9.0} for critical/high/moderate, so critical transfers are weighted to tolerate risk far more than moderate ones.

## IV. Methodology

### *A. Urgency Classification*

Patient urgency is scored by a deterministic rule-based service using condition type, oxygen saturation band, blood pressure band, consciousness level, and ventilator requirement, producing a numeric score in [0,1] mapped to critical/high/moderate classes. This rule-based score is the production baseline; a comparison against learned classifiers trained on the same feature space is reported below, used to validate the rule design rather than to replace it. Explainability in triage was prioritised over marginal accuracy gains from a black-box model.

The 1,000 synthetic training scenarios used in that comparison are labeled by thresholding this same deterministic rule-based score (≥0.75 critical, ≥0.50 high), with independently randomized input features and ±0.08 label noise. Consequently, this classifier comparison demonstrates that the rule's decision boundary is learnable and internally consistent across model families — a sanity check on the rule's structure — rather than an independent validation of the rule's clinical correctness, which would require labels from an external source such as clinician-adjudicated cases.

### *B. Multi-Objective Graph Routing*

For a candidate hospital, two routes are computed over the road graph using A* search with a zero heuristic: shortest-time (edge weight = travel time only), and urgency-aware (edge weight = travel time plus the risk penalty defined above). Each candidate hospital's final ranking score combines normalised time, risk, and distance using urgency-specific weights (critical: 0.75/0.20/0.05; moderate: 0.45/0.40/0.15), plus a small bed-availability bonus.

### *C. Ambulance Dispatch Heuristic*

On transfer acceptance, the eight nearest available ambulances (by straight-line distance) are shortlisted, then re-scored using real road-network pickup ETA and risk, a base-coverage penalty (avoiding stripping a hospital of its own on-site coverage), and a same-base preference. This two-stage design keeps the expensive graph-routing computation bounded to a small candidate set while still correcting for cases where straight-line "nearest" disagrees with real-road "nearest".

### *D. Capacity Forecasting and Simulation*

A deterministic forecast projects each hospital's bed occupancy forward at 1/3/6/12-hour horizons from current occupancy, inbound transfers, and recent discharge/cleaning rates, thresholded into stable/elevated/high/critical pressure levels. A separate, read-only scenario simulator applies the same scoring to four built-in demand scenarios (baseline, evening surge, mass casualty, respiratory wave) without mutating live data, intended as an early-warning tool consistent with prior EMS dispatch simulation work that incorporates destination bed availability into operational decisions [17] .

## V. Evaluation And Results

Five evaluations were run directly against the production service classes, not a separate offline simulation, each reproducible via a dedicated script.

### *A. Urgency Classification Model Comparison*

Trained on 1,000 synthetic transfer scenarios (80/20 stratified split, 200-sample test set). Table II reports accuracy and critical-class recall; Fig. 2 visualises the comparison.

TABLE II
URGENCY MODEL COMPARISON

| Model | Accuracy | Critical Recall |
|---|---|---|
| Logistic Regression | 84.5% | 0.74 |
| Random Forest | 84.0% | 0.74 |
| Gradient Boosting | 85.0% | 0.79 |

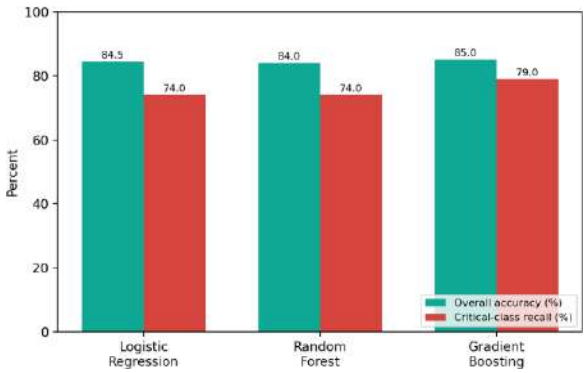



Fig. 2. Urgency model comparison: accuracy vs. critical-class recall.

Gradient Boosting achieved both the highest overall accuracy and the best recall on the critical class (0.79 vs. 0.74), the more clinically relevant metric since a missed critical case is costlier than an over-triaged moderate one. Table III gives the full confusion matrix for the selected model.

TABLE III
GRADIENT BOOSTING CONFUSION MATRIX (N=200 TEST SET)

| True \ Predicted | critical | high | moderate |
|---|---|---|---|
| critical (n=43) | 34 | 9 | 0 |
| high (n=86) | 8 | 76 | 2 |
| moderate (n=71) | 0 | 11 | 60 |

Critical-class recall of 0.79 corresponds to a 20.9% false-negative rate on critical patients; notably, none of these nine missed cases were under-triaged all the way to moderate - all nine were classified as high, the adjacent class. Gradient Boosting was selected specifically to minimize this critical-class miss rate rather than to maximize overall accuracy, since a missed critical case is clinically costlier than an over-triaged moderate one. In production, this classifier is not itself in the decision path - the deterministic rule-based score is - so this comparison bounds how well a learned model could approximate that rule, not the system's actual triage error rate.

### *B. Routing Strategy Evaluation*

24 route comparisons were run across 8 real hospital pairs at all three urgency levels. Table IV and Fig. 3 report the average extra travel time and road-risk reduction relative to the shortest-time baseline, using the auditable per-edge risk formula defined earlier.

TABLE IV
ROUTING TRADEOFF BY URGENCY

| Urgency | Extra Time | Risk Reduction |
|---|---|---|
| Critical | +0.000 min | 13.3% |
| High | +0.037 min | 15.7% |
| Moderate | +0.125 min | 17.7% |

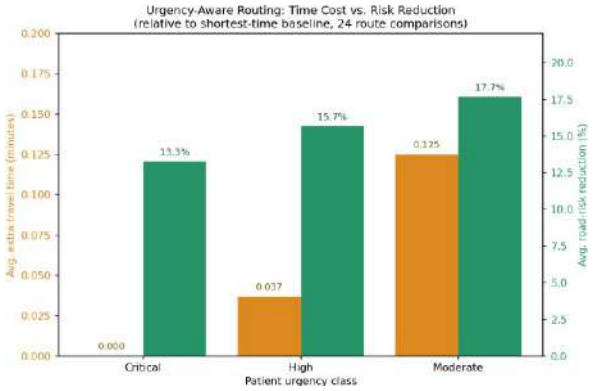



Fig. 3. Urgency-aware routing time cost vs. road-risk reduction.

Critical transfers obtain a lower-risk route at effectively zero time cost, while moderate transfers accept more time for a larger risk reduction, confirming the urgency-weighted design behaves as intended in practice, not only by construction. These reported risk reductions describe reduction in the engineered risk score defined earlier, not a measured reduction in real-world transport risk; a weighting-sensitivity analysis against the live road graph is identified as follow-up work in the discussion below.

### *C. Traffic Prediction Model Comparison*

Trained on 21,504 fully synthetic observations spanning 14 simulated days, generated to match the distributional shape of Google Routes API responses (duration, congestion ratio, and route-structure features); no live Google Routes API data was used in model training for this evaluation. Table V reports $R^2$, MAE, RMSE, and MAPE under both a random 80/20 split (the split used elsewhere in this paper) and a temporal split (last 3 of 14 simulated days held out).

Random Forest was selected for congestion ratio and Gradient Boosting for duration; both are deployed in the production backend. The near-identical fit across random and temporal splits indicates the high $R^2$ is not an artifact of data leakage from random splitting, but a consequence of training on fully synthetic, formula-generated data with no real-world sensor or GPS noise. Validation of this model against real collected Google Routes observations - for which the data-collection tooling exists but has not yet been run at scale - is identified as priority future work.

TABLE V
TRAFFIC MODEL VALIDATION: RANDOM VS. TEMPORAL SPLIT

| Target / Split | $R^2$ | MAE | RMSE | MAPE |
|---|---|---|---|---|
| Congestion — random | 0.978 | 0.032 | 0.043 | 2.61% |
| Congestion — temporal | 0.972 | 0.034 | 0.046 | 2.75% |
| Duration — random | 0.991 | 19.38s | 26.29s | 3.07% |
| Duration — temporal | 0.990 | 19.44s | 26.38s | 3.14% |

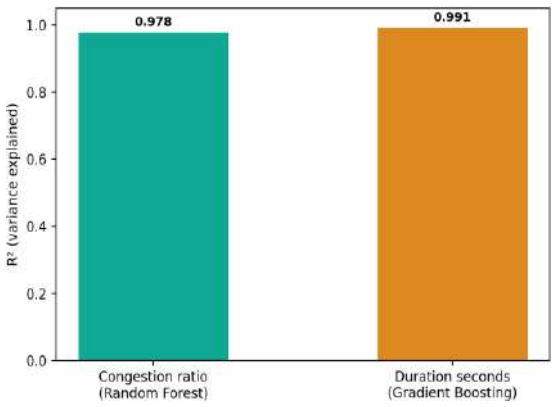



Fig. 4. Traffic prediction model fit ($R^2$), random split.

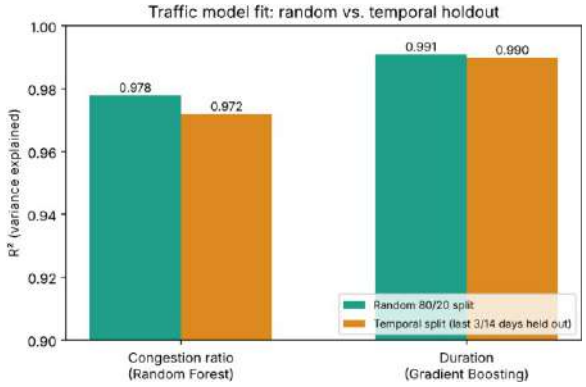



Fig. 5. Traffic model fit under random vs. temporal holdout - the near-identical results indicate the high $R^2$ reflects the dataset's synthetic, noise-free construction rather than train/test leakage.

### *D. Ambulance Dispatch: Heuristic vs. Naive Nearest-Distance*

10 scenarios were run with each origin hospital's own base ambulances excluded, forcing a genuine cross-hospital choice. The heuristic differed from the naive baseline in 2 of 10 cases (Fig. 6): once accepting +0.58 km of pickup distance for a 35.4% pickup-risk reduction, and once selecting an ambulance that was simultaneously closer by real road distance and lower-risk than the straight-line-nearest choice, a distinction only visible when actual road routes, not straight-line distance, are used to rank candidates. This evaluation currently covers 10 scenarios; systematic scaling across ambulance starting position, hospital availability, congestion level, and network load is identified as follow-up work rather than a claim of general superiority over the naive baseline.

### *E. Capacity Simulation*

All four built-in stress scenarios were run against the live simulation service (Fig. 7). At the seeded fleet/bed capacity, the network avoided a hard shortage even under the mass-casualty scenario, but the pressure indicator correctly escalated to critical before any actual shortage occurred in three of the four non-baseline scenarios, the intended early-warning behaviour.

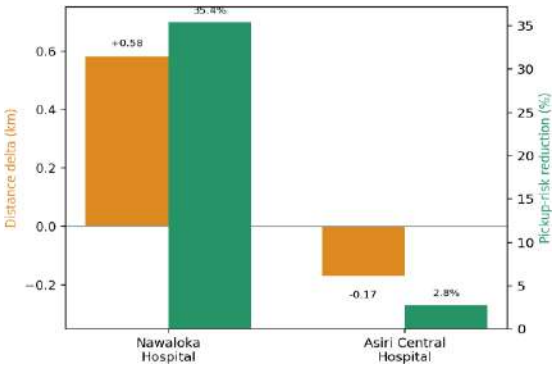



Fig. 6. The two dispatch cases where the heuristic diverged from naive-nearest.

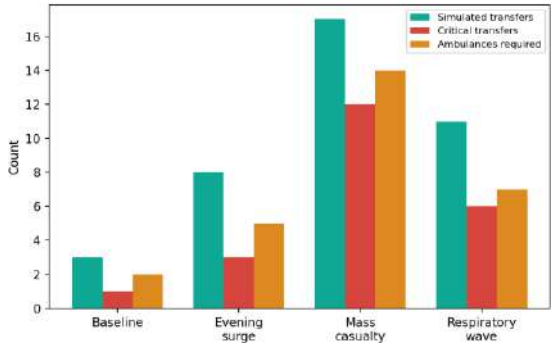



Fig. 7. Simulated demand across four network stress scenarios.

## VI. Discussion And Limitations

A concrete engineering finding emerged during deployment: the development database (SQLite) does not enforce foreign-key constraints by default, which silently masked an insert-ordering defect in which a dependent audit/event row could be written before the parent row it referenced had been flushed to the database. Migrating to a production-grade PostgreSQL/PostGIS deployment, run in a containerized environment, correctly rejected these operations, surfacing three concrete instances of the defect for a fix. This is offered as a specific, evidenced example of why development-database leniency should not be relied upon as a substitute for testing against the target production database engine.

Beyond this, the evaluation is limited in several ways that are stated explicitly here rather than left implicit. First, the traffic-prediction model is trained and evaluated entirely on synthetic data calibrated to resemble Google Routes API output; it has not yet been evaluated against real collected traffic observations, despite the data-collection tooling being implemented and available for that purpose. Second, the urgency-classifier comparison validates internal consistency of the rule-based score rather than its external clinical correctness, since training labels derive from that same rule; an independent clinical validation would require labels from an external source such as clinician-adjudicated cases. Third, the road-risk metric is an engineered, auditable proxy, not a measurement validated against real transport-safety outcomes; the reported 13.3–17.7% risk reductions describe reduction in this engineered score only, and a sensitivity analysis of the routing result to alternative risk

weightings has not yet been run. Fourth, the dispatch-heuristic evaluation covers 10 scenarios and would benefit from systematic scaling across ambulance location, congestion, and network load before any claim of general superiority over the naive baseline. More broadly, hospital and patient-urgency data are seeded/synthetic throughout, and no real clinical outcome dataset or formal user study with hospital staff was used to validate the system's recommendations. The risk-weighting and score-combination coefficients are deliberately hand-set rather than learned, prioritising auditability for a safety-critical recommendation over marginal optimisation gains; this is a design choice appropriate to the current evidence base, not a claim that the coefficients are optimal.

## VII. Conclusion And Future Work

This paper presented an ICU capacity-aware emergency transfer decision-support system for the Colombo hospital network, combining rule-based urgency triage, a learned traffic model, an urgency-weighted multi-objective router over a real road graph with a fully specified risk formula, and a graph-aware ambulance dispatch heuristic, with five quantitative evaluations run directly against production code. Future work includes: validating the traffic model against real, live-collected Google Routes observations rather than synthetic data; clinical validation of the urgency rule against real patient outcome data or clinician-adjudicated labels, independent of the rule used to generate training labels; a sensitivity analysis of routing outcomes to alternative road-risk weightings; scaling the dispatch evaluation to 50–100 systematically varied scenarios; a formal usability study with hospital and ambulance staff; native live GPS/bed-status feed integration in place of the current fleet simulation; and concurrency-safe bed/ambulance claiming under PostgreSQL row-level locking for multi-dispatcher production use.

## References


[1] M. Marzook, M. Rajapaksha, T. Chanakya, and U. Dampage, "A Mobile Application that Employs Artificial Intelligence to Forecast the Required Number of Beds for Intensive Care Units in Sri Lanka," in *2023 IEEE 17th International Conference on Industrial and Information Systems, ICIIS 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 221–226. doi:

[2] P. D. T M *et al.*, "RescueMed: Real-Time Health Data Exchange Through a Secure Mobile and Web-Based Emergency Platform", doi: 10.47772/IJRISS.

[3] C. Olivos and H. Caceres, "MULTI-OBJECTIVE OPTIMIZATION OF AMBULANCE LOCATION IN ANTOFAGASTA, CHILE," *Transport*, vol. 37, no. 3, pp. 177–189, Jun. 2022, doi: 10.3846/transport.2022.17073.

[4] H. Nozari, A. Szmelter-Jarosz, and H. R. Irani, "Designing an ambulance routing optimization model using the combination of machine learning and genetic algorithm in conditions of uncertainty," *Systems and Soft Computing*, vol. 7, Dec. 2025, doi: 10.1016/j.sasc.2025.200276.

[5] M. Almalki, E. Aldhahri, and N. Aljojo, "Ambulance routing optimization based on emergency medical service availability," *Discover Artificial Intelligence*, vol. 5, no. 1, Dec. 2025, doi: 10.1007/s44163-025-00352-3.

[6] C. Selvan, B. H. Anwar, S. Naveen, and S. T. Bhanu, "Ambulance route optimization in a mobile ambulance dispatch system using deep neural network (DNN)," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-95048-0.

[7] S. A. Mahdiraji, M. Juninger, N. Narvell, J. Holmgren, R. C. Mihailescu, and J. Petersson, "Implementing Dynamic Travel Time Calculation in EMS Simulations: Impacts on Prehospital Stroke Care and Transportation," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 781–788. doi: 10.1016/j.procs.2025.02.179.

[8] A. Ferri, "Politecnico di Torino Hybrid Optimization and Simulation Framework for Static and Dynamic Ambulance Allocation A Case Study on Piedmont's emergency medical services."

[9] A. Sivagnanam, A. Pettet, H. Lee, A. Mukhopadhyay, A. Dubey, and A. Laszka, "Multi-Agent Reinforcement Learning with Hierarchical Coordination for Emergency Responder Stationing," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2405.13205

[10] A. Ardiansyah, A. M. Nasution, and M. Iqbal, "Comparative Analysis of Dijkstra and A* Algorithms for Determining the Shortest Route," *bit-Tech*, vol. 8, no. 2, pp. 2974–2983, Dec. 2025, doi: 10.32877/bt.v8i2.3474.

[11] Z. Grujic and B. Grujic, "Optimal Routing in Urban Road Networks: A Graph-Based Approach Using Dijkstra's Algorithm," *Applied Sciences (Switzerland)*, vol. 15, no. 8, Apr. 2025, doi: 10.3390/app15084162.

[12] T. Alghamdi, S. Mostafi, G. Abdelkader, and K. Elgazzar, "A Comparative Study on Traffic Modeling Techniques for Predicting and Simulating Traffic Behavior," *Future Internet*, vol. 14, 2021, doi: 10.3390/10.3390/fi14100294.

[13] H. Ulvi, M. A. Yerlikaya, and K. Yildiz, "Urban Traffic Mobility Optimization Model: A Novel Mathematical Approach for Predictive Urban Traffic Analysis," *Applied Sciences (Switzerland)*, vol. 14, no. 13, Jul. 2024, doi: 10.3390/app14135873.

[14] N. Kumaresh and N. Vanitha, "AI-Powered Multi-Objective Dynamic Route Optimization for Smart Urban Mobility and Logistics." [Online]. Available: www.ijfmr.com

[15] L. Cheng, N. Zhao, and K. Wu, "Stochastic Multi-Objective Multi-Trip AMR Routing Problem with Time Windows," *Mathematics*, vol. 12, no. 15, Aug. 2024, doi: 10.3390/math12152394.

[16] M. Lafuente *et al.*, "A multistate model and its standalone tool to predict hospital and ICU occupancy by patients with COVID-19," *Heliyon*, vol. 9, no. 2, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13545.

[17] N. Stasinos *et al.*, "A Tri-Model Prediction Approach for COVID-19 ICU Bed Occupancy: A Case Study," *Algorithms*, vol. 16, no. 3, Mar. 2023, doi: 10.3390/a16030140.

[18] F. Wan, J. Fondrevelle, T. Wang, and A. Duclos, "Two-stage multi-objective optimization for ICU bed allocation under multiple sources of uncertainty," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-45777-x.

[19] J. Phua *et al.*, "Critical Care Bed Capacity in Asian Countries and Regions," *Crit. Care Med.*, vol. 48, no. 5, pp. 654–662, May 2020, doi: 10.1097/CCM.0000000000004222.

[20] Y. Y. Xu *et al.*, "The emergency medical service dispatch recommendation system using simulation based on bed availability," *BMC Health Serv. Res.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12913-024-12006-8.

[21] A. Karunamurthy and S. Sivasubramanian, "INTERNATIONAL JOURNAL OF COMPUTER APPLICATION AI-Powered Hospital Resource Management and Patient Allocation System Using Weighted Multi-Criteria Decision Modelling and Geospatial Proximity-Capacity Balancing 'AI-Powered Hospital Resource Management and Patient Allocation System Using Weighted Multi-Criteria Decision Modelling and Geospatial Proximity.

[22] F. Da Ros *et al.*, "Supporting Fair and Efficient Emergency Medical Services in a Large Heterogeneous Region," *J. Healthc. Inform. Res.*, vol. 8, no. 2, pp. 400–437, Jun. 2024, doi: 10.1007/s41666-023-00154-1.

[23] V. M. Ton Nathália C O da Silva Angel Ruiz José E Pécora Jr Cassius T Scarpin Valérie Bélanger *et al.*, "Real-Time Management of Intra-Hospital Patient Transport Requests: An Empirical Study," 2023.

[24] A. Qudus Hungbo, C. Adeyemi, and O. Oluwanifemi Ajayi, "Power BI-Based Clinical Decision Support System for Evidence-Based Nurse Decision-Making," *International Journal of Advanced Multidisciplinary Research and Studies*,

# An AI Powered Triage and Dynamic Queue Management System for Outpatient Departments in Sri Lankan Government Hospitals

W. M. H. S. Karunarathna
*Dept. of Software Engineering and Computer Security,*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*

K. K. P. Subashini
*Dept. of Software Engineering and Computer Security,*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
pavithras@nsbm.ac.lk

***Abstract*—Outpatient departments in Sri Lankan government hospitals face persistent overcrowding because patients often need to arrive early and wait physically to obtain a queue position, while conventional first come first served ordering does not account for clinical severity or patient vulnerability. This study proposes an AI Powered Triage and Dynamic Queue Management System that addresses both problems through remote registration and intelligent queue prioritization. Patients can register remotely and receive an estimated arrival time, reducing the need for prolonged physical waiting at the hospital. At registration, an AI assisted severity prediction model evaluates reported symptoms and provides a severity assessment to support prioritization. A Priority Engine then combines symptom severity, self-reported pain, and four vulnerability factors comprising pregnancy, physical disability, older age, and chronic disease, together with a queue ageing mechanism to reduce repeated deferral. A Session Based Queue Generator uses these factors to dynamically construct consultation queues under staff control. The system was implemented using a React frontend, a Node.js and Express backend with MongoDB, and an independent Flask based AI service. Staff interviews and outpatient department observations informed the vulnerability and queue management requirements, while a discrete event simulation was used to evaluate the proposed prioritization approach against conventional first come first served ordering under the same remote registration conditions. In the simulation, priority ordering substantially reduced the estimated waiting time for high-severity patients under the same session-capacity assumptions as FCFS. These results demonstrate simulated operational feasibility, not clinical effectiveness or measured real-world throughput. The proposed framework therefore integrates remote registration, AI assisted triage, vulnerability aware prioritization, and dynamic queue management into a unified approach suited to resource constrained public healthcare settings.**

***Keywords—online OPD registration, estimated arrival time, AI assisted triage, dynamic queue management, vulnerability aware fairness***

## I. Introduction

Outpatient departments (OPDs) are among the busiest points of contact with government hospitals in Sri Lanka, where high patient volumes, long waits, and overcrowded waiting areas place pressure on both patients and staff. A key operational problem is that patients commonly have to arrive at the hospital and complete registration before obtaining a queue position or any indication of when they may be seen. Studies at the National Hospital of Sri Lanka associate prolonged OPD waiting with manual administrative procedures, the absence of computer-based registration, and limited prioritization [1], [2], while research at Lady Ridgeway Hospital for Children reports increased waiting time associated with inefficient registration and documentation [3]. A digital queue system introduced at Galle National Hospital's pharmacy counters in 2026 still issued queue numbers only after patients arrived [4]. These findings indicate that digitizing the queue display alone does not remove the incentive for patients to arrive early and wait physically.

Online registration can address this first source of congestion by allowing patients to submit their information before travelling to the hospital. A retrospective study of a web-based appointment system involving 5,641 outpatients reported lower total waiting time and higher patient satisfaction than conventional walk-in queuing [5], while broader appointment-scheduling research indicates that advance scheduling can improve patient flow but may not adequately represent changing clinical urgency in walk-in-heavy public OPDs [6]. The Sri Lankan OPD-PMMAS prototype further demonstrates the feasibility of remote registration and mobile alerts in the local context [7]. However, registration alone mainly determines how patients enter the system; it does not determine who should be seen first when patients have different levels of clinical need.

Under conventional FCFS queuing, consultation order is largely determined by registration or arrival time, even when patients differ substantially in symptom severity or vulnerability. Regional evidence identifies additional access barriers for pregnant women and persons with disabilities [8], while a structured triage intervention in a Sri Lankan secondary-care hospital reported reductions in OPD waiting time after patient prioritization [9]. These findings support the need for a queue that considers clinical need rather than treating all registered patients as equivalent. In this study, symptom severity is the dominant component of priority, while vulnerability provides additional protection for four predefined groups: pregnant patients, persons with physical disabilities, patients aged 65 years or above, and patients with chronic disease. A queue-ageing term further reduces the risk that lower-priority patients are repeatedly deferred.

Recent research shows that AI can support symptom-based triage and urgency classification when used with appropriate safeguards and human oversight [10]. Other work has explored continuous AI-assisted queue optimization [11] and dynamic triage in low-resource settings [12]. However, the reviewed approaches do not combine online registration with a session-based, vulnerability-aware queue that also explicitly accounts for repeated non-selection across sessions. This study therefore addresses two connected problems: reducing the need for early physical arrival through online registration and improving the fairness and clinical responsiveness of the resulting queue through AI-assisted prioritization. The proposed framework is designed for the operational constraints of Sri Lankan government hospital OPDs and is intended to support, rather than replace, hospital staff and clinical judgement.

This paper makes the following contributions.

An online OPD registration process that allows patients to submit their details before visiting the hospital and receive their queue position and estimated arrival time once the consultation session queue is generated.

An AI-assisted severity-prediction model trained on a re-engineered symptom-severity dataset, supported by a rule-based safety mechanism for high-severity red-flag symptoms.

A Priority Engine that combines symptom severity, self-reported pain, four patient vulnerability factors, and queue ageing into an auditable priority score.

A Session-Based Queue Generator that creates staff-controlled consultation queues, assigns estimated arrival times, and carries deferred patients into subsequent sessions.

A fully implemented three-service prototype and a discrete-event simulation that evaluates priority-based ordering against FCFS under the same online-registration conditions.

## II. Related Work

Digital OPD systems, non-AI structured triage, discrete-event simulation, and AI-assisted queue optimization have each been studied largely in isolation, typically addressing either the physical-queuing problem or the ordering problem, but not both together. Web-based appointment systems reduce total waiting time and improve patient satisfaction by allowing patients to book ahead of time [5], and OPD-PMMAS shows that Sri Lankan patients and staff are receptive to remote, app-based registration specifically [7]. Both, however, treat patients purely by registration or appointment order once they are in the system, without any severity- or vulnerability-based reordering. Discrete-event simulation of a dual-practice specialist clinic shows that matching consultation start times to patient arrival can reduce turnaround time by up to 40% without additional staff or infrastructure, but again does not incorporate severity-based prioritization [13]. Ensemble machine learning methods, particularly XGBoost and Random Forest, consistently outperform simpler statistical models on healthcare classification and resource-allocation tasks structurally similar to severity prediction [14], [15], a finding that directly motivated the model comparison presented in Section III-C. Vulnerability-aware access research from South Asia confirms that disability and pregnancy produce compounding, measurable access barriers in comparable healthcare settings [8], supporting the four-factor vulnerability construct adopted here rather than a single generic special-needs flag.

None of the reviewed systems address session-to-session fairness for patients who are repeatedly not reached. The agentic AI framework continuously re-evaluates urgency but does not explicitly protect against repeated non-selection [11], and low-resource dynamic triage research similarly emphasizes continuous reassessment over inter-session fairness [12]. The queue-ageing mechanism presented in Section III-D is therefore not an adaptation of an existing published mechanism, but a novel component motivated directly by this gap.

## III. Proposed System

### A. System Architecture

The framework is deployed as three independently running services that communicate only over HTTP, shown in Fig. 1. These comprise a React single-page frontend, a Node.js/Express REST API that implements authentication, the Priority Engine, and the Queue Generator, backed by MongoDB, and a Flask microservice that loads the trained severity model and exposes a single POST /predict-severity endpoint. Separating the AI service from the transactional backend isolates the machine-learning dependency stack, comprising scikit-learn, XGBoost, and pandas, from the API stack, comprising Express and Mongoose, so each can be developed and restarted independently.

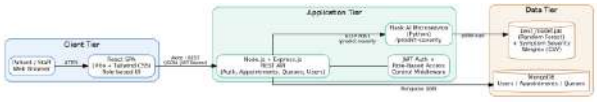


Fig. 1. Three-tier system architecture, comprising a React client, a Node.js/Express/MongoDB backend, and an independent Flask AI microservice.

### B. Remote Registration and Estimated Arrival Time

Online registration addresses the physical-registration problem identified in Section I. A patient, or a family member acting on the patient's behalf, can submit the required details through the web-based registration process before travelling to the hospital. The appointment initially remains in a waiting state. When hospital staff generate a consultation session, the system recalculates the Final Priority Score for all waiting patients, assigns queue positions to the available session capacity, and calculates an Estimated Arrival Time from the session start time, queue position, and average consultation duration. This separates the act of registering from the need to wait physically for registration and allows patients to arrive closer to their expected consultation time.

Remote registration is technically plausible for the population targeted by this research. National statistics report mobile penetration exceeding 140% and internet penetration of approximately 53.6% [16], [17]. However, population-level connectivity does not guarantee independent digital access for every OPD patient. Research involving older Sri Lankans identifies device ownership, connectivity, interface complexity, and language as practical barriers to digital health use [18], while hospital-based evidence indicates that general internet use

does not necessarily translate into use of dedicated mHealth applications [19]. The proposed system therefore allows registration on behalf of a patient by a family member or hospital staff, and its notification design can support channels beyond a smartphone-only workflow.

*C. Primary Evidence Informing the Design*

The system requirements were informed not only by the literature but also by primary evidence collected from OPD staff and non-participant field observation. The semi-structured staff interview examined current patient registration and queue practices, causes of congestion and long waiting times, existing approaches to identifying patients who need earlier attention, the appropriateness of the four vulnerability factors, repeated deferral across sessions, concerns about AI-assisted prioritization, required session-control parameters, and the practicality of a digital staff dashboard. These questions were used to connect operational requirements directly to the Priority Engine and Session-Based Queue Generator rather than treating the software design as a purely technical exercise.

The observation checklist provided an independent view of the pre-system OPD environment. It covered long queues, overcrowding, limited seating and congestion around waiting areas; whether patients were generally served according to arrival order; whether vulnerable patients waited with the general population; and whether patients showed uncertainty about queue progress or expected waiting time. Together, the staff interviews and observations provided contextual evidence for the need for structured prioritization, session-based queue control, and estimated arrival information.

*D. Research and Evaluation Approach*

The study follows a Design Science Research approach in which the software artifact is designed in response to a defined OPD problem and then evaluated through technical and operational evidence. The quantitative component consists of severity-model evaluation and discrete-event simulation, while the qualitative component consists of structured staff interviews and non-participant OPD observation used to inform the operational requirements. This combination is appropriate because technical performance alone cannot establish whether a queue-management design is compatible with the workflow of a government hospital. The primary evidence was therefore used at the design stage to connect observed registration, prioritization, waiting, and session-management issues to explicit system requirements. The conference paper focuses on the implemented artifact and its quantitative evaluation; the patient questionnaire component of the wider research project is intentionally outside the evidence base reported here.

The evaluation was structured around two linked outcomes. The first is the quality of the AI-assisted severity estimate, measured using MAE, RMSE, and severity-class accuracy on held-out data. The second is the effect of the Priority Engine on the distribution of waiting time. For this purpose, the proposed ordering policy was compared with FCFS under identical online-registration conditions. This distinction is important: online registration addresses the physical crowding created by arrival-before-registration, whereas priority ordering addresses the fairness and clinical-responsiveness problem that remains after patients have entered the digital queue. The study therefore avoids attributing the entire waiting-time improvement to the AI component alone.

*E. AI-Assisted Severity Prediction*

The source dataset, comprising 4,920 records covering 41 diseases and 131 symptoms, provides disease labels but no severity measure. Per-symptom severity weights were first re-scored using a triage red-flag categorization informed by Manchester Triage and Emergency Severity Index concepts, grouping symptoms into five urgency bands ranging from non-urgent to life-threatening. A per-patient severity label was then computed as

$$Severity = 0.6 \cdot max(w) + 0.4 \cdot mean(w), \text{ capped at } 10 \quad (1)$$

where w is the set of severity weights of the symptoms present, so that a single red-flag symptom dominates the score rather than being diluted by minor co-occurring symptoms. Three regressors were trained and compared on a held-out test set, reported in Table I. Random Forest was selected as the production model. Because the dataset covers only 41 fixed disease-symptom templates, symptom combinations absent from any single template can be under-predicted. This was observed directly during development, where a stroke-consistent combination scored 4.86 out of 10 under the raw model. A rule-based safety net therefore floors the final score at the weight of any individual reported symptom with a weight of 8 or higher, reflecting standard practice of not relying on a pure statistical model in isolation for safety-critical decisions [10]. The severity targets are derived from the study's symptom-weighting procedure rather than independently assigned clinical labels. Accordingly, the held-out metrics indicate agreement with these constructed targets and should not be interpreted as clinical diagnostic or triage accuracy. Since template-level separation of training and test data is not established here, generalization to unseen disease templates also remains unverified.

The triage component is designed as decision support rather than autonomous clinical decision-making. The severity score is generated from the information supplied during registration and is used to support queue prioritization; it is not presented as a diagnosis. The rule-based safety mechanism is an explicit safeguard for high-severity symptom patterns that may be under-represented in the training data. In addition, the queue-generation workflow remains under staff control: reception staff define the consultation session parameters and can manage the generated queue within the system's authorized workflow. These controls are particularly important in a public healthcare setting, where the model's output should remain interpretable and subject to professional judgement.

TABLE I. SEVERITY-PREDICTION MODEL COMPARISON (N=42 HELD-OUT)

| Model | MAE | RMSE | Class Acc. |
|---|---|---|---|
| Decision Tree | 0.074 | 0.307 | 97.6% |
| Random Forest | 0.049 | 0.247 | 97.6% |
| XGBoost | 0.051 | 0.262 | 97.6% |

*F. Priority Engine and Vulnerability-Aware Fairness*

The Final Priority Score combines four components.

*Priority = Severity + Pain + Vulnerability + Queue Aging (2)*

Pain contributes 0, 1, or 2 points for self-reported low, moderate, or high pain. Vulnerability contributes 1 point for each of four factors: pregnancy, physical disability, age 65 years or above, and chronic disease, giving a range of 0 to 4. These four factors and the practical session-control parameters were informed by structured interviews with OPD registration and nursing staff. Non-participant OPD observation independently examined current patient flow, overcrowding, prolonged queues, and the experience of vulnerable patients. The evidence from these primary sources was used together with the literature to define the queue-management requirements. Aging contributes 0.5, 1, or 1.5 points for one, two, or three or more prior missed sessions, reducing the risk that patients who are repeatedly not reached are permanently disadvantaged.

### *G. Session-Based Queue Generator*

At queue generation, staff specify a queue name, session start time, average consultation time, and capacity. The system takes every currently waiting patient, including those carried over from a previous session and any newly registered since, recalculates each patient's priority score, and sorts by priority in descending order, then by severity in descending order as a tie-break, then by registration time in ascending order as the final tie-break. The top-capacity patients receive sequential queue numbers and an Estimated Arrival Time of session start + (position − 1) × average consultation time, as introduced in Section III-B. Remaining patients stay in Waiting status with an incremented missed-session count, raising their Aging term at the next round. This carry-forward logic, rather than a fixed appointment slot, is what allows the framework to represent the dynamically changing, walk-in-heavy nature of a public OPD that pre-arranged scheduling alone cannot [6].

## IV. Implementation

The system was implemented incrementally and verified module by module against worked numerical examples before integration. The frontend uses React, Vite, and Tailwind CSS with role-based visual theming, using distinct accent colors for patient, receptionist, and admin views, implemented through CSS custom properties so that a single component tree serves all three roles. The backend uses Node.js and Express with MongoDB through Mongoose, chosen over a relational schema because the nested, evolving priority-score sub-objects on each appointment record, comprising severity, pain adjustment, vulnerability, and aging, map naturally onto a document model. Authentication uses JSON Web Tokens with role-based access control middleware that restricts queue-generation and administrative endpoints to the receptionist and admin roles, independently of what the frontend chooses to display. The interface deliberately does not surface a Generate Queue action for the admin role, since queue generation is a real-time, physical-context decision tied to front-desk operations rather than system administration, although the endpoint itself remains authorized for both roles for operational flexibility.

The AI microservice is a Flask application exposing POST /predict-severity, invoked once per registration by the backend. The two services communicate only over a local HTTP call, with no shared code, so the machine-learning and transactional stacks can be developed, tested, and restarted independently. All six modules, comprising registration, AI triage, the Priority Engine, the Queue Generator, Estimated Arrival Time, and the staff dashboard, were verified end-to-end against worked numerical examples prior to evaluation.

### *A. Operational Safety and Human Oversight*

Because the system handles health-related inputs, the prototype separates clinical support from administrative control. The AI service returns a severity score, but the score is not treated as a diagnosis and is not intended to replace clinical judgement. The safety-net rule provides an additional check for high-weight red-flag symptoms, while the queue remains controlled by authorized hospital staff. Role-based access control restricts queue-generation and administrative operations, and the AI microservice is isolated from the transactional backend so that model failures do not require the core appointment and queue services to share the same runtime dependencies. The design also avoids requiring patients to remain physically present while waiting: after registration and queue generation, patients receive an estimated arrival time, while staff retain control of session capacity and consultation parameters. These choices support a cautious deployment model in which automation assists patient-flow management without transferring clinical responsibility to the software.

## V. Evaluation and Results

### *A. AI Severity Model*

Table I reports the held-out comparison of the three severity models. All three models achieved the same severity-class accuracy of 97.6%, while Random Forest produced the lowest MAE (0.049) and RMSE (0.247) and was therefore selected as the production model. The result indicates that model selection was driven by the precision of the continuous severity score used by the Priority Engine rather than by a difference in coarse severity-class accuracy.

### *B. Simulation-Based Comparison with FCFS*

The simulation evaluates the effect of priority-based ordering after online registration, rather than attempting to measure the separate benefit of removing physical registration queues. Both the proposed system and the FCFS baseline therefore assume that patients have registered remotely before arrival and are evaluated under the same session capacities and patient volumes. The question is whether assigning consultation order using the proposed Priority Engine provides a measurable advantage over simple registration order. The implemented Priority Engine and Session-Based Queue Generator were reproduced as a discrete-event simulation and compared with FCFS under otherwise identical conditions.

A synthetic OPD population was generated for each simulated day, comprising an initial 55 patients followed by 30 new registrations before each of 6 session rounds, with a capacity of 28 patients per session. Severity was drawn from a realistic case mix of 55% low, 30% moderate, and 15% high severity on the 1-10 scale, and vulnerability flags were drawn independently at empirically plausible rates of 20% for elderly patients, 22% for chronic illness, 8% for pregnancy, and 7% for disability. Waiting time was measured as the queue position

minus one, multiplied by the average consultation time, consistent with the system's own Estimated Arrival Time semantics described in Section III-E. Results were averaged over 200 independently seeded simulated days. Across the six rounds, this scenario introduces 235 potential patient arrivals (55 initial patients plus 30 before each round) against 168 available service places (28 per round). It therefore represents a capacity-constrained scenario in which a backlog may remain after the final round; the reported results should not be read as showing that every arrival is served within the simulated horizon.

### *C. Evaluation Protocol and Fairness Measures*

Model selection was based on continuous severity-prediction error as well as severity-class agreement. Mean absolute error (MAE) and root mean squared error (RMSE) were used because the predicted score is consumed as a continuous component of the Final Priority Score; severity-class accuracy was retained as an interpretable secondary measure. The held-out test set was not used during model selection. For the queue experiment, the proposed and FCFS policies were exposed to the same patient arrivals, severity distribution, vulnerability distribution, session capacity, consultation duration, and number of sessions. This controlled comparison isolates the effect of ordering policy from the effect of online registration. Waiting time was defined consistently with the system's Estimated Arrival Time calculation, allowing the simulation output to be interpreted directly in terms of the implemented queue semantics.

Fairness was considered in two complementary ways. First, the four vulnerability factors provide additive protection for patients who may experience additional barriers while waiting, without allowing vulnerability alone to override a substantial difference in symptom severity. Second, the Queue Aging Factor increases the priority of patients who remain unserved across sessions. The simulation therefore reports both overall waiting time and group-specific waiting time rather than relying on a single average. This is important because a priority queue can improve the experience of high-severity patients while increasing deferral for some lower-priority patients. Reporting this trade-off makes the evaluation more transparent and provides a basis for future sensitivity analysis of the vulnerability and ageing weights. The waiting-time values are queue-position-based estimates derived from the assumed average consultation duration, not observed elapsed arrival-to-service times. The present experiment does not separately report the final unserved backlog, the full waiting-time distribution or low-priority waiting-time tails, patient dissatisfaction, or prospectively observed triage-error events.

Under the same simulated patient arrivals and session-capacity assumptions, the reported overall average waiting-time estimate is identical at 67.5 minutes for both approaches. The benefit of priority ordering is therefore not a reduction in total system workload but a different distribution of waiting time across patient groups. High-severity patients wait 76% less on average and 54% less in the worst case under the proposed ordering. Vulnerable patients also show an improvement, with average waiting time decreasing from 67.9 to 63.4 minutes. The online-registration mechanism addresses the need for early physical arrival, while AI-assisted priority ordering determines which patients should receive earlier consultation within the resulting queue.

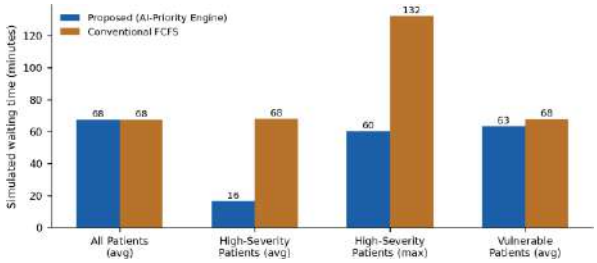



Fig. 1. Simulated waiting time, proposed Priority Engine vs. conventional FCFS, averaged over 200 simulated OPD days.

TABLE II. SIMULATION RESULTS: PROPOSED VS. FCFS (200-DAY AVERAGE)

| Metric | Proposed | FCFS |
|---|---|---|
| Avg. wait, all patients | 67.5 min | 67.5 min |
| Avg. wait, high-severity | 16.5 min | 68.1 min |
| Max wait, high-severity | 60.4 min | 132.4 min |
| Avg. wait, vulnerable | 63.4 min | 67.9 min |

The difference between the severity and vulnerability effects follows from the scoring design. Severity ranges from 1 to 10 and therefore has greater influence on the Final Priority Score, whereas the vulnerability component is capped at 4. Vulnerability is consequently intended to provide additive protection and resolve close cases rather than override substantial differences in symptom severity. This design keeps clinical urgency dominant while ensuring that pregnancy, physical disability, older age, and chronic disease are represented in the queue decision.

The proposed ordering also introduces a fairness trade-off: patients with lower priority may be deferred more often than under strict FCFS. The queue-ageing term is designed to bound this effect by increasing the priority of patients who miss successive sessions, rather than eliminating the trade-off entirely. This is an important property of the proposed design because clinical urgency and strict arrival-order fairness cannot always be maximized simultaneously. Because the ageing contribution is capped, it does not mathematically guarantee that every patient will be served within a fixed number of sessions, particularly under sustained demand exceeding capacity. A formally specified maximum-wait escalation rule and prospective evaluation of staff override procedures are therefore still required.

## VI. DISCUSSION AND LIMITATIONS

Three limitations bound the current evaluation. First, the severity model is trained on a 41-disease symptom-template dataset rather than a large, clinically labelled OPD dataset. Although the rule-based safety mechanism reduces the risk of under-prediction for high-severity red-flag symptoms, broader clinical validation is required. Second, the waiting-time comparison is based on discrete-event simulation using synthetic but empirically motivated patient populations and has not been tested in a live hospital environment. Third, the staff interviews and non-participant observations informed the

vulnerability and queue-management design, but a larger multi-site study is needed to establish whether these factors and workflow assumptions generalize across Sri Lankan government OPDs. The remote-registration design also requires evaluation with patients who have limited digital access or who depend on caregivers for online services. In addition, the derived severity targets have not been independently validated by clinicians, and the current evaluation does not establish generalization across unseen disease templates, languages, or hospital sites. The simulation reports queue-position-based waiting estimates rather than complete arrival-to-service times and does not quantify the final unserved backlog, low-priority waiting-time tails, patient dissatisfaction, or actual triage-error events. The study also does not compare the learned model directly with a deterministic symptom-weight baseline. The capped ageing score cannot guarantee service within a fixed time; maximum-wait escalation and staff override procedures require formal specification and testing. Before deployment, prospective evaluation in at least one OPD should include clinical review of triage outputs, monitoring of served and unserved patients, and assessment of patient safety and satisfaction. Practical adoption would additionally require staff training, accessible multilingual interfaces, and assisted registration for patients who lack devices, connectivity, or digital confidence.

## VII. Conclusion and Future Work

This paper presented an AI-powered triage and dynamic queue management framework for Sri Lankan government hospital OPDs. The framework separates two related problems: online registration reduces the need for patients to arrive early simply to obtain a place in the queue, while AI-assisted priority ordering determines consultation order using symptom severity, pain, vulnerability, and queue ageing. The four vulnerability factors (pregnancy, physical disability, age 65 years or above, and chronic disease) were incorporated with input from staff interviews, field observation, and the literature. In a 200-day discrete-event simulation, priority-based ordering reduced average waiting time for high-severity patients by 76% and worst-case waiting time by 54% compared with FCFS, under identical simulated session-capacity assumptions. The results support the technical feasibility of combining remote registration with vulnerability-aware priority queuing under the simulated conditions; they do not establish clinical effectiveness, patient safety, or real-world throughput. The disclosed fairness trade-off, potential residual backlog, and absence of live clinical deployment indicate the need for further validation. Future work should evaluate the system with larger independently and clinically labelled datasets, template-aware validation, broader patient and staff samples, waiting-time distribution and backlog measures, comparison with a deterministic symptom-weight baseline, sensitivity and uncertainty analysis, formal escalation and override procedures, notification-channel integration, and prospective pilot deployment in government hospital OPDs.

## References


[1] Jayawardena DBAS, “Patients waiting time at out Patient’s Department at the National Hospital Sri Lanka,” *J. Community Med. Public Health*, vol. 1, no. 2, Nov. 2017, doi: 10.29011/2577-2228.100013.

[2] R. Sabhapathige, D. Deerasinghe, and G. Ranasinghe, “Causes of Obstruction in the Outpatient Department of National Hospital, Sri Lanka: A Qualitative Study.”

[3] M. Krishanth, K. Fernando, and D. Liyanage, “Factors Contributing to Increased Patient Waiting Time at Out Patient Department of Lady Ridgeway Hospital for Children in Sri Lanka and Strategies to Minimize the Waiting Time,” *International Journal Of Scientific Advances*, vol. 2, no. 4, 2021, doi: 10.51542/ijscia.v2i4.2.

[4] Dialog Axiata PLC, “Doc990 implements digital queue system for pharmacy services at Galle National Hospital,” [Online]. Available: https://dialog.lk/news/doc990-implements-digital-queue-system-for-pharmacy-services-at-galle-national-hospital.

[5] W. Cao *et al.*, “A web-based appointment system to reduce waiting for outpatients: A retrospective study,” *BMC Health Serv. Res.*, vol. 11, 2011, doi: 10.1186/1472-6963-11-318.

[6] A. Ala and F. Chen, “Appointment Scheduling Problem in Complexity Systems of the Healthcare Services: A Comprehensive Review,” 2022, *Hindawi Limited*. doi: 10.1155/2022/5819813.

[7] M. A. C. Akmal Jahan, S. Rathnayake, M. Kalansuriya, and F. Subair, *OPD-PMMAS: Patient Management and Mobile Alert System for OPDs in Sri Lankan Hospitals-A Prototype*. 2021.

[8] H. R. Devkota, E. Murray, M. Kett, and N. Groce, “Are maternal healthcare services accessible to vulnerable group? A study among women with disabilities in rural Nepal,” *PLoS One*, vol. 13, no. 7, 2018, doi: 10.1371/journal.pone.0200370.

[9] D. H. Liyanage and C. C. Kulathilake, “Impact of a Structured Triage and Patient Flow Intervention on Outpatient Waiting Time in a Secondary Care Hospital in Sri Lanka: A Quasi-Experimental Quality Improvement Study,” *International Journal of Scientific Advances (IJSCIA)*, vol. 7, no. 3, pp. 417–420, 2026

[10] A. Da’Costa, J. Teke, J. E. Origbo, A. Osonuga, E. Egbon, and D. B. Olawade, “AI-driven triage in emergency departments: A review of benefits, challenges, and future directions,” *Int. J. Med. Inform.*, vol. 197, May 2025, doi: 10.1016/j.ijmedinf.2025.105838.

[11] R. Gupta, S. Kumar, and M. Dang, “Agentic AI for Clinical Urgency Mapping and Queue Optimization in High-Volume Outpatient Departments: A Simulation-Based Evaluation,” Mar. 2026, [Online]. Available: http://arxiv.org/abs/2604.00215

[12] U. Bashir, “Exploring an AI-driven dynamic triage system for real-time patient risk reassessment in emergency departments in low-resource settings,” *Front. Digit. Health*, vol. 8, 2026, [13] W. H. Fun *et al.*, “Applying Discrete Event Simulation to Reduce Patient Wait Times and Crowding: The Case of a Specialist Outpatient Clinic with Dual Practice System,” *Healthcare (Switzerland)*, vol. 10, no. 2, Feb. 2022, doi: 10.3390/healthcare10020189.

[14] X. Fan *et al.*, “Prediction of outpatient rehabilitation patient preferences and optimization of graded diagnosis and treatment based on XGBoost machine learning algorithm,” *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1473837.

[15] S. Sezik, M. Ö. Cingiz, and E. İbiş, “Machine Learning-Based Model for Emergency Department Disposition at a Public Hospital,” *Applied Sciences (Switzerland)*, vol. 15, no. 3, Feb. 2025,

[17] T. R. C. of Sri Lanka, “Telecom Statistics of Sri Lanka - Q1 2025,” 2025. [Online]. Available: https://www.trc.gov.lk/content/files/statistics/TSR2025Q1TelecommunicationsStatisticalreport21052025331pmforupdateTRCSLweb.pdf

[18] B. J. Karunarathna, S. Maduwage, and M. S. D. Wijesinghe, “Age-friendly eHealth strategies: Qualitative insights from older persons in Sri Lanka on digital health information access

[19] A. G. T. Sithijaya, T. Suwathijah, D. Tharmarajah, I. P. Liyanage, and S. Rathnayake, “eHealth Literacy and the Use of Internet and mHealth Applications in Health Information Seeking Among Patients Attending the Dental Teaching Hospital, Peradeniya, Sri Lanka,” 2023, [Online]. Available: https://www.researchgate.net/publication/374913834_eHealth_Literacy_and_the_Use_of_Internet_and_mHealth_Applications_in_Health_Information_Seeking_Among_Patients_Attending_the_Dental_Teaching_Hospital_Peradeniya_Sri_Lanka

# Multimodal Deep Learning for NSCLC Recurrence Prediction: Gated Fusion of Histopathology and Clinical Data with Clean Label Derivation and Explainable AI

Piyara Morawakaarachchi
*Department of Sofware Engineering and Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
ptmorawakaarachchi@students.nsbm.ac.lk

Thisarani Wickramasinghe
*Department of Sofware Engineering and Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

***Abstract*—Non-Small Cell Lung Cancer (NSCLC) accounts for approximately 85% of all lung cancer cases, with postoperative recurrence rates of 30 to 55% representing a critical unmet clinical need. Existing prognostic models predominantly rely on single-modality inputs and frequently employ publicly available datasets without rigorous label quality verification, introducing systematic bias into reported performance metrics. This paper presents an adaptive multimodal deep learning framework for NSCLC recurrence prediction that combines ResNet-50 histopathology image features with structured clinical data through a Gated Multimodal Unit (GMU) fusion architecture with SHAP explainability. Recurrence labels are derived cleanly as progression within two years of diagnosis from the curated TCGA Pan-Cancer Clinical Data Resource, excluding patients censored earlier. A leakage audit of The Cancer Genome Atlas (TCGA) clinical records shows that vital status, follow-up length and outcome-dependent missingness raise AUC from 0.652 with six baseline features to 0.779. On 281 patients with embeddings from their own slides, evaluated with 5 × 5 repeated cross-validation and fold-internal preprocessing, a clinical-only model reached AUC 0.613 (logistic regression 0.661), while GMU fusion reached 0.579 and performed no better than a permuted-image control (0.589). SHAP identified pathological stage as the most influential predictor. The results show that feature and label quality, rather than fusion architecture, determine performance in public-data recurrence prediction, and that ImageNet features from predominantly frozen TCGA slides add no measurable prognostic signal.**

***Keywords*—NSCLC, recurrence prediction, multimodal deep learning, data leakage, explainable AI**

## I. Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide, with Non-Small Cell Lung Cancer (NSCLC) accounting for approximately 85% of all diagnosed cases [1]. Despite advances in surgical resection and adjuvant therapy, a substantial proportion of patients experience recurrence, frequently detected at a late and less treatable stage. Accurate individual-level recurrence risk stratification at the time of surgery could enable personalized surveillance, earlier intervention, and more informed adjuvant therapy decisions.

A systematic review of published NSCLC risk prediction models reported AUC values ranging from 0.50 to 0.90 [2]. Most rely on a single data modality, either structured clinical variables or imaging-derived features, limiting their ability to capture both tumor biology and patient context simultaneously. Multimodal approaches that integrate histopathology with clinical data offer a more biologically grounded alternative [3], [4].

A further concern largely unaddressed in the literature is the integrity of labels and features in publicly available datasets. The Cancer Genome Atlas (TCGA) is the most widely used public resource for NSCLC prognostic modelling, yet its clinical records combine variables collected at diagnosis with variables updated during follow-up. When such variables are used as predictors, models learn the outcome rather than predict it, a form of data leakage recognised as a major cause of irreproducible results [5].

This paper makes three contributions. First, a leakage audit of TCGA recurrence modelling quantifies how vital status, follow-up length and outcome-dependent missingness inflate performance, and a clean label is derived from a curated endpoint with a defined prediction time and horizon. Second, five strategies, clinical-only, image-only, weighted fusion, concatenation fusion and GMU fusion, are compared on patients with their own TCGA whole-slide images under a single leakage-free protocol, with permuted-input controls and paired comparisons. Third, the framework provides SHAP-based explanation of clinical predictions and an analysis of GMU gating behavior.

Section II reviews related work in NSCLC prognostic modelling, multimodal learning, and label quality public cancer datasets. Section III describes dataset construction, label derivation, preprocessing, system architecture, and experimental setup. Section IV reports results. Section V discusses findings, limitations, and future directions. Section VI concludes the paper.

## II. Related work

### A. Clinical Background and Artificial Intelligence in Non-Small Cell Lung Cancer

NSCLC treatment has evolved from surgery and platinum-based chemotherapy to targeted therapies directed at driver mutations such as EGFR, ALK and ROS1 [6], and to immune checkpoint inhibitors [7]. Reliable biomarkers for recurrence nevertheless remain an unresolved clinical challenge. Convolutional neural networks learn hierarchical representations from histopathology images; Coudray et al. established a benchmark for NSCLC subtype classification and mutation prediction from whole-slide images [3], and reviews of deep learning on lung histology highlight the computational cost of whole-slide analysis [8].

### B. Multimodal Learning and Fusion Strategies

Combining complementary modalities often improves predictive performance over unimodal approaches. Mobadersany et al. combined histology and genomics to predict cancer survival [4], and integration of histopathology embeddings with clinical records has improved diagnostic classification [9]. Fusion strategies are categorised as early, intermediate, late or adaptive. The Self-and-Mutual Attention network introduced bidirectional interaction between CT and genomic features for NSCLC recurrence prediction [10]. Steyaert et al. noted that the benefit of fusion depends on cohort size and on each modality carrying patient-level signal [11]. The Gated Multimodal Unit (GMU) of Arevalo et al. learns per-sample modality weighting through a lightweight sigmoid gate [12]. Because whole-slide images contain thousands of patches, offline patch embedding and weakly supervised aggregation make slide-level learning feasible on limited hardware [13].

### C. Explainability and Label Quality

SHAP applies Shapley values to quantify individual feature contributions [14], and resource-aware, interpretable AI is considered essential for clinical deployment [15]. The integrity of outcome labels and predictors in public datasets has received less attention. Kapoor and Narayanan catalogued leakage in machine-learning-based science, including features unavailable at prediction time and preprocessing outside the training fold [5]. For TCGA, Liu et al. curated standardised survival endpoints, including progression-free interval (PFI), precisely because raw clinical fields are inconsistent [16]. Reporting guidance for clinical prediction models requires a defined prediction time, outcome and handling of missing data [17].

### D. Research Gap

Systematic comparison of fusion strategies on patient-linked TCGA slides under a leakage-free protocol remains limited, and leakage through outcome variables and missingness patterns in TCGA NSCLC recurrence modelling has not been quantified in prior work.

## I. methodology

### E. Dataset Construction and Label Derivation

Clinical data for the TCGA-LUAD and TCGA-LUSC cohorts were obtained from the Genomic Data Commons (GDC). After deduplication, 1,025 patients were retained (521 LUAD, 504 LUSC). The prediction time was fixed at initial pathological diagnosis, so only information available at that point may be used as a predictor. A fixed random seed of 42 was used throughout.

Clean labels were derived from the PFI endpoint of the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) [16] with a recurrence horizon of two years. A patient was labelled positive if a PFI event (recurrence or progression) occurred within 730 days of diagnosis and negative if PFI-free follow-up exceeded 730 days. Patients censored before 730 days (n = 395) were excluded, because their two-year outcome is unknown. This gave 630 labelled patients (267 positive).

The four-tier labels (GOLD, SILVER, BRONZE, NEG) of the submitted version were retained for validation only. As implemented, GOLD denoted imaging- or biopsy-confirmed recurrence, SILVER recurrence with a recorded date, BRONZE recurrence without a date and NEG no recurrence recorded; vital status was not used in the tiers, no minimum follow-up was applied to NEG, and this description corrects the submitted text. Against the two-year label, 36 of 328 NEG patients had progressed within two years, which shows why a censoring-aware label is needed.

### F. Clinical Feature Extraction and Feature-Timing Audit

Every candidate predictor was audited for when it is recorded relative to the prediction time (Table I). Vital status, follow-up length and treatment fields are recorded after diagnosis and were removed. The audit also found two extraction defects. The subtype variable took a single value for all patients and was rebuilt from the TCGA project. Missing values were outcome-dependent: 72% of the 226 patients with missing stage had a positive tier label, against 18% of the rest, with the same pattern for T, N, laterality and prior malignancy, whereas smoking variables showed no dependence. Stage was therefore taken from TCGA-CDR, which agreed with the original values in 99% of patients where both existed, and the affected TNM fields were dropped. Six features were retained: age, sex, AJCC pathological stage, subtype, smoking status and pack-years. Median imputation and standardisation were fitted on training folds only.

TABLE I
FEATURE-TIMING AUDIT

| Feature | Source | Recorded | Decision |
|---|---|---|---|
| Age, sex | GDC | Diagnosis | Keep |
| AJCC pathological stage | TCGA-CDR | Resection | Keep (re-sourced) |
| Histological subtype | TCGA project | Diagnosis | Keep (fixed) |
| Smoking status, pack-years | GDC exposures | Baseline | Keep |
| T, N, M, laterality, prior malignancy | GDC | Baseline* | Drop |
| Vital status | GDC | Last contact | Drop (outcome) |
| Follow-up length | GDC | After outcome | Drop (outcome) |
| Prior / subsequent treatment | GDC | Post-diagnosis | Drop |

*Baseline values, but their missingness in the original extraction was strongly associated with the outcome.

### G. Patient-Linked Histopathology

In place of the subtype-matched LC25000 proxy images of the submitted version, each patient's own TCGA slide was used. For 463 patients, one slide per patient (files of 200 MB or less) was downloaded from the GDC, processed and deleted, keeping only one slide on disk at a time. One hundred 256 × 256-pixel patches with more than 60% tissue were sampled at approximately 20× magnification, encoded with an ImageNet-pretrained ResNet-50 [18], and mean-pooled into a 2048-dimensional slide embedding. Of the 281 labelled patients with a slide, 42 had formalin-fixed diagnostic slides and 226 frozen tissue sections (13 unidentified). Within each training fold, embeddings were standardised and reduced to 32 principal components.

### H. System Architecture

Fig. 1 illustrates the framework. The clinical encoder maps the six features through Linear(6, 32), ReLU, Dropout(0.3), Linear(32, 16) and ReLU. The image encoder applies the same structure to the 32 image components. GMU fusion [12] computes a per-patient gate from the image and clinical embeddings h_img and h_clin:

$$g = \sigma(W_g\,[h_{img}\,;\,h_{clin}] + b_g) \quad (1)$$

$$z = g \odot \tanh(W_i\,h_{img}) + (1-g) \odot \tanh(W_c\,h_{clin}) \quad (2)$$

The fused representation z passes through a linear layer and a sigmoid to give a recurrence probability. Two fusion baselines were evaluated: scalar weighted fusion, which mixes the two branch outputs with a single learnable weight for all patients, and concatenation fusion, which applies a linear layer to the concatenated embeddings.

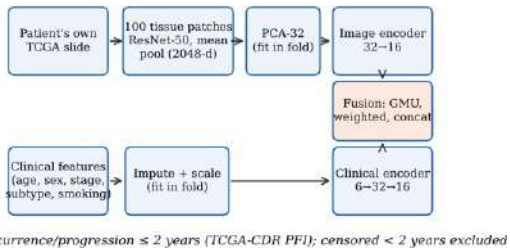



Fig. 1. Revised multimodal framework: patient-linked slide features and clinical features are encoded separately and combined by GMU, weighted or concatenation fusion. All preprocessing is fitted within training folds.

### I. Experimental Setup and Evaluation Protocol

Models were trained with AdamW (learning rate $10^{-3}$, weight decay $10^{-3}$) and class-weighted binary cross-entropy for up to 300 epochs, with early stopping (patience 30) on an inner 20% validation split of each training fold. All hyperparameters were fixed before evaluation. Evaluation used 5-fold stratified cross-validation repeated five times (25 folds) at patient level. Two non-neural baselines, logistic regression on clinical features and stage alone, used the same folds. Negative controls were logistic regression on permuted labels and GMU fusion with slide embeddings permuted across patients. AUC is reported as mean ± SD over folds with the 2.5th–97.5th percentile range; pooled AUC, Brier score and expected calibration error (ECE, 10 bins) were computed from predictions averaged over repeats. Same-patient model comparisons are reported as paired fold-wise AUC differences. All numbers are generated by a single script from one results file. Training ran on a CPU in minutes, so a consumer GPU is no longer required. To assess overfitting, AUC on the training fold was recorded alongside test-fold AUC for every model. Sensitivity analyses repeated the main comparison with a three-year horizon and with the original tier labels, with and without BRONZE patients.

## I. results

### J. Main Performance Results

Table II and Fig. 2 report performance on the 281 patients with both modalities (108 positive). The clinical-only MLP reached AUC 0.613 ± 0.089, and logistic regression on the same features 0.661 ± 0.078, close to stage alone (0.651). The image-only model reached 0.478 ± 0.076, indistinguishable from chance. All three fusion models scored below the clinical-only MLP: weighted fusion 0.569, concatenation 0.560 and GMU 0.579. Paired differences against the clinical-only MLP were -0.034 for GMU (fold range -0.156 to 0.132), -0.052 for concatenation and -0.044 for weighted fusion. GMU with permuted slide embeddings reached 0.589, so real embeddings gave no advantage over permuted ones (paired difference -0.010). On the larger clinical cohort (n = 630), the clinical-only MLP reached 0.621 ± 0.050 and logistic regression 0.652 ± 0.040.

TABLE II
CROSS-VALIDATED PERFORMANCE (N = 281, 5 × 5 CV)

| Model | AUC mean ± SD | 2.5–97.5% | Brier | ECE |
|---|---|---|---|---|
| Stage only | 0.651 ± 0.053 | 0.578–0.754 | – | – |
| Logistic regression (clinical) | 0.661 ± 0.078 | 0.508–0.780 | 0.220 | 0.044 |
| Clinical-only MLP | 0.613 ± 0.089 | 0.454–0.743 | 0.234 | 0.110 |
| Image-only MLP | 0.478 ± 0.076 | 0.319–0.607 | 0.262 | 0.121 |
| Weighted fusion | 0.569 ± 0.070 | 0.428–0.660 | 0.242 | 0.111 |
| Concatenation fusion | 0.560 ± 0.080 | 0.411–0.676 | 0.245 | 0.108 |
| GMU fusion | 0.579 ± 0.065 | 0.476–0.695 | 0.237 | 0.105 |
| GMU, permuted images | 0.589 ± 0.077 | 0.440–0.726 | 0.233 | 0.138 |

Brier score of a constant prediction at the prevalence: 0.237. Ranges are the 2.5th–97.5th percentiles of 25 fold AUCs.

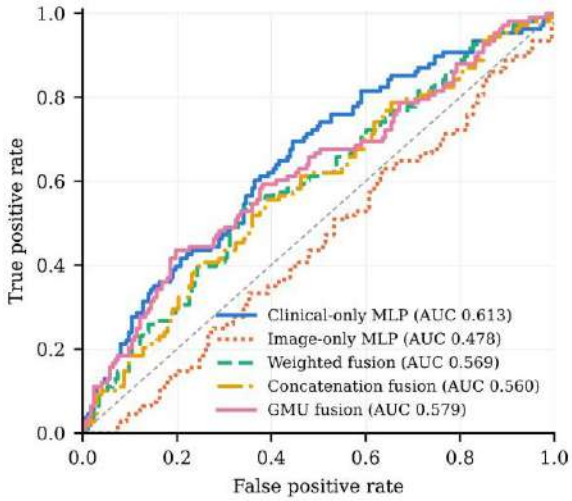



Fig. 2. ROC curves of the five models on the patient-linked cohort (predictions averaged over repeats; legend shows mean fold AUC).

### *K. Label Contamination and Leakage Ablation*

Table III quantifies leakage on the 630 labelled patients using logistic regression, so that differences reflect inputs rather than model capacity. The 12 features of the submitted version reached AUC 0.779 ± 0.040. Vital status alone reached 0.696, above the six clean features combined (0.652 ± 0.040), and adding it to the clean features raised AUC to 0.751. The missing-value pattern of the original extraction alone reached 0.687, and the original extraction without vital status still reached 0.714. Follow-up length reached 0.993, because it nearly determines a horizon-based label. The permuted-label control gave 0.521. This analysis replaces the prior-treatment ablation of the submitted version.

TABLE III
LEAKAGE ABLATION (N = 630, LOGISTIC REGRESSION)

| Features | AUC mean ± SD |
|---|---|
| Submitted 12 features (incl. vital status) | 0.779 ± 0.040 |
| Original extraction, vital status removed | 0.714 ± 0.041 |
| Vital status only | 0.696 ± 0.043 |
| Missingness pattern only | 0.687 ± 0.047 |
| Follow-up length only | 0.993 ± 0.007 |
| Clean 6 features + vital status | 0.751 ± 0.040 |
| Clean 6 features | 0.652 ± 0.040 |
| Stage only | 0.630 ± 0.039 |
| Clean 6 features, permuted labels | 0.521 ± 0.093 |

### *L. Explainability Analysis*

SHAP values for the clinical logistic model (Fig. 3) identified pathological stage as the most influential predictor (mean |SHAP| 0.436), followed by subtype (0.381); age, sex, pack-years and smoking status contributed little. Higher stage increased predicted risk, consistent with the established prognostic value of stage. The Grad-CAM analysis of the submitted version is not reported, because the image encoder is now a frozen ImageNet network and no pathologist review of attention maps was available.

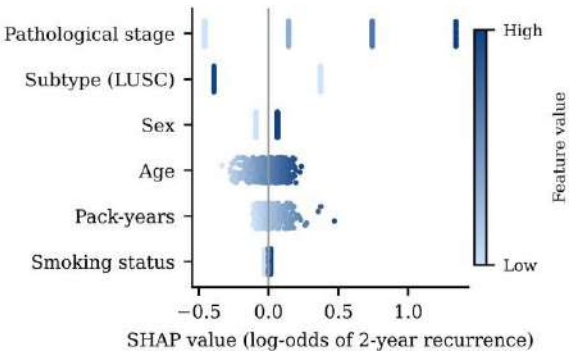



Fig. 3. SHAP values of the clinical logistic model (n = 630). Each point is a patient; colour shows the feature value.

### *M. Gate Behaviour*

Fig. 4 shows the GMU image gate averaged over repeats. The mean image weight was 0.492 (range 0.407–0.569), so the gate stayed close to an even split. It showed a weak association with stage (r = 0.128, p = 0.033) and none with recurrence (r = 0.019, p = 0.756). Together with the permuted-image result, this indicates that the gate did not find patient-specific image information to exploit.

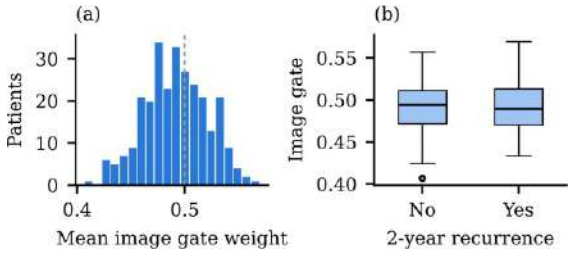



Fig. 4. GMU gating: (a) distribution of the mean per-patient image gate weight; (b) image gate weight by two-year recurrence status.

### *N. Overfitting and Sensitivity Analyses*

Table IV compares training-fold and test-fold AUC. Logistic regression showed the smallest gap (0.036; train 0.698, test 0.661), followed by the clinical-only MLP (0.068). Every model with an image branch showed a much larger gap: 0.214 for the image-only MLP and 0.154–0.178 for the fusion models. The image branch therefore fitted patterns in the training folds that did not generalise, which explains why fusion scored below the clinical-only model.

TABLE IV
TRAINING- VS TEST-FOLD AUC (N = 281)

| Model | Train AUC | Test AUC | Gap |
|---|---|---|---|
| Logistic regression (clinical) | 0.698 | 0.661 | 0.036 |
| Clinical-only MLP | 0.681 | 0.613 | 0.068 |
| Image-only MLP | 0.692 | 0.478 | 0.214 |
| Weighted fusion | 0.743 | 0.569 | 0.174 |
| Concatenation fusion | 0.738 | 0.560 | 0.178 |
| GMU fusion | 0.733 | 0.579 | 0.154 |
| GMU, permuted images | 0.753 | 0.589 | 0.164 |

Mean over 25 folds. Gap = training-fold AUC minus test-fold AUC.

Table V repeats the main comparison under alternative label definitions. With a three-year horizon (n = 234, 127 positive), logistic regression reached 0.631 and GMU 0.558, again

matching its permuted-image control (0.555). With the original tier labels, which count patients with short follow-up as negatives, all models were close to chance (logistic regression 0.537 with all tiers and 0.524 without BRONZE), consistent with these labels being noisier than the censoring-aware definition. In every setting, GMU fusion did not exceed logistic regression on clinical features, and its difference from the permuted-image control was small relative to fold-to-fold variation (at most 0.033 AUC). The main conclusion therefore does not depend on the choice of horizon or label tier.

TABLE V
SENSITIVITY TO LABEL DEFINITION (TEST AUC, 5 × 5 CV)

| Label definition | n (pos.) | LR | MLP | GMU | GMU perm. |
|---|---|---|---|---|---|
| 2-year horizon (main) | 281 (108) | 0.661 | 0.613 | 0.579 | 0.589 |
| 3-year horizon | 234 (127) | 0.631 | 0.591 | 0.558 | 0.555 |
| Tier labels, all | 463 (138) | 0.537 | 0.493 | 0.522 | 0.489 |
| Tier labels, no BRONZE | 368 (43) | 0.524 | 0.488 | 0.500 | 0.487 |

LR: logistic regression on clinical features; MLP: clinical-only MLP; GMU perm.: GMU with slide embeddings permuted across patients. Tier labels apply no censoring rule.

## I. discussion

### *O. Label Quality as the Primary Bottleneck*

The leakage audit shows that most of the performance of the submitted version came from information recorded after the outcome. Vital status and follow-up length encode the outcome directly. Outcome-dependent missingness is subtler: when stage and TNM fields are missing more often for patients who recurred, the missing-value pattern itself predicts the label, and median imputation does not remove this signal. Fold-internal preprocessing does not protect against any of these pathways; only auditing feature timing and missingness does. Recurrence studies using TCGA clinical records should report such audits.

### *P. Clinical Features Drive Current Performance*

Under leakage-free conditions, clinical features reached AUC 0.652 on 630 patients, and much of this signal came from pathological stage. Logistic regression performed at least as well as the MLP (paired difference 0.049 in favour of logistic regression on the slide cohort), which is expected with six features and a few hundred patients. Calibration was modest: Brier scores were close to that of a constant prediction, so predicted risks carry limited individual information.

### *Q. Fusion Strategies with Patient-Linked Images*

With patient-linked slides, no fusion strategy improved on the clinical-only model, and GMU matched its permuted-image control. Several factors likely explain this. Most slides were frozen sections, whose morphology is poorer than that of diagnostic formalin-fixed slides, and restricting to files of 200 MB or less favoured them. Patch features came from an ImageNet encoder rather than a pathology-specific model, and averaging 100 random patches dilutes focal tumour regions. With 108 positive cases, fusion models with more parameters are also more prone to overfitting than the clinical-only model. The results therefore show that this image pipeline adds no measurable signal, not that histology lacks prognostic value. The higher GMU AUC reported in the submitted version should not be interpreted as a multimodal benefit, because its image input was not patient-specific and its clinical inputs included vital status. The training–test gaps in Table IV support this: the image branch added capacity that fitted the training folds but did not generalize.

### *R. Limitations*

The cohort is limited to TCGA, a predominantly US-based academic population, and no external validation was performed. Excluding patients censored before two years reduces sample size and may introduce selection bias. TNM components were dropped rather than re-extracted. Repeated cross-validation folds overlap, so fold percentile ranges are descriptive rather than formal confidence intervals. The image analysis is limited by slide type, encoder and pooling.

### *S. Future Work*

Four extensions follow. Diagnostic formalin-fixed slides should replace frozen sections. A pathology foundation model [19] with attention-based multiple-instance aggregation [20] should replace ImageNet features and mean pooling. Genomic variables such as EGFR, KRAS and TP53 status could form a third modality. Validation on independent cohorts is required before any clinical use.

## I. conclusion

This paper presented a multimodal framework for NSCLC recurrence prediction combining patient-linked histopathology features and clinical data through GMU fusion, with clean label derivation and SHAP explainability. A leakage audit showed that vital status, follow-up length and outcome-dependent missingness raise AUC from 0.652 to 0.779 in TCGA recurrence modelling. With a two-year horizon from a curated endpoint and patients' own slides, clinical features reached AUC 0.661 (logistic regression) and 0.613 (MLP), while GMU fusion reached 0.579 and did not exceed a permuted-image control. Label and feature quality, rather than fusion architecture, determine performance in public-data recurrence prediction, and patient-specific, diagnostic-quality histology features are the essential next step.

## ACKNOWLEDGMENT

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

## REFERENCES

[1] American Cancer Society, "Cancer Facts and Figures 2024," Atlanta, GA, 2024.

[2] Samuel et al., "Risk prediction models for recurrence after curative treatment of early-stage or locally advanced lung cancer: a systematic review," 2026, doi: 10.1002/aac2.70023.

[3] N. Coudray et al., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," Nature Medicine, vol. 24, no. 10, pp. 1559–1567, 2018.

[4] P. Mobadersany et al., "Predicting cancer outcomes from histology and genomics using convolutional networks," Proc. Natl. Acad. Sci. USA, vol. 115, no. 13, pp. E2970–E2979, 2018.

[5] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, p. 100804, 2023.

[6] Makarem and Jänne, "Top advances of the year: targeted therapy for lung cancer," Cancer, 2024, doi: 10.1002/cncr.35423.

[7] M. Rakaee et al., "Deep learning model for predicting immunotherapy response in advanced non-small cell lung cancer," JAMA Oncology, vol. 11, no. 2, pp. 109–118, 2025.

[8] A. Davri et al., "Deep learning for lung cancer diagnosis, prognosis and prediction using histological and cytological images: a systematic review," Cancers, vol. 15, no. 15, p. 3981, 2023.

[9] Khandelwal et al., "Multimodal fusion of histopathology images and electronic health records for early breast cancer diagnosis," arXiv:2604.17122, 2026.

[10] Ai et al., "SAMA: a self-and-mutual attention network for accurate recurrence prediction of non-small cell lung cancer using genetic and CT data," IEEE J. Biomed. Health Inform., 2025, doi: 10.1109/JBHI.2024.3471194.

[11] S. Steyaert et al., "Multimodal data fusion for cancer biomarker discovery with deep learning," Nature Machine Intelligence, vol. 5, pp. 351–362, 2023.

[12] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in Proc. ICLR Workshop, 2017, arXiv:1702.01992.

[13] M. Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," Nature Biomedical Engineering, vol. 5, pp. 555–570, 2021.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, pp. 4765–4774, 2017.

[15] M. Haghighat, G. Aglin, and P. Vepakomma, "Resource-aware multimodal intelligence for medical image analysis," Medical Image Analysis, vol. 101, p. 103432, 2025.

[16] J. Liu et al., "An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics," Cell, vol. 173, no. 2, pp. 400–416, 2018.

[17] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, p. e078378, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770–778.

[19] R. J. Chen et al., "Towards a general-purpose foundation model for computational pathology," Nature Medicine, vol. 30, pp. 850–862, 2024.

[20] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Proc. ICML, 2018, pp. 2127–2136.

# MediHealth Predict: An Explainable Machine Learning Web Platform for Early Multi-Disease Risk Prediction

G.M.A.V. Senevirathna
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University, Sri Lanka*
gmavsenavirathna@students.nsbm.ac.lk

Thilini Bakmeedeniya
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University, Sri Lanka*
thilini.b@nsbm.ac.lk

***Abstract*—Early detection of chronic non-communicable diseases remains a major public-health challenge, particularly in low-resource settings where a large proportion of cases remain undiagnosed until complications appear. This paper presents MediHealth Predict, a full-stack web platform that estimates an individual's risk of three chronic conditions, namely type 2 diabetes, cardiovascular (heart) disease and chronic kidney disease, from routinely available clinical and lifestyle inputs. The system couples a React single-page application with a Flask REST backend, role-based access control for six roles, and three independently trained machine-learning models. The contribution is one of system integration and evaluation practice rather than a new learning algorithm: a leakage-safe twelve-stage training pipeline performs the train-test split before any supervised step, then applies feature selection, SMOTE class balancing and standardisation inside the training data only, and selects among logistic regression, random forest and XGBoost candidates by stratified cross-validation on the training partition, with the test partition used once. On the held-out test set the diabetes model achieved [86.1] percent accuracy and a ROC-AUC of [0.93], and the heart model achieved [81.6] percent accuracy, each reported with [95% confidence intervals, PR-AUC and Brier score]. Every prediction is accompanied by SHAP-based feature attributions and natural-language guidance generated by a large language model; the explanation layer is checked for stability but has not been evaluated with clinicians. The kidney model, although superficially accurate, achieved a ROC-AUC close to 0.5, is reported as a negative result and is withheld from clinical use. The results indicate that integrated, explainable multi-disease screening tools are technically feasible, but that data quality and external validation ultimately govern clinical validity.**

***Keywords—machine learning, disease risk prediction, explainable AI, SHAP, diabetes, cardiovascular disease, chronic kidney disease, web application, class imbalance***

## I. Introduction

Chronic non-communicable diseases are now the leading cause of death and disability worldwide, and three of the most significant contributors are diabetes, cardiovascular disease and chronic kidney disease. According to the International Diabetes Federation, approximately 589 million adults aged between 20 and 79 were living with diabetes in 2024, equivalent to about one in nine adults, and this figure is projected to rise to 853 million by 2050 [1]. Crucially, an estimated 43 percent of these cases, around 252 million people, remain undiagnosed. Chronic kidney disease shows a similar pattern: it is estimated to affect more than 800 million people globally, yet in many low-resource settings as many as nine in ten of those affected are unaware of their condition [2]. Cardiovascular disease, meanwhile, remains the single largest cause of death globally [3]. A common thread across all three conditions is that early intervention substantially improves outcomes, whereas late detection leads to irreversible damage and far higher treatment costs.

These pressures are felt acutely in developing health systems such as that of Sri Lanka, where specialist screening is not always within easy reach and where the asymptomatic early phase of these diseases allows them to progress undetected. An accessible tool that can flag elevated risk from commonly measured indicators could therefore encourage earlier clinical consultation and lifestyle change.

The economic argument reinforces the clinical one. Once these conditions reach an advanced stage they demand resource-intensive interventions such as dialysis, cardiac surgery or the management of diabetic complications, each of which is far costlier than the screening and lifestyle measures that can avert them; systematic economic evaluations of chronic-disease screening report favourable cost-effectiveness ratios under a wide range of settings, particularly for at-risk sub-populations [15]. Opportunistic screening, in which risk is estimated whenever a person interacts with the health system or a self-service tool, is therefore an attractive strategy, provided that the estimate is both trustworthy and explained well enough to motivate action.

Despite the clear value of early screening, several practical barriers persist. First, many existing decision-support tools address only a single disease, forcing clinicians and patients to consult fragmented systems. Second, high-performing machine-learning models are frequently opaque, and a prediction that cannot be explained is difficult for a clinician to trust or act upon. Third, screening tools are often built as offline scripts rather than accessible, secure, multi-user applications suitable for a real clinical workflow.

This paper presents MediHealth Predict, an integrated web platform that addresses these barriers. The system screens for diabetes, heart disease and chronic kidney disease through a single interface, pairs every prediction with an explanation of

the factors that drove it, and enforces role-based access control so that patients, nurses, doctors, laboratory technicians and administrators each see an appropriate view. We state the scope of the contribution plainly: the individual components (logistic regression, random forests, XGBoost, SMOTE, SHAP and large-language-model guidance) are established techniques, and no new learning algorithm or explanation method is proposed. The contributions are (i) an integrated, deployable multi-disease platform in which models, scalers and feature orderings are persisted and reapplied consistently at prediction time; (ii) a leakage-safe training and evaluation protocol in which the data are split before any supervised step, all selection and tuning is confined to training folds, and the test partition is used once, reported together with majority-class and candidate-model baselines, an ablation, calibration and confidence intervals; (iii) the coupling of per-prediction SHAP attributions with constrained language-model guidance, together with a stability check of the attributions; and (iv) an explicit negative result, in which a model that is accurate but non-discriminative is identified and withheld from clinical use. The platform that operationalises this pipeline is described briefly in Section V.

The remainder of this paper is organised as follows. Section II reviews related work. Section III states the requirements. Section IV describes the system design, and Section V the implementation. Section VI reports the results and evaluation, Section VII discusses limitations and ethical considerations, and Section VIII concludes with directions for future work.

## II. Related Work

### A. Machine learning for early detection

Supervised machine learning has been widely applied to structured clinical data for risk prediction. For tabular datasets of the kind used in this work, three families of models recur in the literature: regularised logistic regression, which offers interpretable coefficients and a strong baseline; random forests, an ensemble of decision trees that reduces variance through bagging [4]; and gradient-boosted trees such as XGBoost, which often deliver state-of-the-art accuracy on structured data through additive boosting of weak learners [5]. A recurring practical difficulty is class imbalance, because positive (diseased) cases are usually a minority. The Synthetic Minority Over-Sampling Technique (SMOTE) addresses this by generating synthetic minority examples along feature-space line segments rather than simply duplicating records [6]. Standard implementations of these methods are provided by the scikit-learn library [11], which underpins the present work.

### B. Explainable artificial intelligence

As predictive models grow more complex, explainability becomes essential in clinical contexts. Two model-agnostic techniques dominate. LIME approximates a model locally with an interpretable surrogate to explain individual predictions [7]. SHAP, based on Shapley values from cooperative game theory, attributes a prediction to its input features in an additive and theoretically consistent manner [8]. Because SHAP provides both local per-prediction attributions and globally consistent feature importance, it is well suited to a screening tool that must justify each individual result. The broader literature emphasises that the value of artificial intelligence in medicine depends not only on accuracy but on transparency, clinician trust and careful validation [9], [10].

### C. Class imbalance and evaluation metrics

Because diseased cases are usually a minority, the handling and measurement of class imbalance deserve particular attention. A classifier trained on imbalanced data can minimise its error simply by favouring the majority class, producing a high accuracy that conceals an inability to identify the very cases of interest [13]. For this reason, accuracy is widely regarded as an inadequate sole metric on imbalanced problems, and discriminative measures are preferred. The area under the receiver operating characteristic curve summarises performance across all decision thresholds, while the precision-recall curve is often still more informative when the positive class is rare, because it focuses on the minority class directly [14]. These insights directly shaped both the resampling strategy and the reporting choices adopted in this work, and they explain why a model with high accuracy was nonetheless judged unfit once its ROC-AUC was examined.

### D. Existing systems and the research gap

A large number of published systems report high accuracy for the prediction of an individual disease, and several public web demonstrations expose such models through a simple form. However, comparatively few combine multi-disease screening, per-prediction explanations, secure multi-role access and a deployable web interface within one coherent platform. Reported accuracies are also frequently presented without scrutiny of dataset quality, leakage or discriminative measures such as ROC-AUC, which can mask poor performance on imbalanced data.

Recent studies illustrate both the progress and the remaining gaps (Table I). Halder et al. built a Flask web application around a random-forest classifier for chronic kidney disease (CKD) trained on the 400-record UCI CKD dataset [28], but did not report explanations of individual predictions. Nguycharoen combined SHAP, prototypes, anchor rules and counterfactuals for CKD risk in cardiovascular patients using 491 outpatient records, but did not report a validation scheme or a deployed application [29]. Allani deployed an interactive Dash application for diabetes risk with SHAP and LIME explanations, trained on survey data from the CDC Behavioral Risk Factor Surveillance System [30], and Willson et al. combined SHAP with GPT-4o-generated narratives in an Android front end for diabetes risk, evaluated with functional tests and a 12-participant comprehension study [31]. Chen et al. developed an ensemble model for CKD progression with SHAP explanations and described an external cohort from three centres, but without a deployed interface [32]. Each of these studies addresses a single disease.

TABLE I. Comparison with Recent Related Studies

| Study | Disease | Data and validation | XAI | Web app |
|---|---|---|---|---|
| [28] 2024 | CKD | Split and k-fold CV; UCI, n=400 | None reported | Flask |
| [29] 2024 | CKD in CVD patients | Not specified; n=491 | SHAP, prototypes, anchors | No |

| Study | Disease | Data and validation | XAI | Web app |
|---|---|---|---|---|
| [30] 2025 | Diabetes | Stratified CV; BRFSS, about 250k | SHAP, LIME | Dash app |
| [31] 2026 | Diabetes | Functional tests; 12-user study | SHAP + LLM text | Android app |
| [32] 2025 | CKD course | CV; external cohort described | SHAP | No |
| This work | Diabetes, heart, CKD | Split-first CV; internal test only | SHAP + LLM guidance | React/Flask, RBAC |

MediHealth Predict does not claim higher predictive accuracy than these single-disease models. The gap it addresses is the combination of three disease models behind one role-controlled interface, per-prediction SHAP attributions with constrained language-model guidance, and an evaluation protocol that keeps model selection separate from the test set and reports a failed model openly. Relative to the studies above its limitations are also clear: unlike Chen et al. [32] it has no external validation, and unlike Willson et al. [31] it has no user study; both are identified as future work.

Three strands of the methodological literature shape the evaluation reported here. First, applying supervised feature selection to the full dataset before splitting leaks label information into the test partition and produces optimistic estimates, so selection must be treated as part of the model and refitted inside each training fold [21], [22]; more generally, preprocessing choices that touch the target are a recognised source of leakage [27]. Second, a probability produced by a classifier is not necessarily a calibrated risk, and calibration should be assessed separately from discrimination [23]. Third, post-hoc attributions such as SHAP describe the model rather than the underlying disease, and their use in high-stakes settings has been questioned [24], while efficient tree-based SHAP estimators make per-prediction attributions practical for ensembles [25]. Reporting guidance for clinical prediction models, such as TRIPOD, asks for the data sources, handling of missing data, validation approach and performance measures to be stated in full [26]; we follow it as far as the available data permit.

A further observation from the literature is that a predictive model only delivers value once it is embedded in software that real users can reach securely. Engineering concerns such as authentication, persistence, auditability and a maintainable code structure are therefore as important to a clinical decision-support tool as the underlying model, and established software-engineering practice informs the present design [12].

## III. Requirements

Requirements were elicited from the perspective of the principal stakeholders, namely patients seeking early awareness of their risk and clinical staff who review and act upon results. They were recorded as functional and non-functional requirements to guide design and to provide a basis for evaluation.

### *A. Functional requirements*

The core functional requirements are that the system shall allow users to register and authenticate; shall accept clinical and lifestyle inputs and return a risk estimate for each of the three target diseases; shall present an explanation identifying the factors that most influenced each result; shall generate plain-language lifestyle and medical guidance; shall store an assessment history that users can revisit; shall provide a constrained conversational assistant for general health questions; and shall provide administrators with the means to verify clinical staff and manage accounts.

### *B. Non-functional requirements*

The principal non-functional requirements concern security, usability and maintainability. Sensitive data shall be protected through hashed credentials and token-based authentication; access shall be restricted according to user role; the interface shall be usable by non-specialists and shall present risk in an immediately interpretable form; and the model-training process shall be reproducible so that results can be regenerated and audited.

## IV. System Design and Methodology

### *A. Architecture*

The platform follows a classic three-tier architecture, illustrated in Fig. 1: a React single-page presentation tier, a Flask REST API application tier that hosts authentication, prediction, explanation and administration, and a SQLAlchemy data tier portable from SQLite in development to PostgreSQL in production. The tiers communicate over JSON, and the trained models are loaded by the backend as serialised artifacts at start-up.

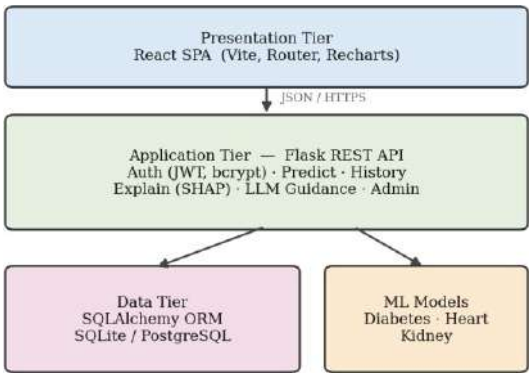



Fig. 1. Three-tier architecture of the MediHealth Predict platform.

### *B. Datasets and preprocessing*

Each model is trained on a separate, publicly available tabular dataset whose features reflect the indicators clinically associated with the disease in question: glucose-related measures and body-mass index for diabetes [18], blood-pressure and cholesterol-related measures for heart disease [19], and serum-chemistry and urinalysis indicators for chronic kidney disease [20]. Table III reports the size of each dataset after cleaning and the class balance. [Insert, for each dataset: exact

source and version/URL, licence, number of raw and cleaned records, positive/negative counts and the final feature list.] The diabetes and kidney datasets are positive-class minority problems, which motivates the resampling strategy justified in Section IV-D, while the heart dataset is comparatively balanced but small. Cleaning is deliberately conservative. Records with impossible or missing values are handled by [state rule: removal or imputation, and the number affected], exact duplicates are removed before splitting so that the same record cannot appear in both the training and test partitions, and any field that would not be available at the moment of prediction, or that trivially encodes the target, is excluded to prevent leakage. The stratified train-test split is then performed. Univariate analysis-of-variance (ANOVA) scoring, which uses the labels, is fitted on the training partition only (inside each cross-validation fold during model selection) and the retained feature set is then applied unchanged to the test partition, so that no test-set label information influences feature selection.

*C. Model selection rationale*

Three model families were selected as pipeline candidates rather than a single fixed algorithm, and each was included for a specific, literature-supported reason rather than by default. Regularised logistic regression was retained as an interpretable baseline: its coefficients map directly onto clinically meaningful odds ratios, it is inexpensive to train on the small heart-disease sample, and it establishes a lower bound against which the more flexible models must justify their added complexity. Random forests were included because bagged decision trees reduce the variance of a single tree, handle non-linear feature interactions without explicit engineering, and tend to remain robust when the training sample is small or noisy [4]. XGBoost was included because regularised gradient-boosted trees frequently match or exceed the accuracy of alternative model families on structured, medium-sized datasets, while built-in L1/L2 regularisation and shrinkage help limit overfitting [5]. Broader empirical benchmarking supports this choice: tree-based ensembles remain highly competitive with, and often outperform, deep-learning alternatives on tabular data of this size [16]. These are plausible priors rather than guarantees, so no family is assumed to be best: candidates are compared empirically on each dataset by stratified cross-validation on the training partition, using ROC-AUC as the primary criterion with PR-AUC and the cross-validated train-validation gap as secondary criteria. The test partition plays no part in this choice. Section VI-E reports the resulting comparison of all candidates, together with a majority-class baseline, on each dataset.

*D. Resampling strategy: justification of SMOTE*

The diabetes and kidney datasets are both positive-class minority problems, so the training pipeline must correct for imbalance before a model is fitted. SMOTE was chosen over the available alternatives for three reasons. First, compared with random oversampling, which duplicates existing minority records verbatim, SMOTE synthesises new examples along feature-space line segments between neighbouring minority instances, which densifies the minority region without creating exact duplicates that would otherwise encourage the model to memorise specific rows [6]. Second, compared with random under-sampling, which discards majority-class records, SMOTE preserves the full majority class, which matters because discarding data would be especially damaging for the small heart dataset and for the kidney dataset. Third, compared with adaptive variants such as ADASYN, which bias synthetic-sample generation toward minority instances that are hardest to classify, SMOTE gives more even coverage of the minority class and is less sensitive to label noise near the decision boundary, which matters for clinical features that are themselves measured with some error [17]. Class weighting is a resampling-free alternative; rather than asserting that it is inferior, we include it as an ablation configuration in Section VI-F so that its effect can be read directly from the results. SMOTE is applied to training data only: after the train-test split, and, during model selection, inside each training fold of the cross-validation (through a pipeline object that refits it per fold). Synthesising minority examples before splitting would let synthetic neighbours of a held-out record appear in the training data, inflating the reported performance. The validation and test folds are never resampled, so they retain the natural class distribution.

*E. Data model and access control*

Authentication uses JSON Web Tokens with bcrypt-hashed passwords, and the relational model defines six roles: patient, nurse, doctor, laboratory technician, hospital administrator and super administrator: together with an audit log that records sensitive actions. This role-based access control ensures that each class of user is granted only the privileges appropriate to its function, which matters because the platform handles sensitive health data.

*F. Prediction workflow*

A prediction proceeds through a well-defined sequence. The authenticated client submits the user's clinical and lifestyle inputs to the prediction endpoint. The backend validates the request, derives any model features that were not supplied directly, scales the feature vector with the scaler stored for the relevant disease, and obtains a probability from the corresponding model. It then computes the feature attributions, maps the probability to a risk band, requests natural-language guidance, persists the assessment, and raises a notification if the result is high risk. The composed response, containing the score, the band, the leading factors and the guidance, is returned to the client for display. Persisting each assessment in this way also enables the analytics view to plot how a user's risk changes over time.

*G. Development methodology*

The project was developed iteratively, with requirements refined across successive increments covering authentication, the prediction engine, the explanation layer and the administrative features, each validated against its requirements before the next was undertaken.

*H. Scalability and persistence*

The system was designed so that scale was not an afterthought: database access passes exclusively through the SQLAlchemy mapping layer, so the same code operates over PostgreSQL in production without modification, and the stateless, token-based API can be replicated behind a load balancer. The trained models are loaded once at start-up and shared across requests, keeping prediction latency independent of user count.

## V. Implementation

### A. Implementation overview

The implementation pairs a Python machine-learning ecosystem (scikit-learn and XGBoost) with a Flask REST API, organised into blueprints for authentication, prediction, explanation, chat, administration and notifications, and a React/Vite single-page front end. Beyond serving predictions, the backend generates natural-language guidance and drives a conversational assistant that is explicitly restricted from issuing diagnoses or prescriptions, raises notifications when a high-risk result is recorded, and exports assessment reports. The front end presents role-appropriate dashboards, a guided prediction flow, and analytics view in which a colour-coded risk gauge, ranked attribution bars and a trend chart make each result and its drivers immediately interpretable to non-specialists.

### B. Model-training pipeline

Each disease model is produced by a single reproducible pipeline of twelve stages, governed by a fixed random seed. The stages are: (1) loading the raw dataset; (2) converting the target to a binary label; (3) cleaning invalid and missing values; (4) removing duplicate records; (5) performing a stratified seventy-five to twenty-five train-test split, before any step that uses the labels; (6) selecting the most informative features by univariate ANOVA scoring, fitted on the training partition only; (7) applying SMOTE to the training partition only; (8) standardising the features with a scaler fitted on the training partition; (9) training and comparing the candidate models, namely logistic regression at two regularisation strengths, a regularised random forest and a regularised XGBoost classifier, with hyperparameters and the final candidate chosen by stratified [5]-fold cross-validation on the training partition (steps 6–8 being refitted inside every fold), using ROC-AUC as the primary criterion; (10) refitting the chosen candidate on the full training partition and calibrating its probabilities by [sigmoid/isotonic] calibration fitted within the training data; (11) evaluating the final model once on the untouched test partition; and (12) persisting the model, the scaler, the calibrator and the ordered feature list as artifacts. The earlier practice of ranking candidates by closeness to a fixed accuracy target and by train-test gap computed on the test partition has been removed, because it allowed the test set to influence model choice. Selecting features, resampling and tuning entirely within the training data, and removing leak-prone fields, are deliberate measures to prevent data leakage and over-optimistic estimates.

The use of four candidate configurations, rather than a single fixed algorithm, allows the pipeline to adapt to the characteristics of each disease dataset. Because the cross-validation criterion is discriminative performance rather than an accuracy target, the selection rule favours models that rank cases correctly, and the cross-validated train-validation gap is examined as a guard against overfitting. Because the same procedure, seed and evaluation protocol are applied to every disease, the three models are produced consistently and their results are directly comparable. The hyperparameter search space and the values finally selected for each disease are listed in [Table: search grids and selected values].

Reproducibility is treated as a first-class concern. Fixing the random seed across splitting, resampling and model initialisation ensures that a training run can be repeated to obtain identical artifacts, which is essential if results are ever to be audited or defended. On completion, the pipeline persists not only the model but also the fitted scaler, the calibrator and the ordered list of selected features, so that the exact transformation applied during training is reapplied at prediction time. Storing these artifacts together removes a common source of silent error, in which a model is served with features scaled or ordered differently from how it was trained. The library versions used and the training code are available at [repository URL / DOI]. Because the fixed seed fixes a single split, the variability due to the split is quantified separately by repeated stratified cross-validation and by bootstrap confidence intervals on the test set (Section VI-G).

### C. Prediction and feature derivation

In practice a user rarely supplies every variable that a model expects. The prediction service therefore derives missing model features from the clinical inputs that are available, [state the exact derivation rules, e.g. defaults or population medians, per feature], applies the stored scaler, and produces a risk score. Whether this score is a well-calibrated probability is assessed empirically in Section VI-G rather than assumed, and the effect of derived features on the output is quantified there. Each score is mapped to a colour-coded band, green for low, amber for moderate and red for high risk, so that results are immediately interpretable by non-specialists. The band thresholds are [state values] and have not yet been validated against clinical outcomes.

### D. Explanation and guidance layer

For every prediction the system computes SHAP feature attributions using [the explainer used, e.g. TreeExplainer or LinearExplainer] and surfaces the three factors that contributed most to the result, presented to the user as labelled bars. This per-prediction transparency is intended to allow a clinician to sanity-check the model's reasoning; Section VI-H reports what has and has not been verified about it. In addition, a large language model ([model name and version]) generates structured, plain-language guidance covering diet, activity and general medical advice, and a constrained conversational assistant answers general health questions. The assistant is explicitly restricted from issuing diagnoses, so that the tool supports rather than replaces professional judgement. The kidney model is disabled in the deployed prediction endpoint and the interface reports it as unavailable, so that its output cannot be mistaken for a validated result [confirm this is implemented in the code].

## VI. Results and Evaluation

Each model was evaluated once on its untouched test partition using accuracy, precision, recall, F1-score, the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC). ROC-AUC is reported because, unlike accuracy, it does not depend on a single decision threshold or directly on class prevalence; PR-AUC is reported alongside it because ROC-AUC can look optimistic when the positive class is rare [14]. Intervals are 95 percent bootstrap confidence intervals over the test partition. Table II summarises the headline results, Table III describes the

datasets on which the models were trained, and Table IV compares all candidates and a majority-class baseline.

TABLE II. PERFORMANCE OF THE THREE DISEASE MODELS ON HELD-OUT TEST DATA

| Disease | Best model | Acc. % | F1 % | AUC | Gap |
|---|---|---|---|---|---|
| Diabetes | Logistic reg. | 86.1 | 86.0 | 0.93 | -0.1 |
| Heart | Logistic reg. | 81.6 | 81.5 | 0.89 | +3.8 |
| Kidney | Random forest | 80.0 | 71.1 | 0.51 | +5.3 |

[a] Gap denotes the train minus test accuracy difference; small values indicate little overfitting. [All values must be regenerated with the corrected pipeline (split before ANOVA selection; CV-based model choice) and reported with 95% CIs, PR-AUC and Brier score.]

TABLE III. CHARACTERISTICS OF THE TRAINING DATASETS

| Disease | Train size | Test size | Note |
|---|---|---|---|
| Diabetes | 75,000 | 25,000 | Positive-class minority; SMOTE on train only |
| Heart | 226 | 76 | Small after de-duplication |
| Kidney | 24,648 [reconcile] | 5,135 [reconcile] | Imbalanced classes |

Sizes refer to records after cleaning and the stratified split. The kidney partition (24,648/5,135, i.e. about 83/17) does not correspond to the stated 75/25 split. [Reconcile: either correct the figures or state the split actually used; add the positive/negative counts for every dataset.]

### *A. Diabetes and heart models*

The diabetes model was the strongest performer. Trained on a large dataset of one hundred thousand records, it achieved [86.1] percent accuracy, an F1-score of [86.0] percent and a ROC-AUC of [0.93] on the test partition, with a negligible train-test gap, which is consistent with good discrimination and little overfitting on this dataset. The heart model achieved [81.6] percent accuracy and a ROC-AUC of [0.89]. However, after removing duplicate records only around three hundred unique samples remained (226 training and 76 test records), so the test partition contains few cases, the confidence interval around these figures is wide [report it], and the results, while encouraging, rest on a small sample and a modest positive train-test gap; they should be regarded as preliminary.

### *B. Kidney model and the limits of accuracy*

The kidney model illustrates why accuracy alone is an unreliable indicator. It reported 80 percent accuracy, which appears acceptable, yet its ROC-AUC was only 0.51, essentially equivalent to random guessing, and its accuracy should be read against the majority-class baseline of [state %] given in Table IV. This pattern is characteristic of a model that predicts mostly the majority class: it can attain high accuracy while failing to discriminate between patients. Class imbalance is the most likely explanation, but we cannot exclude others, such as weak signal in the available features, label noise or the way the dataset was assembled, and the confusion matrix [add] is needed to confirm the behaviour. The model is therefore reported as a negative result and is not fit for clinical use; it is withheld from the deployed platform. Reporting this outcome transparently is, in our view, more valuable than presenting a misleadingly high accuracy figure, and it directly motivates the data-quality work proposed in Section VIII.

### *C. Comparative analysis*

Fig. 2 contrasts accuracy with ROC-AUC across the three models and makes the central finding visually plain. For the diabetes and heart models the two measures agree and remain well above the chance line, which is consistent with genuine discriminative ability, subject to the confidence intervals reported above. For the kidney model, however, accuracy and ROC-AUC diverge sharply: accuracy remains high while ROC-AUC falls to the chance level. This divergence is the clearest single indication that the kidney model's apparent performance reflects the class distribution rather than evidence of clinical usefulness.

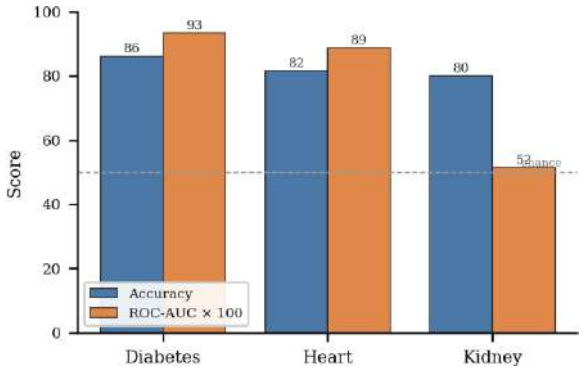



Fig. 1. Accuracy compared with ROC-AUC for the three models; the dashed line marks chance-level discrimination.

### *D. Precision, recall and risk bands*

Headline accuracy and F1-score do not by themselves reveal how a model behaves at the decision boundary, which matters for a screening tool, where missing a true case (low recall) and raising false alarms (low precision) have different costs. Precision, recall and PR-AUC for each model, together with the confusion matrices at the operating threshold, are reported in [Table II extension / Appendix]. For the kidney model, the relatively high nominal recall is consistent with over-prediction of a single class, which is precisely the failure mode that the precision-recall view is designed to expose. The mapping from probability to a three-level risk band was chosen so that borderline cases fall into the moderate band and are encouraged to seek confirmation, rather than being given false reassurance or undue alarm; refining these thresholds against clinical outcomes is a natural extension of the present work.

### *E. Baseline comparison*

To verify that the selected model for each disease is the best-justified of the candidates evaluated in the training pipeline (Section V-B), rather than a product of the selection rule, Table IV reports all four candidates and a majority-class baseline side by side for every disease. The test-set figures are reported for transparency only; the choice of candidate was made on cross-validation within the training partition.

[Complete Table IV with the measured accuracy, F1-score and ROC-AUC of all four candidates and the majority-class baseline for each of the three diseases, from the corrected pipeline. Then state which candidate each disease would have selected under a pure accuracy-maximising rule versus the cross-validated ROC-AUC rule actually used, and whether this changes any conclusion of the paper.]

TABLE IV. BASELINE COMPARISON OF ALL CANDIDATE MODELS ON HELD-OUT TEST DATA

| Disease | Candidate | CV AUC | Test Acc. % | Test F1 % | Test AUC |
|---|---|---|---|---|---|
| Diabetes | Majority class | N/R | N/R | N/R | N/R |
| | LogReg (C1) | 0.9431 | 88.74 % | 89% | 0.9339 |
| | LogReg (C2) | N/R | N/R | N/R | N/R |
| | Random forest | 0.9456 | 91.98 % | 92% | 0.9429 |
| | XGBoost | 0.9444 | 91.99 % | 92% | 0.9423 |
| Heart | Majority class | N/R | N/R | N/R | N/R |
| | LogReg (C1) | 0.9278 | 83.61 % | 84.85 % | 0.9048 |
| | LogReg (C2) | 0.9254 | N/R | N/R | N/R |
| | Random forest | 0.9075 | N/R | N/R | N/R |
| | XGBoost | N/A | N/A | N/A | N/A |
| Kidney | Majority class | N/R | N/R | N/R | N/R |
| | LogReg (C1) | N/A | N/A | N/A | N/A |
| | LogReg (C2) | N/A | N/A | N/A | N/A |
| | Random forest | N/R | 80.01 % | 71.13 % | N/R |
| | XGBoost | N/R | 77.41 % | N/R | N/R |

CV AUC is the cross-validated ROC-AUC on the training partition used for selection. Test columns are reported once, after selection. [Mark the candidate chosen by the CV rule and the one a pure accuracy rule would choose.]

### *F. Ablation study*

The ablation study is designed to investigate the individual contributions of class imbalance handling and feature selection techniques to the overall performance of the proposed multi-disease prediction system. By comparing different configurations, the study aims to identify how these components influence predictive performance across diabetes, heart disease, and kidney disease classification tasks.

### *G. Calibration, uncertainty and missing-feature sensitivity*

Because the platform presents a score as a risk to the user, calibration was assessed in addition to discrimination. For each retained model we report the Brier score and a reliability curve on the test partition [add values and figure] [23]. Uncertainty in every headline metric is expressed as a 95 percent bootstrap confidence interval over the test partition, and, to show sensitivity to the particular split, as the spread across [10] repeated stratified splits. Finally, because users rarely supply every input, we quantify the effect of the derived-feature mechanism by randomly replacing an increasing fraction of the measured inputs with the derived defaults and re-scoring the test partition; the resulting change in ROC-AUC and in the proportion of cases whose risk band changes is reported in [Fig. 3 / Table]. SMOTE changes the class prior seen during training, so uncalibrated outputs are expected to over-estimate risk for minority-class diseases; calibration is therefore fitted within the training data as described in Section V-B.

### *H. Validation of the explanation layer*

The explanation component is evaluated at three levels, and its limits are stated. First, a sanity check: the globally most influential features by mean absolute SHAP value are compared with the clinically established risk factors for each disease [list them and report agreement]. Second, stability: the rank correlation of attributions across bootstrap-refitted models and across random seeds is computed [report Spearman ρ], since unstable attributions would undermine their use as an explanation. Third, SHAP explains the behaviour of the model and not the causes of disease [24]; the attributions should therefore not be read as causal. No user study with patients or clinicians has been carried out, so the claim that the attributions help clinicians trust or act on a result is a design motivation that remains untested, and a usability and comprehension study is identified as future work. The language-model guidance has likewise not been evaluated for factual accuracy beyond the constraints described in Section VII.

### *I. Functional testing*

The core workflows behaved as specified, confirming that the predictive components are correctly integrated within a working multi-user system. Assessed against the requirements of Section III, the platform meets its functional goals for registration, prediction, explanation, guidance, history and administration, and its non-functional goals for role-based security, interpretable presentation of risk and a reproducible training process. The principal shortfall is not in the software but in the clinical validity of one model, which the evaluation reports openly rather than concealing.

### *J. Threats to validity*

Several threats to the validity of these results should be acknowledged. The metrics are computed on held-out partitions of the same datasets used for training, so they measure internal rather than external validity; performance on a different population could differ, and no external dataset has yet been used. An earlier version of the pipeline applied supervised feature selection before the split and selected candidates using test-partition behaviour; both have been removed, but the test partition was seen during development, so a further untouched dataset remains the only fully clean confirmation. The heart result, resting on a few hundred unique records, is especially sensitive to the particular split, and a different random partition could shift its figures noticeably. The diabetes dataset in particular may be partly synthetic or curated, so its very high performance may not transfer to routine clinical data [confirm provenance]. SMOTE generates synthetic training records, which may not reflect real patient variability. The practice of deriving absent features from available inputs, while necessary for usability, means that a deployed prediction may be based on estimated rather than measured values; its effect is quantified in Section VI-G but only on the test data. Because the datasets were

assembled for research rather than drawn from the target clinical setting, label definitions and measurement conventions may not match local practice, and no Sri Lankan data are included. Finally, the explanation and guidance layers have not been evaluated with users. These threats motivate the external validation proposed below and caution against over-interpreting the headline numbers.

## VII. Discussion, Limitations and Ethics

The evaluation supports a cautious observation: the clinical validity of a screening tool appears to be governed at least as much by the quality of the underlying data as by the choice of algorithm. The diabetes model performed well on a large dataset with a reasonable number of positive cases, whereas the kidney and heart models were constrained by imbalance and limited sample size respectively; we did not isolate these causes experimentally, so this remains an interpretation rather than a demonstrated result. The contribution of the work is one of integration and careful evaluation rather than algorithmic novelty. Several limitations follow. The system is a screening aid and not a certified medical device, and it has not undergone prospective clinical validation. The datasets used for training may not reflect the demographic characteristics of the population in which the tool would be deployed, so external validation is required before any clinical reliance. Finally, the derivation of missing model features from available inputs, while practical, introduces approximation whose effect on risk bands should be monitored in deployment.

Ethical considerations were central to the design. Because the platform handles sensitive health data, access is restricted by role, credentials are hashed and sensitive actions are logged. The conversational assistant is constrained so that it does not present itself as a diagnostic authority, reducing the risk that a user acts on automated output without professional advice. These measures reflect the principle that an artificial intelligence tool in healthcare must be transparent, accountable and subordinate to clinical judgement.

Questions of deployment readiness and data governance follow directly. A tool of this kind would, in most jurisdictions, need to satisfy data-protection obligations covering consent, storage and the right to erasure, and any move from a screening aid towards a clinical claim would invite scrutiny under medical-device regulation. The architecture anticipates some of these requirements through its audit log and role separation, but formal compliance, together with bias auditing across demographic subgroups, would be a prerequisite for real-world use. Reporting an unfit model openly, as done here, is itself part of responsible governance, since it prevents a weak model from being mistaken for a validated one.

The use of a large language model for guidance warrants its own caution. While it produces fluent, accessible advice, such a model can in principle generate plausible but incorrect statements. The design mitigates this by confining the model to general lifestyle and informational guidance, by grounding its output in the structured prediction rather than free diagnosis, and by explicitly preventing it from issuing diagnoses or prescriptions. The deterministic SHAP attributions, not the language model, remain the authoritative explanation of why a given risk was estimated; the language model serves only to communicate, not to decide.

## VIII. Conclusion and Future Work

This paper presented MediHealth Predict, an explainable, multi-disease risk-prediction web platform that combines a secure multi-role architecture, a reproducible and leakage-safe model-training pipeline, SHAP-based explanations and large-language-model guidance. The diabetes model showed strong discrimination on internal held-out data, the heart model showed encouraging but preliminary results, and the kidney model was transparently reported as a negative result and withheld from use. The work demonstrates that integrated, explainable screening tools are technically achievable, while underlining that data quality and external validation are decisive. Future work will focus on external validation on independent cohorts, sourcing a higher-quality, balanced kidney dataset, obtaining a larger heart-disease sample, a clinician and user study of the explanations, prospective validation with clinical partners, extending the platform to additional conditions, and investigating privacy-preserving training so that models can improve without centralising sensitive patient data.

Beyond improving the individual models, future work will consider how the platform could fit into a clinical pathway. This includes integration with electronic health records so that inputs are populated from measured rather than self-reported data, the addition of continuous monitoring so that a deployed model is re-evaluated as new outcomes accrue, and a structured bias audit across age, sex and other demographic groups to ensure that performance is equitable. Each of these steps moves the system further from a proof of concept and closer to a tool that a clinician could responsibly rely upon, while preserving the explainability and transparency that distinguish the present design.

## References


[1] International Diabetes Federation, IDF Diabetes Atlas, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025.

[2] Jha et al., “The state of global kidney health,” Nature Reviews Nephrology, vol. 20, pp. 473–485, 2024.

[3] World Health Organization, “Cardiovascular diseases (CVDs),” Fact sheet, WHO, Geneva, Switzerland, 2021.

[4] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[8] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.

[9] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” Nature Medicine, vol. 25, pp. 44–56, 2019.

[10] L. Beam and I. S. Kohane, “Big data and machine learning in health care,” JAMA, vol. 319, no. 13, pp. 1317–1318, 2018.

[11] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[12] Sommerville, Software Engineering, 10th ed. Harlow, U.K.: Pearson,

2016.
[13] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[14] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLoS ONE, vol. 10, no. 3, e0118432, 2015.
[15] Ambroise and G. J. McLachlan, “Selection bias in gene extraction on the basis of microarray gene-expression data,” Proc. Natl. Acad. Sci. USA, vol. 99, no. 10, pp. 6562–6566, 2002.
[16] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” BMC Bioinformatics, vol. 7, art. 91, 2006.
[17] Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proc. 22nd Int. Conf. Machine Learning (ICML), 2005, pp. 625–632.
[18] Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.
[19] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, vol. 2, pp. 56–67, 2020.

# Systematic Method To Manage Patient Flow And Queue For Echocardiography Units In Sri Lankan Public Hospitals

R A Manusha Prabath Ranaweera
*Faculty of Computing, MIS Batch 22.2*
*NSBM Green University*
*Homagama, Sri Lanka*
ampranaweera@students.nsbm.ac.lk

Hirushi Dilpriya
*Senior Lecturer, Faculty of Computing*
*NSBM Green University Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Echocardiography (ECO) units in Sri Lankan public hospitals largely rely on manual, paper-based registration and queue management, which leads to long and unpredictable waiting times, uncertainty about queue position, limited communication of delays, and reactive resource planning. This paper presents the design and prototype implementation of an integrated queue-management and wait-time-prediction system for a public-hospital ECO unit. The system provides three coordinated interfaces (a patient mobile application, a staff web portal and an administrator web portal) and supports digital registration with automatic ticket generation, real-time queue tracking, staff-controlled priority handling of urgent cardiac cases, and a machine-learning (ML) prediction service (Random Forest and XGBoost) for patient waiting time and hourly patient inflow. Design Science Research with Agile prototyping was followed, using requirements gathered through interviews with ECO unit staff, Figma wireframes, and twenty interface screens. Because no historical patient-level queue data were available, the ML service has only been functionally tested; no predictive-accuracy results are reported, and an evaluation protocol for real operational data is given instead. The system has not been validated in a live clinical deployment, so improvements in waiting time and resource planning are expected benefits rather than measured outcomes; a pilot evaluation is planned. Within the literature reviewed, we found no existing system that combines these features for an ECO unit in Sri Lanka. Limitations and future work, including reinforcement learning (RL)-based adaptive prioritization and non-smartphone access channels, are discussed.**

***Keywords—Queue management, Machine learning, Echocardiography, Wait-time prediction, Healthcare informatics***

## I. Introduction

Optimizing patient flow is a persistent challenge in public services, especially in diagnostic units where equipment and skilled staff time are limited [18]. Sri Lanka's public healthcare is free and heavily utilized, and outpatient departments (OPDs) often depend on manual, paper-based registration, with studies documenting long and variable OPD waiting times in Sri Lanka [1], [2] and elsewhere [3]. The ECO unit is particularly constrained by a limited number of sonographers and machines and by examinations of variable duration [28], [29]. International research has begun to apply queueing theory and machine learning (ML) to patient flow in imaging units [11]–[17], [21], but we found no reported application to Sri Lankan public-hospital ECO units, where registration and prioritization remain manual.

Based on a preliminary investigation (observation of the manual process and stakeholder interviews) and the Sri Lankan OPD literature [1], [2], the ECO unit faces interrelated problems: manual registration that can be lost or duplicated; long, unpredictable queues without reliable wait estimates; no formal queue-number system; difficulty identifying and prioritizing urgent patients; poor communication of delays to patients; no real-time visibility of queue length and station status for management; and reactive rather than evidence-based staffing. Each problem is

documented separately in the literature, but we found no Sri Lankan or international system that addresses them together in a single diagnostic-imaging setting such as ECO.

National platforms such as the Hospital Health Information Management System (HHIMS) and eChannelling [6], [9], [10] address hospital-wide information management and appointment booking, but none offers unit-level real-time queue prediction and priority-based management for a high-demand diagnostic service. Meanwhile, ML-based waiting-time forecasting with Random Forest, XGBoost and other models is well established [11]–[17], [26], [27].

This paper presents the design and prototype implementation of an ECO queue-management and prediction system for a Sri Lankan public hospital. Its contributions are: (i) an integrated three-interface architecture (patient, staff, administrator) for registration, ticketing, real-time tracking and priority handling; (ii) a staff-controlled priority mechanism for urgent cardiac cases; (iii) an ML prediction service (Random Forest and XGBoost) for waiting time and hourly inflow, with a defined evaluation protocol (accuracy results await real operational data); and (iv) a feature comparison with related systems and an explicit statement of what has and has not been validated. The ML component only predicts waiting time and queue volume from operational data; it does not perform diagnosis or image interpretation and is intended to complement HHIMS and eChannelling. Section II reviews related work, Section III describes the design and methodology, Section IV presents the implementation results, and Section V discusses implications, limitations and future work.

## II. Related Work

### A. Queueing Theory and Simulation

Fomundam and Herrmann [18] survey queueing theory in healthcare, including standard arrival and service models (e.g., M/M/1, M/M/c). A study of a Malaysian public hospital

[19] reported that simple queue control and basic digitization can reduce waiting times without additional staff or equipment, a setting similar to Sri Lanka's. First-come-first-served and static-priority queues can leave low-priority patients waiting indefinitely; an accumulating priority queue with a finite horizon [20] addresses this and is relevant to prioritizing urgent cardiac referrals without starving routine echocardiograms. The methodologically closest study is an ML-enhanced queueing-network simulation of a hospital ultrasound department [21], which uses ML surrogate models of arrival and service times inside a discrete-event simulation. Outside Sri Lanka, a simulation study of OPD patient flow

[3] estimated waiting times and resource utilization for each OPD section.

### B. Sri Lankan Digital Health and OPD Studies

National documents such as the Sri Lanka Digital Health Blueprint [8] and the National Digital Health Guidelines and Standards v2.0 [9] set out the interoperability and governance expectations for new health systems, and a national review documents two decades of digitization [10]. Empirically, Dilrukshi et al. [1] used simulation to analyze bottlenecks and reduce OPD waiting time at the National Hospital of Sri Lanka. OPD-PMMAS [2] digitizes registration and sends SMS status alerts, while computer-vision methods count patients [4] and recognize clinic signage [5]. eChannelling [6] and Doc990 [7] are booking and mobile-access channels rather than live, in-hospital predictive queue tools. None of these studies targets ECO or includes ML-based prediction.

### C. Machine Learning for Wait-Time Prediction

Supervised ML is widely used to forecast outpatient waiting time. Tree-based ensembles outperformed linear regression in a tertiary children's hospital [11], and related work [12]–[17] addresses feature engineering, appointment delay, dissatisfaction-aware and interpretable prediction. Guo et al. [15] predicted waiting times for laboratory and radiology tasks, including echocardiography, in a pediatric hospital and found queue-related predictors, especially the number of queuing patients, to be the most influential, which informs our rolling queue-length feature. Across these studies, tree-based ensembles generally outperform simpler baselines on structured hospital-operations data, although a simple linear model performed acceptably for some tasks, including echocardiography, in [15]. We therefore compare Random Forest [26], which is robust on noisy, mixed-type data, with XGBoost [27], which uses second-order gradients and regularization and is a standard choice for structured prediction.

### D. Priority Queueing,

Mobile Systems and Echo WorkflowBeyond prediction, some studies actively manage queues. A reinforcement learning (RL) approach to priority-queue optimization [22] outperformed static policies; MediQueue

[23] combines queue prediction with AI-based optimization; and Kavitha et al. [24] combine classical queueing metrics with AI-based optimization. An early mobile queue-tracking system [25] established the pattern of remote ticketing, push notifications and real-time position tracking used in our patient app. Echocardiography workflow studies [28], [29] address digital image management, measurement automation and report turnaround time rather than patient queueing, while demand forecasting for healthcare resources is reported in [30]. The gaps identified are summarized in Table I together with the corresponding system features.

TABLE I
. Research Gaps and Corresponding System Features

| Gap | Description | System feature |
|---|---|---|
| G1 | No Sri Lankan system combines digital registration with ML wait-time prediction [1]–[17] | Registration + ML wait-time prediction |
| G2 | Priority-queue/RL methods [20], [22]–[24] not applied to ECO | Staff-assigned priority queue |
| G3 | Echo workflow studies [28], [29] address imaging/reporting, not patient queueing | Real-time queue-tracking dashboard |
| G4 | No mobile smart queue [25] + ML back end [21] for imaging | Patient mobile app + ML service |
| G5 | Demand forecasting [30] not applied to ECO volumes | Queue-length/inflow forecasting module |

## III. System Design and Methodology

### A. Research Design

The study followed Design Science Research (DSR), since the main contribution is a software artefact that must be purposefully designed and evaluated: identify the problem, define solution objectives, design and build the artefact, demonstrate it, evaluate it, and communicate the results. A mixed-methods approach was used: requirements elicitation through semi-structured interviews with ECO unit staff at the General Hospital (one doctor and four staff nurses) and direct observation of the existing manual process; prototype walkthroughs with ECO staff from different shifts and with patients; and an ML strand for the prediction service.

### B. System Architecture and Requirements

An Agile process was used: requirements gathering, wireframing, prototyping, sprint-based implementation, testing and refinement from stakeholder feedback, one component at a time (patient journey, staff screens, administrator screens). As shown in Fig. 1, the Patient Mobile App (Flutter), Staff Web Portal (React.js) and Administrator Web Portal (React.js) all use one Node.js/Express REST API, which handles login, ticket issuing, notifications and business logic and stores data in a MySQL database. For inference, the API calls a separate ML Prediction Service, developed in Python with scikit-learn and XGBoost, which can be retrained and redeployed without

changing client code. The core functional requirements, derived from the operational problems, Table I and stakeholder interviews, are given in Table II.

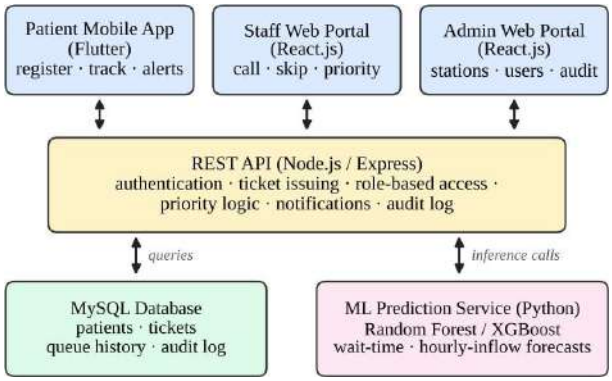



Fig. 1. Overall system architecture: three client interfaces, central REST API, MySQL database and independent ML prediction service.

TABLE II
. CORE FUNCTIONAL REQUIREMENTS AND PRIMARY INTERFACES

| Req. | Functional requirement | Interface |
|---|---|---|
| FR1 | Digitally join a queue and receive a ticket | Patient App |
| FR2 | Live view of queue position and visit stage | Patient App |
| FR3 | Call, skip or complete tickets | Staff Portal |
| FR4 | Flag and prioritize urgent patients | Staff Portal |
| FR5 | Predict patient inflow, peak hours, wait time | Staff Portal |
| FR6 | Search, view and register patient records | Staff Portal |
| FR7 | Generate operational performance reports | Staff/Admin |
| FR8 | Unit-wide overview of stations and alerts | Admin Portal |
| FR9 | Manage rooms/stations and staff accounts | Admin Portal |
| FR10 | Maintain a user-attributed audit log | Admin Portal |

### *C. Identification and Prioritization of Urgent Cases*

At present the ECO unit has no formal urgency-scoring system: staff use clinical judgement and send urgent patients ahead of the queue. The proposed system therefore does not

define clinical criteria or infer urgency automatically; it digitizes the existing staff decision. Priority is deliberately assigned by staff only and is not self-declared in the patient app, because patients who could flag themselves as urgent would be likely to all do so, which would defeat the purpose of prioritization. When registering a patient or reviewing a referral, authorized staff set the ticket's priority category (Routine or Urgent) in the Staff Portal (FR4). Urgent tickets are placed ahead of all waiting Routine tickets, and tickets within each category are served first-come-first-served. Each priority assignment or change is recorded in the audit log with the staff ID and timestamp (FR10), and patients' queue positions are updated in real time (FR2). The current version has no ageing mechanism, so a continuous stream of urgent arrivals could delay routine patients. Standardized urgency criteria agreed with cardiologists, accumulating-priority policies [20] and RL-based adaptive prioritization [22] are future work.

### *D. Data Sources and ML Prediction Service*

Requirements and operational parameters were obtained directly from ECO unit staff at the General Hospital through interviews and observation. No historical patient-level queue records (arrival and service timestamps) were available in digital form. The ML prediction service is designed to use arrival timestamp, referral source, priority category, assigned doctor/room, service start/end time and daily volume, which the system records during registration and queue handling. Only preliminary functional tests of the service were carried out, using sample records and scenarios based on conditions reported by staff: three to four staff on duty at a time, typical patient arrival volumes, how patients are entered into the system, and how urgent patients are prioritized. These tests checked that the prototype and ML service run end to end; they are not an evaluation of predictive accuracy, and no performance metrics are reported.

The planned model development (Table III) follows [12],

[14]: data cleaning, numerical encoding of categorical attributes, and derived features such as day of week, hour of day and rolling queue length. Targets are waiting time (minutes, from arrival to service start) and hourly patient arrivals. Random Forest [26] and XGBoost [27] will be compared on MAE, RMSE and $R^2$, and feature importances will be retained so that administrators can see which factors most influence predicted waiting time [14].

TABLE III
. PLANNED ML EVALUATION PROTOCOL (NOT YET EXECUTED)

| Item | Planned protocol |
|---|---|
| Data | Real timestamps collected during a multi-month pilot in the ECO unit; synthetic data, if used, for training only |
| Targets | Waiting time (min); hourly patient arrivals |
| Split | Chronological 80:20 train/test split; test set of real records only |
| Tuning | k-fold cross-validation on the training set |
| Models | Random Forest [26] and XGBoost [27], compared with linear-regression and historical-mean baselines |
| Metrics | MAE, RMSE, $R^2$; feature importance |
| Selection | Lowest cross-validated RMSE |

### *E. Security, Privacy and Ethics*

The system processes personal data (NIC and phone numbers, visit records) covered by Sri Lanka's Personal Data Protection Act, No. 9 of 2022 [31]. The prototype has so far been run only in a test environment without real patient data. The security design comprises: (i) role-based, need-to-know access separating patient, staff and administrator roles, with

patients seeing only their own ticket; (ii) staff authentication by hospital ID/password, with optional Single Sign-On; (iii) a user-attributed audit log of access and changes; (iv) data minimization, as the patient app requires only NIC and phone number, matched against the hospital's existing patient database; and (v) de-identification of any data used for ML development, with access limited to the researchers. Encrypted transport (HTTPS/TLS), salted password hashing, encryption of stored data and a PDPA compliance review are required before any deployment with real patient data; they are not claimed as implemented in the current prototype. Data collection at this stage was limited to interviews and observation with ECO unit

staff; no patient records or identifiable patient data were collected. Formal ethical approval will be obtained before any pilot with real patient data.

### *F. Testing Status and Plan*

Functional testing of the clients, API, ML service and database was carried out in a test environment. Integration and response-time testing, usability testing with representative staff and patients (scenario-based tasks and the System Usability Scale, SUS) and user-acceptance testing with hospital staff remain planned. Usability testing has not yet been completed (Section V-C).

## IV. Results

This section reports the implemented interfaces for the General Hospital ECO unit and the status of the ML evaluation. The figures show prototype screens populated with sample data. Owing to page limits, the Administrator Web Portal (station and staff-account management, audit log, settings) is described only in the underlying thesis; it follows the same architecture.

### *A. Patient Mobile Application*

Patients log in with their registered phone number and NIC (Fig. 2a). The design reflects the digital literacy of the patient population by requiring only these two identification fields and a single clear call to action. The home dashboard (Fig. 2b) shows the assigned queue number and the ML-predicted waiting time, addressing the need to communicate waiting status identified in Section I. The Live Queue screen (Fig. 2c) shows staged, colour-coded progress with real-time position updates and an alert so that patients can leave home when their turn is near rather than waiting at the hospital.

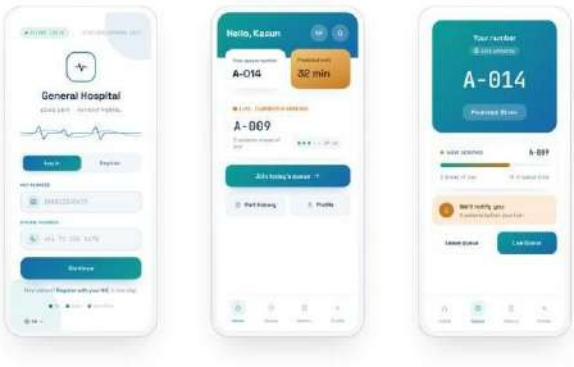



Fig. 2. Patient app screens (sample data): (a) secure login, (b) home dashboard with queue number and predicted wait, (c) live queue tracking.

### *B. Staff Web Portal*

Authorized staff sign in with a staff ID and password, optionally via Single Sign-On (Fig. 3a), and a persistent notice reminds them that the data are patient-identifiable and

for authorized use only. A staff profile dashboard (Fig. 3b) shows individual shift indicators (patients served, average time per patient, hours worked) and an auditable activity feed. The main operational dashboard (Fig. 4) shows the current queue length, patients served that day, the average wait, and an hourly ML-predicted patient-inflow chart that highlights the predicted peak hour, so that staff can plan resources proactively rather than reactively. Together, the patient and staff screens form a patient-side transparency loop and a staff-side operational control loop.

Fig. 3. Staff portal screens (sample data): (a) authentication, (b) staff profile with shift metrics.

Fig. 4. Staff dashboard (sample data) with queue overview and ML-predicted hourly patient inflow.

### *C. ML Component Status*

Predictive-accuracy results (MAE, RMSE, $R^2$) for Random Forest and XGBoost are not reported in this paper. No historical patient-level queue data from the ECO unit were available, and the models have only been functionally tested (Section III-D). The predicted waiting times and inflow values shown in Figs. 2–4 are illustrative sample values, not measured model output. Table III defines the protocol that will be applied once real operational data are collected during the pilot. Preliminary functional tests with sample records confirmed that the prototype can send registration data to the ML service and display the returned prediction values; these outputs were not evaluated for accuracy.

### *D. Stakeholder Feedback and Usability Status*

Prototype walkthroughs with ECO unit staff from different shifts (three to four staff on duty at a time, rotating between shifts) and with patients attending the unit guided iterative refinement of the screens. The number of patient participants was not formally recorded. These were formative sessions, not a formal usability evaluation. No SUS scores are reported, and no claims about usability are made.

## V. Discussion and Conclusion

### *A. Achievement of Research Objectives*

The prototype addresses gaps G1, G3 and G4 of Table I at the design-and-implementation level (registration with ML prediction, real-time tracking, and mobile–ML integration). G2 and G5 are only partially addressed: priority handling is

implemented as staff-assigned priority, with RL-based adaptive prioritization left for future work, and the

forecasting module needs validation on real, multi-month ECO attendance data. The separation of patient, staff and administrator concerns reflects the requirements in Table II.

### *B. Comparison with Existing Systems*

Table IV compares the proposed system with related systems. Ratings of existing systems are based only on their published descriptions and those of the proposed system denote implemented features, not clinical validation. To the best of our knowledge, within the literature reviewed, no system combines all four features for an ECO unit. A quantitative comparison of waiting time against existing systems is not yet possible: they are not available for benchmarking in an ECO setting, and the proposed system has not been deployed. The planned pilot (Section V-C) will measure baseline waiting times under the manual process, from registration and service-start timestamps, and compare them with the pilot period.

TABLE IV
. FEATURE COMPARISON WITH RELATED SYSTEMS

| System | Real-time | ML predict. | Priority | Imaging-specific |
|---|---|---|---|---|
| eChannelling [6] | No | No | No | No |
| OPD-PMMAS [2] | Partial | No | No | No |
| MediQueue [23] | Yes | Yes | Partial | No |
| Ultrasound sim. [21] | Sim. only | Yes | No | Yes |
| Proposed system | Yes* | Yes* | Yes* | Yes* |

Yes = feature described as available in the source (or implemented in our prototype*); Partial = limited form (e.g., SMS status alerts only, or priority handling only in part); Sim. only = shown only in simulation; No = not reported. *Implemented in the prototype; not clinically validated.

### *C. Limitations and Future Work*

The system is a design and prototype implementation, not a live, multi-month clinical deployment; therefore, reductions in waiting time, improved queue management and better resource planning are expected benefits that have not been measured. The ML models have not been evaluated, because no real historical ECO queue data were available, and predictive accuracy must be verified on real attendance data. The study covers one ECO unit, and findings may not generalize to hospitals with different case mix or workflows. The patient app depends on smartphone ownership, which is not universal among public-hospital patients [7]. No formal, statistically sound usability assessment has been completed, and the priority mechanism lacks ageing. Priority is based on staff judgement, as the unit has no standardized urgency criteria. Several safeguards in Section III-E (encryption, hashing, PDPA review) must be implemented before real patient data are used.

Future work will include a time-bound pilot with formal usability (SUS) and predictive-accuracy evaluation; training and evaluating the ML models on real operational data collected during the pilot, benchmarked with MAE, RMSE and $R^2$ on held-out data; a measured comparison of waiting times before and after deployment; an SMS/basic-mobile fallback channel based on [2], [7]; RL-based adaptive prioritization; and extension to other resource-constrained diagnostic units (e.g., radiology, ultrasound) following [21].

### *D. Conclusion*

This paper presented the design and prototype implementation of an intelligent queue-management and prediction system for a Sri Lankan public-hospital echocardiography unit, with digital registration, staff-controlled priority handling of urgent cardiac cases, wait-time and inflow prediction, and real-time patient and staff

dashboards. The design addresses a specific local operational need and an identified research gap, and provides an architectural template for other resource-constrained diagnostic units. Its benefits remain to be confirmed through live pilot evaluation.

## REFERENCES


[1] P. A. D. Dilrukshi, H. D. I. M. Nirmanamali, G. H. J. Lanel, and M. A. S. C. Samarakoon, "A strategy to reduce the waiting time at the outpatient department of the National Hospital of Sri Lanka," *Int. J. Sci. Res. Publ.*, vol. 6, no. 2, pp. 281–285, 2016.

[2] M. A. C. Akmal Jahan, S. L. Rathnayake, M. Kalansuriya, and F. Subair, "OPD-PMMAS: Patient management and mobile alert system for OPDs in Sri Lankan hospitals – a prototype," in *Proc. Int. Conf. Sci. Technol. (ICST)*, South Eastern Univ. Sri Lanka, Jul. 2021.

[3] S. Varshitha, S. D. Kulkarni, V. Roshni, V. Venkataraman, T. Monish, and M. V. Sandeep, "Modelling the patient flow in an Out Patient Department (OPD) of a hospital using simulation techniques," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1059, art. 012041, 2021, doi: 10.1088/1757-899X/1059/1/012041.

[4] M. A. C. Akmal Jahan, "Performance analysis of human detection algorithms for OPDs in public hospitals," in *Proc. 2023 IEEE Int. Workshop Artif. Intell. Image Process. (IWAIIP)*, Indonesia, 2023.

[5] M. A. C. Akmal Jahan and K. M. N. Kalansuriya, "Nameboard detection and text recognition in OPDs of public hospitals," in *Proc. 2023 IEEE Int. Workshop Artif. Intell. Image Process. (IWAIIP)*, Indonesia, 2023, pp. 1–6.

[6] eChannelling PLC / ICTA Sri Lanka, "e-Channelling launches its digital OPD reservation service at district hospitals in Sri Lanka," ICTA Sri Lanka, 2023.

[7] GSMA Mobile for Development, "Doc990: Enhancing patient access to healthcare services in Sri Lanka through mobile," GSMA Case Study, 2018.

[8] Ministry of Health, Nutrition and Indigenous Medicine, Sri Lanka, *Sri Lanka Digital Health Blueprint*, Health Information Unit, 2023.

[9] Ministry of Health, Sri Lanka, *National Digital Health Guidelines and Standards (NDHGS) v2.0*, 2023.

[10] Health Informatics Society of Sri Lanka (HISSL), *Digital Health in Sri Lanka: Sustainable Implementation, 1998–2017*, 2017.

[11] X. Li *et al.*, "Prediction of outpatient waiting time: using machine learning in a tertiary children's hospital," *Transl. Pediatr.*, vol. 12, no. 11, pp. 2030–2043, 2023.

[12] J. Joseph, S. Senith, A. A. Kirubaraj, and J. S. R. Ramson, "Machine learning for prediction of wait times in outpatient clinic," *Procedia Comput. Sci.*, 2022.

[13] C. Curtis, C. Liu, T. J. Bollerman, and O. S. Pianykh, "Machine learning for predicting patient wait times and appointment delays," *J. Am. Coll. Radiol.*, 2018.

[14] J. Shin, D. A. Lee, J. Kim, C. Lim, and B.-K. Choi, "Dissatisfaction-considered waiting time prediction for outpatients with interpretable machine learning," *Health Care Manag. Sci.*, vol. 27, no. 3, pp. 370–390, 2024, doi: 10.1007/s10729-024-09676-5.

[15] L. L. Guo *et al.*, "Predicting waiting times for medical tasks in a pediatric hospital using machine learning: Comprehensive, retrospective, real-

world study," *JMIR Med. Inform.*, vol. 13, art. e77297, 2025, doi: 10.2196/77297.

[16] A. Rastpour and C. McGregor, "Predicting patient wait times by using highly deidentified data in mental health care: Enhanced machine learning approach," *JMIR Ment. Health*, vol. 9, no. 8, art. e38428, 2022, doi: 10.2196/38428.

[17] J. Joseph, "Comparative analysis of machine learning algorithms for predicting consultation wait times in outpatient clinics," *Int. J. Eng. Trends Technol. (IJETT)*, 2025.

[18] S. Fomundam and J. Herrmann, "A survey of queuing theory applications in healthcare," ISR Tech. Rep. 2007-24, pp. 1–22, 2007.

[19] N. R. Ting, S. F. Sufahani, and M. Fahmy-Abdullah, "Enhancing the queuing management system in Malaysian public hospital," in *Proc. 3rd Int. Conf. Trends Comput. Cogn. Eng. (TCCE 2021), Lecture Notes Netw. Syst.*, vol. 348. Singapore: Springer, 2022, pp. 329–339, doi: 10.1007/978-981-16-7597-3_26.

[20] M. Cildoz, A. Ibarra, and F. Mallor, "Accumulating priority queues versus pure priority queues for managing patients in emergency departments," *Oper. Res. Health Care*, vol. 23, art. 100224, 2019, doi: 10.1016/j.orhc.2019.100224.

[21] Y. Pan *et al.*, "A high-fidelity, machine-learning enhanced queueing network simulation model for hospital ultrasound operations," arXiv:2104.07451, 2021.

[22] I. Adhicandra, S. Nurhidayati, and T. R. Fauzan, "Optimization of hospital queue management using priority queue algorithm and reinforcement learning for emergency service prioritization," *Int. J. Softw. Eng. Comput. Sci. (IJSECS)*, vol. 4, no. 2, 2024, doi: 10.35870/ijsecs.v4i2.2772.

[23] H. Kasliwal, A. Dhas, K. Kandhare, and R. Kandurke, "MediQueue: An ML-driven hospital queue management system with real-time wait time prediction," Preprints.org, 2026 (preprint, not peer reviewed), doi: 10.20944/preprints202604.0628.v1.

[24] K. Kavitha, S. Maragathasundari, and K. Karthikeyan, "AI-driven healthcare: A queuing-based framework for optimized patient management," in *Proc. 6th Int. Conf. Data Intell. Cogn. Informat. (ICDICI)*,Tirunelveli, India, 2025, doi: 10.1109/ICDICI66477.2025.11134866.

[25] M. Ghazal, R. Hamouda, and S. Ali, "A smart mobile system for the real-time tracking and management of service queues," 2016.

[26] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.

[28] B. L. Macrum, "Workflow optimization in echocardiography," *US Cardiol.*, vol. 1, no. 1, pp. 1–3, 2004.

[29] J. S. Solis, F. Falahi, S. Garcia Casado, M. Rodriguez Martinez, and S. Salguero Bodes, "Optimizing echocardiography workflow to increase access to high quality care for patients," *Eur. Heart J. Cardiovasc. Imaging*, vol. 24, Suppl. 1, jead119.094, 2023.

[30] K. A. Taiwo, G. I. Olatunji, and O. O. Akomolafe, "Forecasting hospital resource demand using time series and machine learning models," *Gulf*

[31] *J. Adv. Bus. Res.*, vol. 3, no. 8, 2025. Parliament of the Democratic Socialist Republic of Sri Lanka, *Personal Data Protection Act, No. 9 of 2022*, 2022.

# A Dual-Method Framework Integrating AHP-TOPSIS and Temporally Validated Machine Learning for Export Market Recommendation: Evidence from Sri Lanka's Apparel Sector

UGD Nijai
*Department of Computer and Data Science*
*Faculty of Computing,*
*NSBM Green University, Homagama, Sri Lanka*
gamagedinuri@gmail.com

RWKT Rajapaksha
*Department of Computer and Data Science*
*Faculty of Computing,*
*NSBM Green University, Homagama, Sri Lanka*
gamagedinuri@gmail.com

***Abstract*—Sri Lanka's apparel exports increased from approximately USD 4.0 billion in 2010 to USD 5.6 billion in 2024, but this growth lagged behind several Asian competitors. To address this competitive gap, this study develops an integrated decision-support framework to identify realistic export destinations for HS 61-62 apparel. The framework integrates AHP-weighted TOPSIS, temporally evaluated XGBoost, and a commercial-readiness screening pipeline to optimize market selection. The analysis uses 211,333 exporter–importer-year observations for 2010-2024. Methodologically, the TOPSIS component evaluates market attractiveness based on six structural criteria. The machine-learning model estimates whether an exporter will meet or exceed the median three-year market-share change among same-year peers with a similar starting share. The selected XGBoost model used 91 features. It achieved a mean annual ROC-AUC of 0.6679 in the 2020 and 2021 retrospective evaluations, with a pooled ROC-AUC of 0.6683. Integration of the 2024 rankings revealed 53 eligible markets, with 52 demonstrating readiness for expansion. Combining market readiness and ranking support narrowed these to 25 actionable destinations. Among expansion-ready markets, Poland was the only country to receive top-quartile support from both assessment methods. Using the median Sri Lankan market share of 0.4908% as the benchmark, markets were classified into established and under-penetrated groups. The established pool was led by Türkiye, Poland, Thailand, Austria and France. The underpenetrated group was dominated by the Philippines, Denmark, the Republic of Korea, Chile, and Kazakhstan. These findings differentiate markets based on structural attractiveness, predicted outperformance, and demonstrated commercial accessibility.***

***Keywords—export market selection, AHP-TOPSIS, XGBoost, temporal validation, trade diversification.***

## I. Introduction

Sri Lanka's apparel industry is a major source of export earnings and is recognized for its commitment to ethical manufacturing, labour compliance, and value-added production. However, these strengths have not translated into export growth compare to several regional competitors. Between 2010 and 2024, Sri Lankan apparel exports increased from approximately USD 4.0 billion to USD 5.6 billion, which represents growth of only 1.4 times. In comparison, apparel exports increased 3.04 times in Bangladesh, 3.54 times in Vietnam, and 3.90 times in Cambodia [1]. During the same period, Sri Lanka's share of global apparel exports changed only marginally, from 1.21% to 1.31%. Therefore, the focus should shift from merely retaining current buyers to strategically identifying markets for either expanding an existing destination or testing credible new export destinations.

Export success cannot be explained by geographic diversification alone. Bangladesh expanded rapidly while remaining strongly oriented toward Europe, whereas Vietnam grew through a more diversified market structure. As a result, effective market selection should also consider factors such as demand scale, market growth, tariffs, competition, and the exporter's own position. A large or fast-growing destination may still be difficult to enter, and a high model score can be misleading when commercial evidence is weak.

This study addresses the export market selection challenge through three linked but separately interpretable stages. AHP-TOPSIS evaluates current structural attractiveness. XGBoost estimates future peer-relative outperformance using chronological validation. A final screening stage checks whether a statistically promising market has evidence of access through competitor presence, market size, recent trade engagement, and Sri Lanka's current foothold. This study contributes a transparent, three-stage prioritization procedure that distinguishes between structural attractiveness and predicted performance while testing their commercial actionability.

## II. Literature Review

Sri Lanka's apparel sector has successfully transitioned from quota-supported production to a competitive model centered on quality, compliance, ethical manufacturing, and higher-value apparel products [2]. However, its performance remains sensitive to imported raw materials, external demand, trade preferences, and concentration in major western destinations [3]. Evidence from Sri Lanka indicates a long-run association between market diversification and export performance [4], while Bangladesh's expansion into non-European markets contributed to strong export growth [5].

These studies support destination analysis, but they do not show that geographic expansion guarantees export success.

Revealed comparative advantage measures have been widely used to assess changes in export specialization among apparel exporters [6, 7]. However, their usefulness for market selection is limited because they reflect past trade performance rather than identify future market opportunities. International market selection also requires the evaluation of multiple criteria simultaneously. Previous literature on textile market selection combined AHP weighting with TOPSIS ranking [8], and TOPSIS has been applied to country selection problems more broadly [9]. AHP depends on subjective judgment, and TOPSIS provides only a static view of observed conditions [10, 11].

Recent machine-learning studies show that nonlinear models can improve trade prediction by capturing complex relationships among market variables [12, 13], but existing studies primarily aim to forecast trade outcomes rather than support export market recommendation. Model performance can also be overstated when temporally related observations are randomly divided between training and testing samples. Time-aware validation is therefore important in panel trade data [14]. A further challenge relates to the choice of performance measure. Percentage growth can make small-base markets appear disproportionately attractive, whereas absolute growth favors established markets with large export bases. Therefore, the existing approaches lack a framework that jointly evaluates structural attractiveness, predicts peer-relative future performance, and verifies the commercial feasibility of the resulting market selections.

## III. Methodology

### *A. Data and Feature Construction*

Bilateral HS 61-62 trade data covering the period 2010-2024 were obtained from UN Comtrade [1]. Macroeconomic indicators, including GDP growth, GDP per capita, and population, were sourced from the World Development Indicators [15]. Tariff information was collected from WITS/TRAINS [16], and gravity and policy variables were retrieved from CEPII [17]. After country-code standardization, aggregate removal, and duplicate checks, the final dataset contained 211,333 unique exporter-importer-year observations, covering 235 exporters and 181 importers.

Derived variables included market size, exporter share, supplier count, Sri Lanka's share, HHI on the 0-10,000 scale, and exact three-year market CAGR. Extreme CAGR values were clipped to reduce the influence of outliers. Missing tariff observations were filled using a predefined sequence: observed bilateral tariffs, previous bilateral tariffs, importer-year median tariffs, and importer historical median tariffs. Bilateral distance was standardized, macroeconomic gaps were imputed with audit flags, and skewed size, income, and population measures were log transformed.

For retrospective ML evaluation, tariff observations without prior observed bilateral data were treated as missing. Similarly, importer-median macroeconomic values were excluded when their historical availability could not be verified. Data imputation, feature scaling, and categorical encoding steps were performed separately within each training fold. Remaining contemporaneous imputations and historical source vintages were not fully audited.

For structural market selection, Sri Lanka's 2024 destinations with complete criteria were considered. The assessment took account of concerns about re-export and transshipment activity, export-oriented production networks, and territorial status. Destinations such as Hong Kong, Macao, Singapore, Panama, Nicaragua, Honduras, El Salvador, and Mayotte were excluded to improve comparability among destination markets. This defined the study's scope and did not imply that the excluded destinations lacked commercial potential. The median Sri Lankan market share was 0.4908%, calculated across the remaining complete cases before applying the minimum import-market-size threshold of USD 500 million. Destinations with Sri Lankan market shares above this median were grouped into the established pool, while those at or below the median formed the under-penetrated pool. TOPSIS rankings were calculated separately within each pool.

### *B. AHP-TOPSIS Ranking*

TOPSIS was applied using six criteria, namely log market size, three-year growth, Sri Lanka's share, tariff, HHI, and supplier count. Market size and growth were treated as benefits while tariff, HHI, and supplier count were treated as cost criteria. Sri Lanka's market share was considered a benefit in established markets because it reflected an existing foothold. But in the under-penetrated pool it was treated as a cost because a lower share indicated greater room for expansion.

The AHP weighting process ranked the criteria in the following order of importance: market size, growth, market share, tariff, market concentration, and supplier count. The researcher assigned the pairwise judgments for this study, informed by related textile and export-market-selection literature [8, 18]. These studies provided guidance on relevant evaluation criteria. However, the numerical values in the pairwise comparison matrix were determined specifically for this study and were not adopted directly from previous research. The final criterion weights were calculated by normalizing the geometric mean of each matrix row. The pairwise comparison matrix yielded a consistency ratio of 0.012, below the 0.10 threshold [10]. TOPSIS then normalized the criteria, applied the weights, and ranked each pool by relative closeness to the ideal solution [11]. Equal and entropy weights were used as robustness checks.

AHP P

TABLE 1
AIRWISE COMPARISON MATRIX AND NORMALIZED WEIGHTS

| Criterion | Size | Growth | Share | Tariff | HHI | Suppliers | Weight |
|---|---|---|---|---|---|---|---|
| Log market size | 1 | 2 | 2 | 2 | 3 | 4 | 0.3166 |
| Three-year growth | 1/2 | 1 | 1 | 2 | 2 | 3 | 0.1995 |
| Sri Lanka share | 1/2 | 1 | 1 | 1 | 2 | 3 | 0.1777 |
| Tariff | 1/2 | 1/2 | 1 | 1 | 1 | 2 | 0.1318 |
| HHI | 1/3 | 1/2 | 1/2 | 1 | 1 | 2 | 0.1098 |
| Supplier count | 1/4 | 1/3 | 1/3 | 1/2 | 1/2 | 1 | 0.0646 |

Weights were calculated by normalizing the geometric mean of each row. The consistency ratio was 0.012.

For market *i*, the TOPSIS closeness coefficient was

$C_i = \frac{S_i^-}{S_i^+ + S_i^-}$. (1)

where $S_i^+$ and $S_i^-$ are the distances from the positive and negative ideal solutions. A higher coefficient indicates a market closer to the preferred structural profile.

*C. Peer-Relative Target and Temporal Evaluation*

Each exporter-importer record in year $t$ was matched to the same pair in year $t + 3$. Its change in market share was compared with the median change among observations in the same base year and starting-share quintile. Observations that performed at or above the median were assigned a value of 1, while those below the median were assigned a value of 0. This produced an approximately balanced outcome and reduced the bias in favor of markets with very tiny starting shares. The original experiments used 2010-2017 for development, 2018-2019 for validation, and 2020-2021 for later-period comparison. For the reported retrospective evaluation, models were fitted separately for 2020 and 2021. Training data for the 2020 model included base years up to 2016, while training data for the 2021 model included base years up to 2017. This ensured that training-label outcome years preceded the evaluation year. Data preprocessing, including imputation, scaling, and encoding, was performed separately within each training fold. Also, historical-success encodings used the same prior-outcome restriction. The 91-feature model was evaluated on 25,915 observations from 2020-2021, a period that had previously informed model selection.

*D. Model, Integration, and Commercial Screening*

The selected XGBoost classifier used 91 historically constructed predictors. These features included 55 core variables covering trade, market conditions, lags, momentum, policy, gravity, and peer market position. The model also incorporated 15 historical-success encodings, and 21 bilateral relationship-dynamics measures. The final model used 500 trees, learning rate 0.08, maximum depth five, minimum child weight 10, subsample 0.75, column subsample 0.70, and L1/L2 regularization of 0.5/2.0.

To generate the 2024 predictions, the model was refitted using labelled observations from base years 2010-2020 and then applied to Sri Lanka's 2024 records. This approach ensured that all training-label outcomes occurred before the prediction base year. Therefore, the resulting scores estimate peer-relative performance over 2024-2027, for which outcome data are not yet available. Non-sovereign destinations were removed. Within destinations passing the initial accessibility screening, TOPSIS and ML ranks were calculated separately for the established and under-penetrated pools. A destination was considered to have top-quartile support if its rank fell within the top 25% of markets in its group, with ties included at the cutoff. This gave cutoffs of rank 10 for the 37 established markets and rank 4 for the 16 under-penetrated markets.

Markets supported by both methods were classified as Strong Recommend. Support from TOPSIS alone produced Structurally Attractive, while support from ML alone led to ML Growth Signal. Destinations outside both cutoffs were classified as Monitor. To be considered actionable, a market had to receive support from at least one method and be expansion-ready.

The top-quartile rule was used to identify a similar proportion of high-priority destinations within each market group. This was a practical decision rather than an empirically optimal threshold. Sensitivity was assessed using alternative cutoffs of 20% and 30%, holding the rankings and readiness decisions fixed. For within-pool rank $r$, the normalized rank was $1 - (r - 1)/(N - 1)$. The normalized TOPSIS and ML ranks were averaged to obtain the integrated ordering score. This score was used for prioritization and was not interpreted as a probability.

Commercial screening was carried out in two steps. First, an accessibility assessment identified eligible destinations before ranking and integration. Second, a readiness assessment evaluated whether those destinations were suitable for recommendation. The broad accessibility assessment considered six Asian apparel competitors: Bangladesh, Viet Nam, India, Cambodia, Pakistan, and China. Evidence was compiled using data from 2022 to 2024, and at least two years of observations were required. A competitor was materially present in a year if its exports exceeded USD 1 million or its market share reached 0.1%. A destination qualified for further analysis when sufficient data were available and either the average combined competitor market share was at least 1% or at least two competitors were materially present on average.

Later, the readiness assessment was conducted using five comparable Asian apparel exporters, excluding China. This allowed the analysis to focus on exporters that are more similar to Sri Lanka. Among destinations passing the accessibility gate, a market was classified as exploratory when Sri Lanka's 2024 exports were below USD 100,000, the comparable peers' average combined share was below 1%, and their average total annual exports were below USD 10 million. The peer averages used the observed destination-years in 2022–2024. Other destinations with Sri Lankan exports below USD 100,000 were classified as entry-test candidates. Destinations with Sri Lankan exports of at least USD 100,000 were classified as expansion-ready.

Only markets that were expansion-ready and ranked in the top quartile under at least one method were included in the actionable shortlist. These screening thresholds were operational rules rather than empirically validated measures of market-entry success. Predictive performance was assessed separately for each evaluation year and for the pooled observations. The annual ROC-AUCs were also summarized using their unweighted mean and sample standard deviation. Calibration was examined using ten approximately equal-frequency score bins and the Brier score.

A 95% confidence interval for pooled ROC-AUC was estimated using 1,000 bootstrap resamples grouped by exporter–importer pair, keeping observations from the same pair together. This interval describes uncertainty conditional on the fitted models and evaluation years. No recalibration was applied to the deployment scores.

## IV. Results and Discussion

*A. Structural Ranking and Predictive Performance*

In TOPSIS, the leading established markets were Türkiye, Malaysia, India, Mexico, and Brazil. The under-penetrated pool

included Kazakhstan, the Philippines, Romania, Latvia, and Greece. Within the 53 integrated destinations, Spearman correlations between AHP and equal-weight ranks were 0.898 for established markets and 0.947 for under-penetrated markets. The AHP and entropy-weighted rankings generally agreed, with Spearman correlations of 0.701 for established markets and 0.862 for under-penetrated markets. TOPSIS assessed market attractiveness based on the conditions observed in 2024.

The 91-feature XGBoost model achieved ROC-AUCs of 0.6802 in 2020 and 0.6556 in 2021, giving an unweighted annual mean of 0.6679 and a sample standard deviation of 0.0174. Across the pooled 2020–2021 observations, ROC-AUC was 0.6683, average precision was 0.6557, accuracy was 0.61995, and F1-score was 0.64192 (Table 2). A 1,000-resample bootstrap grouped by exporter–importer pair gave a 95% confidence interval of 0.6609-0.6754 for pooled ROC-AUC. Because the 2020–2021 period had previously influenced model selection, these results represent a retrospective comparison rather than an untouched independent-test estimate. Predicted scores are used to prioritize markets, not to guarantee export growth.

TABLE 2
RETROSPECTIVE EVALUATION OF THE 91-FEATURE XGBOOST MODEL

| Metric | Value |
|---|---|
| ROC-AUC, 2020 | 0.6802 |
| ROC-AUC, 2021 | 0.6556 |
| Mean annual ROC-AUC | 0.6679 |
| Pooled ROC-AUC, 2020–2021 | 0.6683 |
| Pooled average precision | 0.6557 |
| Pooled accuracy | 0.61995 |
| Pooled F1-score | 0.64192 |
| Pooled Brier score | 0.23049 |

Calibration was assessed using ten approximately equal-frequency bins from the same 2020–2021 predictions (Fig. 1). The Brier score was 0.2305, compared with 0.2500 for a constant score equal to the evaluation prevalence. In the highest bin, the mean predicted score was 0.851 and the observed positive fraction was 0.762, indicating overconfidence at high scores. No recalibration was applied to the deployment rankings. The scores concern peer-relative outperformance and should not be interpreted as calibrated probabilities of commercial success.

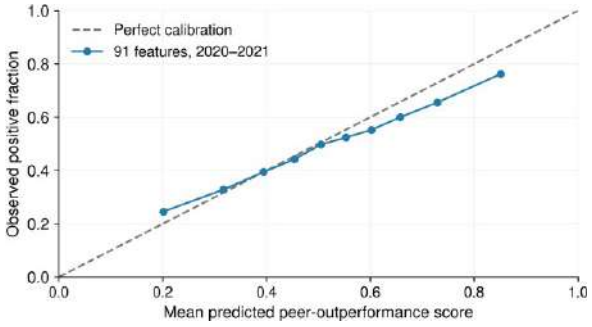



Fig. 1. Reliability of the 91-feature scores across ten approximately equal-frequency bins, using the 2020–2021 predictions.

*B. Integrated and Commercially Screened Recommendations*

TOPSIS and ML shared 53 eligible destinations, including 37 established and 16 under-penetrated markets. All 53 passed the broad competitor-presence gate. The subsequent assessment using comparable exporters classified 52 markets as expansion-ready and Latvia as exploratory.

Among the 52 expansion-ready markets, 25 received top-quartile support from TOPSIS, ML, or both and entered the actionable shortlist. The shortlist comprised 19 established and six under-penetrated destinations. The other 27 expansion-ready markets were assigned to the Monitor category. Poland was the only expansion-ready strict-consensus market that met the cutoff in both rankings.

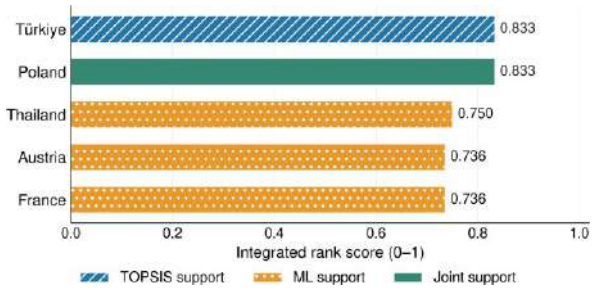



Fig. 2. Top five actionable established markets. Green denotes Strong Recommend, blue denotes Structurally Attractive, and orange denotes ML Growth Signal. Integrated scores are not probabilities.

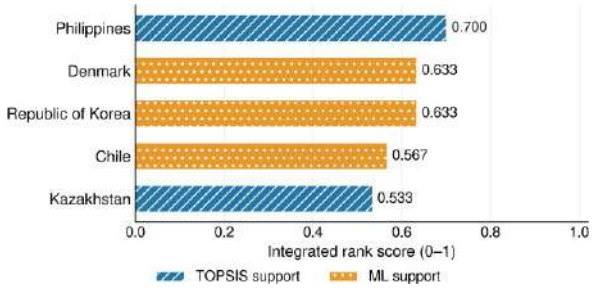



Fig. 3. Top five actionable under-penetrated markets. Blue denotes Structurally Attractive, while yellow denotes ML Growth Signal. Integrated scores are not probabilities.

Poland was the only final Strong Recommend market. It ranked ninth under TOPSIS and fifth under ML, with an integrated score of 0.8333. Its 2024 apparel market was approximately USD 16.14 billion, three-year growth was 8.86%, HHI was 1,338, Sri Lanka already held about 0.95%, and the model-estimated score was 0.7036. The two methods therefore support examining opportunities to deepen Poland's existing foothold.

The leading actionable established markets were Türkiye, Poland, Thailand, Austria, and France (Fig. 2). Thailand, Austria, and France received top-quartile predictive support without top-quartile TOPSIS support. This suggests that the two methods captured different aspects of market potential. Türkiye showed the opposite pattern: it ranked first under TOPSIS but thirteenth under ML. Although Türkiye and Poland both had integrated scores of approximately 0.8333, only Poland met both ranking cutoffs and the readiness condition. This implies that Türkiye's strong structural attractiveness does not

necessarily establish future Sri Lankan share gains. Instead, the market appears more suitable for focused investigation of product opportunities and buyer demand before any expansion decisions are made.

The leading actionable under-penetrated markets were the Philippines, Denmark, the Republic of Korea, Chile, and Kazakhstan (Fig. 3). The Republic of Korea received an ML score of approximately 0.7800. However, this score represents a relative ranking measure and should not be interpreted as a direct probability of commercial success. Latvia illustrates the value of the final screening stage. It received top-quartile support from both methods, but its Sri Lankan shipments were only about USD 3,415, comparable-peer average combined share was 0.62%, and comparable-peer average annual exports were USD 3.44 million. It was therefore treated as an exploratory rather than an actionable opportunity.

Sensitivity analysis revealed that the actionable shortlist changed with the ranking cutoff. Using the same 2024 rankings and readiness classifications, thresholds of 20%, 25%, and 30% identified 22, 25, and 30 actionable destinations, respectively. No expansion-ready market achieved joint TOPSIS and ML support at the 20% level. At the 25% level, only Poland satisfied both criteria, while Poland and Austria did so at 30%. Therefore, these findings suggest that the final shortlist depends on the chosen cutoff level and the results do not establish 25% as an optimal threshold.

### *C. Competitive Context and Practical Meaning*

Sri Lanka's apparel exports increased by 40%, rising from approximately USD 4.0 billion to USD 5.6 billion between 2010 and 2024. Although export earnings continued to grow, the pace of expansion lagged behind several regional competitors. Within the same period, apparel exports expanded by 3.90 times in Cambodia, 3.54 times in Vietnam, 3.04 times in Bangladesh, and 2.67 times in Pakistan, while India recorded a more modest increase of 1.23 times. Sri Lanka's share of global apparel exports increased only marginally, from 1.21% to 1.31%, indicating that Sri Lanka maintained its position but captured international growth more slowly than key competitors.

In 2024, Sri Lanka's apparel exports remained concentrated in a few major destinations. The largest destinations were the United States (40.25%), the United Kingdom (14.18%), Italy (10.13%), Germany (5.38%), and the Netherlands (4.68%). Together, these five markets represented approximately 74.62% of Sri Lanka's covered apparel exports. These destinations remain commercially important, but the concentration also supports the need to identify additional opportunities (Fig. 5).

Geographic concentration alone did not explain the export growth differences. For example, Europe received 49.4% of Sri Lanka's apparel exports and 47.8% of Cambodia's, yet Cambodia achieved considerably faster growth. Bangladesh also achieved strong growth while sending 67.2% of its exports to Europe. Vietnam followed a more geographically diversified pattern, with only 22.4% directed to Europe. These contrasting experiences suggest that export growth is influenced by a broader set of factors such as market access, production scale, pricing, product positioning, delivery capability, and buyer relationships, alongside destination diversification.

The divergence between the methods separates favorable present conditions from modeled peer-relative performance. The framework therefore supports a portfolio approach. Türkiye, Poland, Thailand, Austria, and France provide opportunities to deepen existing market relationships, while the Philippines, Denmark, the Republic of Korea, Chile, and Kazakhstan represent under-penetrated destinations for further exploration. Under-penetrated does not necessarily mean that Sri Lanka has no exports to these countries. It indicates that Sri Lanka's market share was at or below the study median used to define the two pools. These results should be used as decision-support evidence rather than automatic market-entry decisions. Apparel experts should also assess product-level demand, buyer availability, freight costs, trade preferences, rules of origin, sustainability requirements, and domestic production capacity before committing resources.

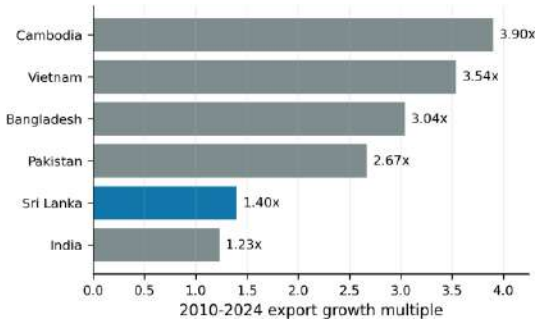



*Fig. 4. Apparel export-value growth multiples for Sri Lanka and selected Asian competitors, 2010–2024.*

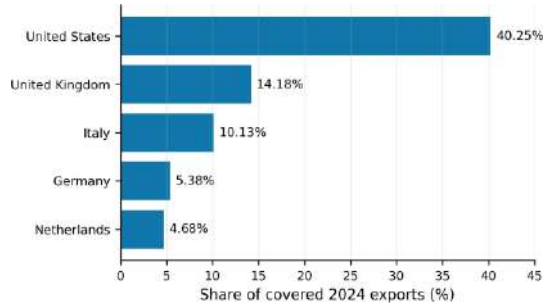



*Fig. 5. Sri Lanka's five largest apparel export destinations in 2024.*

### *D. Limitations*

The target measures peer-relative market-share performance, not revenue, profit, or entry success. The recommendations depend on the six criteria, the AHP weights, the USD 500 million minimum market size, the division into market pools, and the screening rules. The screening thresholds require empirical validation as indicators of entry feasibility. In addition, some tariff and macroeconomic variables were estimated through imputation. Moreover, the use of country-level data does not capture differences across products, buyers, or market segments. Results from later periods had already informed model selection, and evaluating the 2024 deployment requires the corresponding three-year outcomes in 2027.

Training-label outcomes were restricted to periods that occurred before each evaluation base year. But historical data vintages and the remaining same-year imputations require

further auditing. The labelled dataset included only observations with available outcome data in $t + 3$, and some cross-sectional peer predictors were constructed within this retrospectively selected cohort. A prospective evaluation should calculate these predictors using all observations available in the base year and test how missing future outcomes affect the results. The reported results therefore describe a retrospective comparison and do not establish real-time forecasting performance. Future evaluation should use newly observed outcomes and examine logistics, rules of origin, and buyer demand.

## V. Conclusion

This study combined AHP-TOPSIS, temporally evaluated XGBoost, and commercial-readiness screening into a unified export-market recommendation framework. The findings suggest that structural market attractiveness and predicted peer-relative performance provide complementary evidence to inform market selection. AHP-TOPSIS identified Türkiye as the highest-ranked established market and Kazakhstan as the leading under-penetrated market. The machine-learning component provided an additional perspective on potential exporter performance.

The integrated analysis identified 53 eligible destinations, of which 52 were classified as expansion-ready. After considering both the rankings and commercial readiness, the study shortlisted 25 markets for further action. Among these 25 markets, 19 were established markets and six were under-penetrated markets. Poland was the only expansion-ready market receiving top-quartile support from both methods. It ranked ninth in TOPSIS and fifth in ML. The leading actionable established markets were Türkiye, Poland, Thailand, Austria, and France. The leading actionable under-penetrated markets were the Philippines, Denmark, the Republic of Korea, Chile, and Kazakhstan. These findings highlight the importance of considering both existing market conditions and peer-relative performance signals rather than relying on a single ranking method.

The findings provide distinct decision-making implications for the two market pools. In established destinations, exporters can focus on strengthening existing buyer relationships, broadening product offerings, and increasing market share where commercial access has already been demonstrated. In under-penetrated markets, they can study demand, look for buyers, attend trade fairs, and test the market on a small scale before investing more. Export-promotion agencies and policymakers can use the shortlist to prioritize trade missions, buyer-matching programmes, market intelligence, and targeted export-support activities. Destinations supported by both ranking approaches, especially Poland, represent the strongest candidates for further evaluation. Meanwhile, markets supported by only one method need further assessment before major investment decisions are made.

The study contributes to a transparent and evidence-based approach that distinguishes structural attractiveness, modeled performance, and commercial readiness. This framework provides practical support for exporters and policymakers in prioritizing promising export destinations and allocating market-development resources. However, exporters should conduct further product-specific and buyer-focused analysis before making market-entry decisions.Future work should evaluate the recommendations against newly observed outcomes and assess the usability of the framework through engagement with apparel industry professionals and export decision-makers.

## References


[1] United Nations Statistics Division, “UN Comtrade Database.” Accessed: Sep. 9, 2026. [Online]. Available: https://comtradeplus.un.org/

[2] S. Kelegama, Ed., *Ready-Made Garment Industry in Sri Lanka: Facing the Global Challenge*. Colombo, Sri Lanka: Institute of Policy Studies of Sri Lanka, 2004.

[3] N. Ekanayake, “Determinants of external demand for textiles and garments of Sri Lanka,” *Staff Studies*, vol. 43, nos. 1–2, pp. 85–104, 2016, doi: 10.4038/ss.v43i1-2.4691.

[4] T. Bhavan, “Export market diversification and export performance of Sri Lanka: a cointegration analysis,” *Asian Journal of Empirical Research*, vol. 7, no. 4, pp. 75–83, Apr. 2017, doi: 10.18488/journal.1007/2017.7.4/1007.4.75-83.

[5] R. Ahmed, M. T. Islam, and M. Al-Amin, “The Effects of Market Diversification Activities on Bangladesh RMG Export,” *International Journal of Economics and Financial Issues*, vol. 3, no. 4, pp. 938–948, 2013, [Online]. Available: www.econjournals.com

[6] M. R. Hasan, I. S. Swazan, and D. Das, “Beyond the seams: evaluating competitiveness and comparative advantage in Vietnam’s apparel industry,” *Journal of Chinese Economic and Foreign Trade Studies*, vol. 17, no. 1, pp. 4–28, Jun. 2024, doi: 10.1108/JCEFTS-10-2023-0050.

[7] S. Das, M. R. Hasan, and D. Das, “Who is the next China? Comparative advantage analysis from top ten apparel exporting nations,” *Competitiveness Review*, vol. 34, no. 6, pp. 1127–1154, Oct. 2024, doi: 10.1108/CR-06-2023-0143.

[8] J. G. Vanegas-López, J. J. Baena-Rojas, D. A. López-Cadavid, and M. Mathew, “International market selection: an application of hybrid multi-criteria decision-making technique in the textile sector,” *Review of International Business and Strategy*, vol. 31, no. 1, pp. 127–150, Mar. 2021, doi: 10.1108/RIBS-07-2020-0088.

[9] A. V. Christian, Y. Zhang, and C. K. Salifou, “Country Selection for International Expansion: TOPSIS Method Analysis,” *Modern Economy*, vol. 07, no. 04, pp. 470–476, 2016, doi: 10.4236/me.2016.74052.

[10] T. L. Saaty, *The Analytic Hierarchy Process*. New York, NY: McGraw-Hill, 1980.

[11] G.-H. Tzeng and J.-J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. Boca Raton, FL: CRC Press, 2011.

[12] M. Gopinath, F. A. Batarseh, J. Beckman, A. Kulkarni, and S. Jeong, “International agricultural trade forecasting using machine learning,” *Data & Policy*, vol. 3, Art. no. e1, 2021, doi: 10.1017/dap.2020.22.

[13] H. Jošić and B. Žmuk, “A Machine Learning Approach to Forecast International Trade: The Case of Croatia,” *Business Systems Research*, vol. 13, no. 3, pp. 144–160, Oct. 2022, doi: 10.2478/bsrj-2022-0030.

[14] C. Bergmeir and J. M. Benítez, “On the use of cross-validation for time series predictor evaluation,” *Information Sciences*, vol. 191, pp. 192–213, 2012, doi: 10.1016/j.ins.2011.12.028.

[15] World Bank, “World Development Indicators,” *World Bank DataBank*. Accessed: Sep. 9, 2026. [Online]. Available: https://databank.worldbank.org/source/world-development-indicators

[16] World Bank, “World Integrated Trade Solution (WITS).” Accessed: Sep. 9, 2026. [Online]. Available: https://wits.worldbank.org/

[17] M. Conte, P. Cotterlaz, and T. Mayer, “The CEPII Gravity database,” CEPII Working Paper no. 2022-05, Jul. 2022. Accessed: Sep. 9, 2026. [Online].
Available: https://www.cepii.fr/cepii/en/bdd_modele/bdd_modele_item.asp?id=8

[18] A. Yıldız and A. Özbek, “Selection of socks export markets for Turkey using multi-criteria decision making methods,” *Sigma Journal of Engineering and Natural Sciences*, vol. 38, no. 2, pp. 795–815, 2020.

# Factors Affecting Citizens' Acceptance of the Inland Revenue Department's Online Tax Services in Sri Lanka: An Application of the Technology Acceptance Model (TAM)

S.G.N.S. Rupasinghe
*Department of Software Engineering and Computer Security*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
nadeeshar48@gmail.com

Thisarani Wickramasinghe
*Department of Software Engineering and Computer Security*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

***Abstract*—The Inland Revenue Department (IRD) of Sri Lanka has progressively digitized tax administration through the Revenue Administration Management Information System (RAMIS) and associated e-services, yet citizen uptake remains uneven and its behavioural drivers are not well understood. This study extends the Technology Acceptance Model (TAM) with Trust and Computer Self-Efficacy (CSE) and examines how Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Trust and CSE affect Behavioural Intention (BI), and how BI relates to self-reported actual usage. A cross-sectional online survey of 150 registered taxpayers (five-point Likert scale) was analyzed using reliability testing, descriptive statistics, correlation, and multiple regression with multicollinearity diagnostics. All constructs were reliable (Cronbach's alpha 0.80–0.90), but all construct means were below the scale mid-point (2.33–2.68), indicating limited acceptance. The regression model explained 74.2% of the variance in BI ($R^2 = 0.742$, $p < 0.001$); CSE was the only significant determinant ($B = 0.60$, $p < 0.001$), while PU ($B = 0.17$, $p = 0.123$), Trust ($B = 0.16$, $p = 0.187$) and PEOU ($B = -0.005$, $p = 0.955$) were not significant. BI strongly predicted actual usage ($B = 0.79$, $p < 0.001$, $R^2 = 0.684$). Given the small sample, the findings are preliminary; they suggest that building taxpayers' confidence and capability is the most promising lever for increasing adoption.**

***Keywords*—Technology Acceptance Model, e-government, online tax services, perceived usefulness, perceived ease of use, trust, computer self-efficacy, behavioural intention, Sri Lanka.***

## I. Introduction

Governments worldwide use information and communication technologies (ICT) to modernize public service delivery, and tax administration is one of the most active domains [1], [2]. Electronic filing and payment promise convenience, faster processing, lower compliance costs and fewer manual errors, but these benefits depend less on the sophistication of the technology than on whether citizens accept and continue to use it.

In Sri Lanka, the Inland Revenue Department (IRD) introduced the Revenue Administration Management Information System (RAMIS) to automate core tax processes and to deliver e-services such as taxpayer registration, electronic filing of returns and online payment. Successive digitization initiatives, including a national e-invoicing framework integrated with RAMIS, make these services an increasingly central channel between the state and taxpayers. Their effectiveness, however, depends on uptake: an online channel that is available but under-used cannot deliver its intended administrative and revenue benefits. Online tax filing and payment sit at the most advanced, transactional stage of e-government, where demands on usability, security and citizen confidence are greatest.

The Sri Lankan e-government literature indicates that citizen adoption remains low relative to the resources committed, and that empirical research explaining adoption behaviour is scarce [3]. Many taxpayers still rely on intermediaries or manual channels, and adoption is constrained by trust, awareness, perceived risk and limited user capability [3].

TAM and its extensions have been applied to electronic tax systems in Taiwan, Malaysia and elsewhere [2], [13], [15], but little work has examined citizen acceptance of online tax services in Sri Lanka. Ranaweera [3] modelled trust in general e-government using TAM without focusing on tax services, and Shantha [26] examined individual e-tax filing through a UTAUT-based, risk-oriented lens, finding that perceived usefulness, perceived ease of use, perceived risk and facilitating conditions influenced intention. To the best of the authors' knowledge, no prior Sri Lankan study has combined PU, PEOU, Trust and CSE in a single TAM-based model with BI as the predictor of actual usage of the IRD's online tax services. This study addresses that gap.

*Objectives of the study.* The main objective is to investigate the factors affecting citizens' acceptance of the IRD's online tax services using TAM. The specific objectives are: (1) to examine the effect of PU on BI; (2) to examine the effect of PEOU on BI; (3) to determine the effect of Trust on BI; and (4) to determine the effect of CSE on BI; and (5) to evaluate the relationship between BI and actual usage.

*Research questions.* RQ1: How does PU affect BI to use the IRD online tax services? RQ2: How does PEOU affect BI?

RQ3: What is the effect of Trust on BI? RQ4: How does CSE influence BI? RQ5: How does BI affect actual usage?

*Significance and scope.* For the IRD and policymakers, the study provides evidence on which factors most strongly shape citizen acceptance. For the academic community, it provides context-specific evidence on an extended TAM from Sri Lanka, where such studies are limited. The scope is limited to individual citizens who are registered taxpayers and have used, or attempted to use, the IRD's online tax services. Corporate tax agents, the IRD's internal perspective and the technical architecture of RAMIS are excluded. The study is cross-sectional and confined to the four determinants in the conceptual framework.

*Empirical.* evidence from a mandatory, data-sensitive e-government service in a developing-country setting where such evidence is scarce. *Theoretical.* a test of whether Trust and CSE add explanatory power to the TAM core beliefs. The finding that CSE, rather than PU, PEOU or Trust, was the only significant determinant of BI indicates that, among taxpayers with limited acceptance, personal capability may outweigh the instrumental and trust beliefs that dominate prior tax-filing studies. This qualifies the usual primacy of PU in TAM research and motivates further testing with larger samples. *Practical.* identification of taxpayer confidence and capability as the priority area for IRD intervention.

## II. Literature Review

*Contributions of the study.* The Theory of Reasoned Action (TRA) holds that behaviour is driven by intention, which is shaped by attitudes and subjective norms [4]. The Theory of Planned Behaviour (TPB) adds perceived behavioural control [5], and Diffusion of Innovations explains adoption through attributes such as relative advantage and complexity [6]. TAM was proposed specifically for information systems [1] and was later consolidated into UTAUT [7] and UTAUT2 [8]. TAM is the most widely validated of these models because of its parsimony and predictive power [3], [9], and is adopted here as the theoretical foundation.

TAM proposes that acceptance is determined by two beliefs: Perceived Usefulness (PU), the degree to which a person believes that using a system would enhance their performance, and Perceived Ease of Use (PEOU), the degree to which a person believes that using it would be free of effort [1]. PEOU is theorised to influence PU, and both shape intention, which predicts actual use. TAM2 added social-influence and cognitive-instrumental processes [10]. A recognised limitation of TAM is that it omits contextual factors such as trust, perceived risk and user capability, which are salient in e-government and online financial transactions [11], [12]. Extending TAM with domain-specific constructs is established practice, and Trust and CSE were chosen here for their relevance to mandatory, data-sensitive tax transactions and to populations with heterogeneous digital skills.

E-government adoption depends not only on usefulness and ease of use but also on trust in government and the internet, perceived risk and innovation characteristics [11], [17]. TAM-based tax studies show consistent patterns. Hung et al. found that PU, PEOU, perceived risk, trust, compatibility and self-efficacy determined acceptance of Taiwan's online tax filing system [2]. Wu and Chen integrated TAM, TPB and trust and confirmed the role of usefulness and trust [13]. Fu et al. reported that taxpayers were sensitive to the usefulness of a filing method [14]. In Malaysia, PU influenced intention to use online tax payment, with weaker effects of ease of use [15], [17], and system, information and service quality shaped continued use [16]. In Sri Lanka, Shantha found that PU, PEOU, perceived risk and facilitating conditions influenced intention to use e-tax filing [26]. Across these studies, PU is usually the dominant predictor, trust and perceived risk are decisive because the transaction is mandatory and financially sensitive, and individual capability shapes whether citizens feel able to complete the process unaided.

Trust covers confidence in the security and reliability of the website and in the institution operating it [11], [18], [19]; in Sri Lanka, trustworthiness significantly shapes use of e-government services [3]. Computer Self-Efficacy is an individual's belief in their ability to use a computer or online system to accomplish a task, rooted in Bandura's social cognitive theory [20], [21], and was a significant antecedent of intention in the Taiwanese tax context [2]. Finally, in TAM and related models, BI is the immediate antecedent of actual use [1], [7]. Accordingly, BI is positioned here as the immediate predictor of actual usage, operationalised as self-reported use of the IRD's online tax services.

### A. Conceptual Framework and Hypotheses

*Conceptual framework.* Drawing on TAM [1] and its extensions [2], [3], the framework models PU, PEOU, Trust and CSE as predictors of BI, which in turn predicts Citizen Acceptance (actual usage) (Fig. 1).

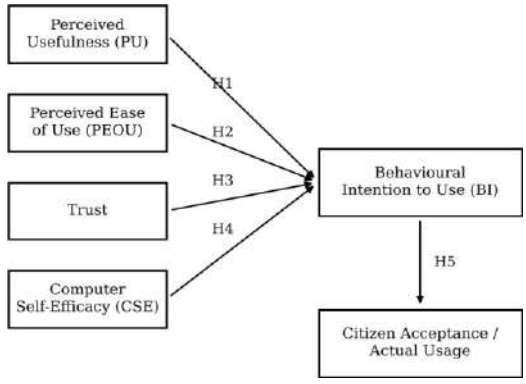



Fig. 1. Conceptual framework: Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Trust and Computer Self-Efficacy (CSE) predict Behavioural Intention (BI) (H1–H4), which predicts actual usage (H5).

*Hypotheses development.* Consistent with prior TAM-based tax research [1], [2], [13], the following hypotheses are proposed:

*H1:* PU positively affects BI to use IRD online tax services.

*H2:* PEOU positively affects BI to use IRD online tax services.

*H3:* Trust positively affects BI to use IRD online tax services.

*H4:* CSE positively affects BI to use IRD online tax services.

*H5:* BI positively affects actual usage of IRD online tax services.

## III. Research Methodology

The study adopted a quantitative, positivist and deductive approach, because it tests pre-defined hypotheses derived from an established theory (TAM) and measures relationships among operationalised constructs [22]. A cross-sectional survey design was used, collecting data at a single point in time with a standardised questionnaire.

*Population and sampling.* The target population comprised Sri Lankan citizens who (i) were registered with the IRD, (ii) had used or attempted to use the IRD's online tax services, and (iii) were eligible taxpayers. Because no complete sampling frame of online taxpayers is publicly available, probability sampling was not feasible, and non-probability convenience and purposive sampling was adopted, as in comparable e-government and e-tax studies [2], [26]. The questionnaire (Google Form) was circulated via e-mail, WhatsApp and social-media taxpayer groups. A screening question at the start retained only registered taxpayers who had encountered the online services. The final sample was N = 150 valid responses. This is smaller than conventional recommendations for regression with four predictors (for example, $N \geq 150 + 8k = 82$ [23]); the study is therefore treated as exploratory, and its findings as preliminary (see Limitations and Future Research). Convenience sampling also limits generalisability.

*Instrument and measures.* Each construct was measured with multiple items adapted from validated studies [1], [2], [3], [7], [18], [21] and reworded for the IRD context, using a five-point Likert scale from 1 ("strongly disagree") to 5 ("strongly agree") (Table 1). The questionnaire began with a consent statement and a demographic section, followed by one section per construct; the full item list is available from the authors on request. *Actual usage* was measured through self-reported questionnaire items on frequency and extent of use of the IRD's online tax services, because objective system logs were not available; this may introduce common-method and social-desirability bias (see Limitations and Future Research). Responses were exported to a spreadsheet, screened for incomplete or patterned answers, and coded numerically.

TABLE 1
QUESTIONNAIRE STRUCTURE AND ITEM SOURCES

| Construct | Items | Adapted from |
|---|---|---|
| Perceived Usefulness | 4–5 | Davis [1]; Hung et al. [2] |
| Perceived Ease of Use | 4–5 | Davis [1]; Hung et al. [2] |
| Trust | 4–5 | Ranaweera [3]; Bélanger & Carter [18] |
| Computer Self-Efficacy | 4–5 | Compeau & Higgins [21] |
| Behavioural Intention | 3–4 | Venkatesh et al. [7] |
| Actual Usage (self-reported) | 3–4 | Davis [1]; Hung et al. [2] |

*Data analysis.* Data were analysed in SPSS. First, descriptive statistics summarised respondent characteristics and construct distributions. Second, internal-consistency reliability was assessed with Cronbach's alpha, with 0.70 or above considered acceptable [24], [25]. Third, Pearson correlations examined bivariate associations among constructs. Fourth, multiple regression tested H1–H5: BI was regressed on PU, PEOU, Trust and CSE, and actual usage was regressed on BI. Multicollinearity was examined using variance inflation factors (VIF) and tolerance, with VIF below 5 taken as indicating no severe multicollinearity [27]. A significance level of 0.05 was applied throughout. Exploratory and confirmatory factor analysis and structural equation modelling were not conducted because the sample was too small for stable estimation; construct validity was therefore supported only by the use of previously validated scales and by reliability testing (see Limitations and Future Research). Mediation was not tested, so BI is treated as the direct predictor of actual usage and no mediation claims are made.

Participation was voluntary and based on informed consent. Responses were anonymised, used solely for academic purposes and stored securely. No sensitive identifying tax information was collected, and respondents could withdraw at any time.

## IV. Results and Discussion

*Respondent profile.* A total of 150 valid responses were obtained (Table 2). Most respondents were aged 25–44 (68%), held a Bachelor's degree or Diploma (88%), and worked in the private or self-employed sector or owned a business. Usage was infrequent: 24% reported never using the services and 52% only sometimes, consistent with the low adoption levels this study set out to explain.

TABLE 2
DEMOGRAPHIC PROFILE OF RESPONDENTS (N = 150)

| Characteristic | Category | n (%) |
|---|---|---|
| Age group | Under 25 | 5 (10%) |
| | 25–34 | 22 (44%) |
| | 35–44 | 12 (24%) |
| | 45–54 | 9 (18%) |

| Characteristic | Category | n (%) |
|---|---|---|
| | 55 and above | 2 (4%) |
| Gender | Male | 27 (54%) |
| | Female | 23 (46%) |
| Education | G.C.E. A/L | 5 (10%) |
| | Diploma | 18 (36%) |
| | Bachelor's degree | 26 (52%) |
| | Postgraduate | 1 (2%) |
| Occupation | Private sector | 18 (36%) |
| | Self-employed | 15 (30%) |
| | Business owner | 12 (24%) |
| | Government sector | 4 (8%) |
| | Other | 1 (2%) |
| Usage frequency | Never | 12 (24%) |
| | Rarely | 5 (10%) |
| | Sometimes | 26 (52%) |
| | Often | 6 (12%) |
| | Always | 1 (2%) |

*Reliability and descriptive statistics.* All six constructs exceeded the 0.70 threshold, indicating reliable scales (Table 3). All construct means were below the scale mid-point of 3.0, indicating generally low levels of perceived usefulness, ease of use, trust, self-efficacy, intention and usage. Mean PEOU (2.33) and PU (2.41) were the lowest.

TABLE 3
RELIABILITY AND DESCRIPTIVE STATISTICS (N = 150)

| Construct | Cronbach's α | Mean | SD |
|---|---|---|---|
| Perceived Usefulness (PU) | 0.86 | 2.41 | 0.69 |
| Perceived Ease of Use (PEOU) | 0.90 | 2.33 | 0.67 |
| Trust | 0.90 | 2.58 | 0.75 |
| Computer Self-Efficacy (CSE) | 0.86 | 2.56 | 0.66 |
| Behavioural Intention (BI) | 0.88 | 2.56 | 0.68 |
| Actual Usage (USE) | 0.80 | 2.68 | 0.65 |

*Correlation analysis.* All constructs were positively correlated (Table 4). The strongest associations were between CSE and BI (r = 0.85) and between BI and actual usage (r = 0.83). CSE was also strongly correlated with Trust (r = 0.81) and PU (r = 0.71), which motivated the multicollinearity diagnostics below.

TABLE 4
CORRELATION MATRIX (N = 150)

| | PU | PEOU | Trust | CSE | BI | USE |
|---|---|---|---|---|---|---|
| PU | 1.00 | | | | | |
| PEOU | 0.44 | 1.00 | | | | |
| Trust | 0.61 | 0.39 | 1.00 | | | |
| CSE | 0.71 | 0.47 | 0.81 | 1.00 | | |
| BI | 0.69 | 0.41 | 0.75 | 0.85 | 1.00 | |
| USE | 0.57 | 0.41 | 0.62 | 0.70 | 0.83 | 1.00 |

*Hypothesis testing.* Model 1 regressed BI on the four determinants and was significant, explaining 74.2% of the variance in BI ($R^2 = 0.742$, adjusted $R^2 = 0.719$, $F(4, 45) = 32.27$, $p < 0.001$). Model 2 regressed actual usage on BI ($R^2 = 0.684$). Table 5 ranks the effects from strongest to weakest. CSE was by far the strongest and the only significant determinant of BI (B = 0.600, $p < 0.001$). PU (B = 0.170, p = 0.123) and Trust (B = 0.157, p = 0.187) had small positive but non-significant effects, and PEOU had essentially no effect (B = −0.005, p = 0.955). BI was a strong and significant predictor of actual usage (B = 0.789, $p < 0.001$).

Multicollinearity diagnostics showed VIF values between 1.32 and 3.84 (tolerance 0.26–0.76), all below the cut-off of 5. Multicollinearity was therefore not severe, even though CSE and Trust were highly correlated (r = 0.81); the CSE finding is not an artefact of collinear predictors.

TABLE 5
REGRESSION RESULTS (N = 150)

| Path | B | p | VIF | Tol. | Result |
|---|---|---|---|---|---|
| CSE → BI | 0.600 | <0.001 | 3.84 | 0.26 | Significant |
| PU → BI | 0.170 | 0.123 | 2.09 | 0.48 | Not sig. |
| Trust → BI | 0.157 | 0.187 | 2.93 | 0.34 | Not sig. |
| PEOU → BI | −0.005 | 0.955 | 1.32 | 0.76 | Not sig. |
| BI → Usage | 0.789 | <0.001 | – | – | Significant |

*Summary of hypothesis testing.* Table 6 summarises the outcome of the hypotheses.

TABLE 6
SUMMARY OF HYPOTHESIS TESTING

| Hypothesis | Statement | Outcome |
|---|---|---|
| H1 | PU → BI | Not supported |
| H2 | PEOU → BI | Not supported |
| H3 | Trust → BI | Not supported |
| H4 | CSE → BI | Supported |
| H5 | BI → Actual Usage | Supported |

*Discussion.* The results confirm the central TAM proposition that BI drives actual usage (H5). Contrary to much

prior TAM research on tax systems [2], [13], [14], PU, PEOU and Trust did not significantly predict BI, while CSE emerged as the dominant determinant (H4). Because VIF values were acceptable, this is unlikely to be a purely statistical artefact of collinearity. Two other explanations are more plausible. First, CSE shared substantial variance with Trust and PU (r = 0.81 and 0.71), which suggests that, for these taxpayers, confidence in their own ability may be closely tied to how useful and trustworthy the system appears; the study design cannot separate these constructs, and discriminant validity could not be tested formally. Second, with N = 150 the regression has limited power to detect small effects, so the positive but non-significant coefficients for PU and Trust (B = 0.17 and 0.16) may reflect low power rather than true absence of effects. The near-zero PEOU effect is consistent with earlier tax studies in which ease of use was a weaker direct predictor than usefulness [15].

Substantively, for these taxpayers, confidence in their own ability to use the online system appears to be the decisive factor in forming the intention to use it. The uniformly low construct means indicate that acceptance remains limited. Given the small sample, these findings are best regarded as preliminary and indicative rather than conclusive.

## V. Limitations and Future Research

The main limitation is the small sample (N = 150), which is below conventional recommendations for regression with four predictors [23]. It limits statistical power and the stability of the estimates, so the results should be interpreted with caution and confirmed in a larger study. The cross-sectional design cannot establish causality or capture how acceptance evolves with experience. Non-probability convenience and purposive sampling with online distribution may under-represent older, rural or less digitally connected taxpayers and limits generalisability. Actual usage was self-reported and may diverge from objective system logs, with potential common-method and social-desirability bias. Exploratory and confirmatory factor analysis and structural equation modelling were not performed because of the sample size, so construct validity rests on validated scales and reliability evidence; discriminant validity between highly correlated constructs (notably CSE and Trust) remains untested. Mediation (indirect effects) was not tested; BI is treated as the direct predictor of actual usage. The model is bounded to four determinants and excludes others such as facilitating conditions, social influence, perceived risk and service quality. Future research should use larger probability samples, objective usage data and longitudinal designs; apply factor analysis and SEM; add the omitted constructs; and compare taxpayer segments, age groups and regions.

## VI. Conclusion

This study examined the factors affecting citizens' acceptance of the IRD's online tax services through an extended TAM, using survey data from 150 registered taxpayers. CSE was the strongest and only significant determinant of BI (B = 0.60), and BI strongly predicted actual usage (B = 0.79). PU, PEOU and Trust were not significant once CSE was included, and multicollinearity diagnostics did not indicate severe collinearity. The generally low construct means indicate that acceptance remains limited.

The implications for the IRD are drawn only from the significant finding. Because CSE drives intention, taxpayer education and support should be prioritised: step-by-step tutorials, help desks and assisted-filing facilities that build users' confidence in completing tax tasks online. This study does not establish whether usefulness, ease of use or trust interventions would raise acceptance; the low means for these constructs suggest they merit attention, but testing this requires larger samples. Because of the sample size, these recommendations should be treated as preliminary and confirmed in future research.

## References


[1] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quarterly 13 (3) (1989) 319–340. doi: 10.2307/249008.

[2] S.-Y. Hung, C.-M. Chang, T.-J. Yu, Determinants of user acceptance of the e-government services: The case of online tax filing and payment system, Government Information Quarterly 23 (1) (2006) 97–122. doi: 10.1016/j.giq.2005.11.005.

[3] H.M.B.P. Ranaweera, Perspective of trust towards e-government initiatives in Sri Lanka, SpringerPlus 5 (2016) 22. doi: 10.1186/s40064-015-1650-y.

[4] M. Fishbein, I. Ajzen, Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research, Addison-Wesley, Reading, MA, 1975.

[5] I. Ajzen, The theory of planned behavior, Organizational Behavior and Human Decision Processes 50 (2) (1991) 179–211. doi: 10.1016/0749-5978(91)90020-T.

[6] E.M. Rogers, Diffusion of Innovations, 5th ed., Free Press, New York, 2003.

[7] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: Toward a unified view, MIS Quarterly 27 (3) (2003) 425–478. doi: 10.2307/30036540.

[8] V. Venkatesh, J.Y.L. Thong, X. Xu, Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology, MIS Quarterly 36 (1) (2012) 157–178. doi: 10.2307/41410412.

[9] F.D. Davis, R.P. Bagozzi, P.R. Warshaw, User acceptance of computer technology: A comparison of two theoretical models, Management Science 35 (8) (1989) 982–1003. doi: 10.1287/mnsc.35.8.982.

[10] V. Venkatesh, F.D. Davis, A theoretical extension of the technology acceptance model: Four longitudinal field studies, Management Science 46 (2) (2000) 186–204. doi: 10.1287/mnsc.46.2.186.11926.

[11] L. Carter, F. Bélanger, The utilization of e-government services: Citizen trust, innovation and acceptance factors, Information Systems Journal 15 (1) (2005) 5–25. doi: 10.1111/j.1365-2575.2005.00183.x.

[12] D. Gefen, E. Karahanna, D.W. Straub, Trust and TAM in online shopping: An integrated model, MIS Quarterly 27 (1) (2003) 51–90. doi: 10.2307/30036519.

[13] I.-L. Wu, J.-L. Chen, An extension of trust and TAM model with TPB in the initial adoption of on-line tax: An empirical study, International Journal of Human-Computer Studies 62 (6) (2005) 784–808. doi: 10.1016/j.ijhcs.2005.03.003.

[14] J.-R. Fu, C.-K. Farn, W.-P. Chao, Acceptance of electronic tax filing: A study of taxpayer intentions, Information & Management 43 (1) (2006) 109–126. doi: 10.1016/j.im.2005.04.001.

[15] S. Anuar, R. Othman, Determinants of online tax payment system in Malaysia, International Journal of Public Information Systems 1 (2010) 17–32.

[16] C.-W. Chen, Impact of quality antecedents on taxpayer satisfaction with online tax-filing systems, Information & Management 47 (5–6) (2010) 308–315. doi: 10.1016/j.im.2010.06.005.

[17] O.K. Lean, S. Zailani, T. Ramayah, Y. Fernando, Factors influencing intention to use e-government services among citizens in Malaysia,

International Journal of Information Management 29 (6) (2009) 458–475. doi: 10.1016/j.ijinfomgt.2009.03.012.

[18] F. Bélanger, L. Carter, Trust and risk in e-government adoption, Journal of Strategic Information Systems 17 (2) (2008) 165–176. doi: 10.1016/j.jsis.2007.12.002.

[19] L.C. Schaupp, L. Carter, The impact of trust, risk and optimism bias on e-file adoption, Information Systems Frontiers 12 (3) (2010) 299–309. doi: 10.1007/s10796-008-9138-8.

[20] A. Bandura, Social Foundations of Thought and Action: A Social Cognitive Theory, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[21] D.R. Compeau, C.A. Higgins, Computer self-efficacy: Development of a measure and initial test, MIS Quarterly 19 (2) (1995) 189–211. doi: 10.2307/249688.

[22] U. Sekaran, R. Bougie, Research Methods for Business: A Skill-Building Approach, 7th ed., Wiley, Chichester, 2016.

[23] S.B. Green, How many subjects does it take to do a regression analysis, Multivariate Behavioral Research 26 (3) (1991) 499–510.

[24] J.C. Nunnally, Psychometric Theory, 2nd ed., McGraw-Hill, New York, 1978.

[25] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, Multivariate Data Analysis, 7th ed., Pearson, Upper Saddle River, NJ, 2010.

[26] K.V.A. Shantha, E-tax filing adoption of individuals in Sri Lanka, Wayamba Journal of Management 15 (1) (2024).

[27] R.M. O'Brien, A caution regarding rules of thumb for variance inflation factors, Quality & Quantity 41 (5) (2007) 673–690.

# UrbanSync: A Centralized Urban Management System for Sri Lanka

D. N. Pevinya
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
nilasipevinya@gmail.com

G. Perera
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
gayanp@nsbm.ac.lk

***Abstract*—Municipal complaint management in Sri Lanka remains fragmented across multiple government authorities, requiring citizens to identify the appropriate organisation when reporting urban service issues. This paper presents UrbanSync, a centralized urban complaint and service request management prototype developed for the Sri Lankan context. The system combines a React Native mobile application for citizens, a React.js administrative dashboard, and a Node.js RESTful backend supported by a MySQL database. Its main functionality is a deterministic routing mechanism that uses the issue category and GPS-derived administrative division to identify the relevant authority within the configured areas. The platform also incorporates complaint status tracking, role-based access control, multilingual support, notifications, and an AI-powered conversational assistant for submission guidance. The prototype was developed and tested in a local environment using selected administrative divisions and authorities in the Colombo and Gampaha districts. Evaluation included manual functional testing, automated testing, and User Acceptance Testing involving three participants representing the main system roles. The results indicate that the implemented functions operated as expected in the evaluated scenarios. However, the limited user sample and absence of live municipal deployment mean that further field evaluation is required before drawing conclusions about operational effectiveness, scalability, or wider public adoption.**

***Keywords—urban complaint management; e-governance; automated routing; civic engagement; smart cities.***

## I. INTRODUCTION

The rapid proliferation of mobile connectivity and cloud infrastructure has created tangible opportunities to modernize the relationship between citizens and municipal authorities. According to the United Nations E-Government Survey 2024, digital transformation at the local government level is now considered central to achieving inclusive public service delivery, particularly in fast-urbanizing developing nations [1]. Sri Lanka illustrates both the opportunity and the challenge: while Datar portal [2] reports smartphone penetration exceeding 60% of the population, urban public service management continues to depend on fragmented, manual reporting channels. Citizens encountering road damage, drainage failures, burst water mains, malfunctioning street lights, or irregular waste collection must independently determine which of several independent government authorities, the Municipal Council, National Water Supply and Drainage Board, Ceylon Electricity Board, or Road Development Authority, is responsible before reporting can even begin.

This structural fragmentation generates three compounding problems. First, complaint misdirection: submissions routed to the wrong authority are returned, delayed, or silently discarded. Second, resolution opacity: citizens receive no systematic feedback after submission, undermining trust in government responsiveness. Third, inter-agency coordination failure: a single incident crossing departmental boundaries requires multiple separate submissions with no coordinated resolution path.

International platforms demonstrate that these problems are manageable. Singapore's OneService application eliminates manual authority identification, routing submissions automatically to the responsible agency [3]. The United Kingdom's FixMyStreet integrates map-based location capture to improve routing precision [4]. India's Centralized Public Grievance Redress and Monitoring System (CPGRAMS) provides a unified submission portal across central government departments [5]. To the best of our knowledge, no comparable unified multi-authority complaint platform has been reported for Sri Lankan municipal governance.

This paper presents UrbanSync, a centralized urban complaint and service request management platform adapted to the Sri Lankan municipal context. Its main contribution is the integration of location-based and issue-category-based routing within a unified platform designed to connect citizens with relevant authorities. Rather than proposing a new routing algorithm, UrbanSync applies an established deterministic lookup approach using GPS-derived administrative divisions and predefined complaint categories. This approach is integrated with a citizen-facing mobile application, an administrative web dashboard, complaint lifecycle tracking, role-based access control, and multilingual support. The study demonstrates how these established technologies and design approaches can be combined and configured to address the fragmented nature of urban complaint management in the Sri Lankan context.

This study addresses four research questions: RQ1 (efficiency), whether centralised multi-category handling reduces the manual effort citizens need to report issues; RQ2 (misdirection), whether deterministic routing assigns complaints to the intended authority in the evaluated scenarios; RQ3 (transparency), whether status tracking and notifications

give citizens visibility of progress; and RQ4 (secure architecture), whether role-based access and layered authentication keep data segregated between user roles.

## II. Literature Review

### A. Evolution of Civic Reporting Systems

The foundational model for structured civic reporting is the 311 service system, which evolved from telephone-only channels to multi-modal digital platforms. Stowers [6] documents this trajectory across multiple U.S. cities, demonstrating that response times vary significantly by request category and confirming that structured categorization is a prerequisite for efficient routing. The United Kingdom's FixMyStreet extends this model through map-based geographic input; Matthews et al. [4] show that spatial visualization substantially improves reporting precision, enabling authorities to locate and address issues more rapidly than text-description-only systems.

### B. Integrated Multi-Authority Platforms

Singapore's OneService application represents the most advanced publicly deployed example of automated, multi-authority routing. The system accepts submissions without requiring citizens to identify the responsible agency, routes them automatically, and provides progress updates throughout resolution [3]. The Ministry of Social and Family Development [7] additionally documents a conversational chatbot that guides users through reporting, improving accessibility for those uncertain how to categories their issues. India's CPGRAMS demonstrates scalability of centralized grievance platforms; Nakkina and Veeraju [5] show that centralization improves accountability and creates measurable audit trails, though breadth can introduce processing delays.

### C. Accessibility and the Digital Divide

Research on developing-country e-governance consistently identifies a tension between technical capability and actual adoption. Dahri et al. [8] evaluate Pakistan's citizen portal and find that registration complexity and navigation barriers suppress usage among populations with limited digital experience. Raza [9] elaborates on structural factors including low digital literacy and unequal internet access, concluding that technical sophistication alone does not guarantee uptake. These findings reinforce the importance of user-centred design: UrbanSync adopts a guided, category-first submission flow and multilingual support (English, Sinhala, and Tamil) to lower the participation barrier for less digitally experienced citizens.

### D. Intelligent Assistance and Alternative Channels

Shah et al. [10] document how citizens increasingly resort to microblogging platforms to express municipal grievances due to platform familiarity. However, social media channels lack structured routing, accountability mechanisms, and enforcement pathways, meaning reports rarely lead to systematic resolution. Zhou et al. [11] examine chatbot-based 311 platforms and identify three critical success factors: transparency in system behaviour, consistent accuracy, and contextual understanding. UrbanSync incorporates an OpenAI API-powered conversational assistant as a supplementary guidance layer but retains a deterministic routing engine as the primary authority assignment mechanism to avoid dependency on probabilistic outputs for compliance-critical operations.

### E. The Sri Lankan Context and Research Gap

Within Sri Lanka, digital transformation has occurred primarily in isolated commercial sectors, with minimal cross-agency integration in the public service domain. The Ceylon Electricity Board operates its own complaint portal [12]; the National Water Supply and Drainage Board and Municipal Councils manage separate channels. A citizen encountering a burst pipe causing road damage and power disruption must file three independent reports across three separate platforms. The World Bank [13] projects continued urban population growth in Sri Lanka, which will intensify service demand and amplify the cost of this fragmentation. This comparative analysis reveals a clear research gap: Sri Lanka lacks a unified, multi-authority complaint platform with automated routing. UrbanSync fills this gap by adapting the design principles of OneService and FixMyStreet to the specific administrative, linguistic, and connectivity conditions of Sri Lankan municipal governance.

### F. E-Government Maturity Perspective

The development of UrbanSync is also considered in relation to the four-stage e-government model proposed by Layne and Lee [14], which describes the progression from online information provision (cataloguing) to electronic service transactions, vertical integration, and horizontal integration across government functions. A staged view of online service maturity is also used in the UN E-Government Survey [15]. UrbanSync primarily supports the transactional stage by allowing citizens to submit complaints digitally and monitor their progress. Its centralized design also introduces elements of cross-agency service integration through predefined authority mappings and role-based complaint management. However, these integrations are implemented within the prototype configuration and do not represent live interoperability between government information systems. The model is therefore used as a theoretical design reference rather than as a formal assessment of government maturity.

## III. Methodology

### A. Research Design

This study adopts an applied engineering design approach, combining requirement validation, iterative system development, and multi-phase evaluation. The research objective is constructive: to design, build, and validate a practical artifact that addresses a documented deficiency in Sri Lankan municipal service delivery. A mixed-methods approach was employed: quantitative survey data validated key design assumptions, while informal requirement-gathering discussions with stakeholders informed the system design.

UrbanSync was developed and evaluated as a functional software prototype using a local development environment. The implementation focused on selected administrative divisions and government authorities within the Colombo and Gampaha districts of Sri Lanka, rather than nationwide coverage. The prototype consists of a React Native mobile application for citizens, a React-based web dashboard for authority officers and system administrators, and a centralized backend and database. The routing configuration was populated according to the

selected geographic areas and authority mappings. The system was not deployed on a public hosting environment or integrated into live government infrastructure. Therefore, the evaluation focused on the functionality and usability of the implemented prototype within its configured scope, rather than measuring its performance or effectiveness in operational municipal environments.

### *B. Requirement Validation Survey*

A structured survey was conducted using Google Forms, yielding 30 valid responses from Sri Lankan urban residents. All 30 respondents reported regular smartphone usage, validating the mobile-first strategy. Although 56.7% had previously submitted complaints, only 43.3% clearly understood the submission process, confirming that engagement does not imply procedural clarity. The most frequently cited barriers were resolution delays (41.7%), absence of any response (33.3%), and uncertainty about whether complaints reached the correct authority (25.0%).

Furthermore, 70% of respondents expressed willingness to adopt a centralised digital platform, with expected benefits including faster resolution (43.3%) and increased public participation (40.0%). These findings directly informed three core system requirements: simplifying the submission process, ensuring accurate complaint routing, and providing transparent real-time status updates. As detailed information on participant recruitment and geographic distribution was not formally recorded, the survey findings are considered indicative of the 30 responses received rather than representative of the wider Sri Lankan urban population.

### *C. Development Methodology*

System development followed an Agile iterative process, chosen for its capacity to accommodate evolving requirements and validate functionality continuously. The initial phase produced a minimal viable prototype validating user input handling and database connectivity. Subsequent sprints introduced complaint categorization, the routing engine, real-time tracking, and the administrative dashboard.

### *D. Evaluation Strategy*

Evaluation was conducted across three complementary phases. Manual functional testing validated all core workflows against defined test cases with specified inputs and acceptance criteria. Automated backend testing using Jest and Supertest with mocked database interactions validated API endpoints, routing logic, and data consistency. Automated frontend testing using Vitest and React Native Testing Library verified component rendering, navigation flow, and role-based view differentiation. User Acceptance Testing (UAT) engaged three participants representing the Citizen, Authority Officer, and Super Administrator roles, each completing predefined task scripts in a controlled environment followed by structured feedback collection.

## IV. Proposed System Design

### *A. System Architecture*

UrbanSync adopts a three-tier client-server architecture comprising a presentation layer, application layer, and data layer. The presentation layer consists of a React Native / Expo mobile application serving the citizen persona across Android and iOS, and a React.js (Vite) web application serving Authority Officers and Super Administrators. The application layer is a Node.js / Express RESTful API handling all business logic, including routing decisions, authentication, and notification dispatch. In the prototype, inter-tier communication ran over HTTP on a local network; HTTPS would be required for any deployment. The data layer is a normalized MySQL relational database storing user profiles, complaint records, authorities linked to departments and administrative divisions, and notifications. Third Normal Form (3NF) normalization reduces redundancy and supports the join-based queries required by the routing engine.

### *B. Automated Routing Engine*

UrbanSync implements a deterministic, two-dimensional routing mechanism that combines issue category and administrative location to identify the responsible authority. When a citizen submits a complaint, the backend uses two parameters: (1) Issue Type, selected by the citizen from a predefined list grouped into categories, where each issue type is mapped to a responsible department (for example local councils, the Public Health Inspector, the Police, the Water Board, or the Ceylon Electricity Board); and (2) Administrative Division, obtained from the selected GPS location: the mobile application reverse-geocodes the coordinates into an address, and the backend matches this address text against the configured division names. The combination of department and division is then looked up in the authorities table. If no authority is configured for the division, the system falls back to an authority of the same department within the same district; if none is found, the complaint is left unassigned and appears in a queue for the Super Administrator to reassign manually. The authority records (department and division) are configurable by administrators through the dashboard, enabling adaptation to boundary changes without code modifications. This deterministic, rule-based design was deliberately chosen over probabilistic classification to support auditability in a governance context. The lookup follows an established rule-based pattern; the contribution lies in its configurable application to multi-authority Sri Lankan municipal administration.

### *C. Functional Components*

The system is decomposed into six functional modules: (1) User Management, handling registration, login with a Firebase SMS OTP second factor, JWT session management, and RBAC; (2) Complaint Submission, guiding citizens through category selection, description entry, image attachment, and GPS-based location input, and rejecting a new submission of the same category within 50 meters of an unresolved complaint as a duplicate; (3) Automated Routing Engine, executing the two-dimensional authority lookup with district-level fallback; (4) Complaint Lifecycle Management, managing status updates (Pending → In Progress → Resolved, with Rejected and Cancelled states); (5) Notification Module, creating an in-app notification for the citizen on each status update; and (6) Administrative Dashboard, providing Authority Officers with jurisdiction-filtered complaints and Super Administrators with cross-authority monitoring and reassignment tools.

### D. Comparison with Existing Systems

Table I compares UrbanSync with three internationally deployed platforms across six features: automated authority routing, multi-domain handling, GPS location capture, real-time tracking, image-based submission, and AI conversational guidance. UrbanSync combines these features within a single platform. The comparison reflects the features available in each system and does not imply technical superiority, as the platforms differ in scale, maturity, and deployment context, and UrbanSync is evaluated only as a prototype.

## V. System Implementation

### A. Technology Stack

React Native was selected for the citizen mobile application to support both Android and iOS from a single codebase, reducing maintenance overhead. Node.js was chosen for its non-blocking I/O model, which scales efficiently for concurrent complaint submissions. MySQL was preferred over a document store because the relational schema enforces referential integrity between complaints, users, authorities, and routing mappings. Firebase Authentication provides SMS OTP two-factor verification; Google Maps SDK enables interactive location selection; and OpenAI API powers the conversational complaint guidance assistant. Automated backend tests used Jest and Supertest with mocked database interactions; frontend tests used Vitest and React Native Testing Library.

TABLE I. Feature Comparison with Existing Civic Platforms

| Feature | CPGRAMS | FixMyStreet | OneService (SG) | UrbanSync |
|---|---|---|---|---|
| Automated routing | No | Partial | Yes | Yes |
| Multi-domain handling | Yes | No | Yes | Yes |
| GPS location capture | No | Yes | Yes | Yes |
| Real-time tracking | Partial | Partial | Yes | Yes |
| Image-based submission | No | Yes | Yes | Yes |
| AI conversational guide | No | No | Yes | Yes |

### B. Mobile Application

The citizen-facing mobile application provides six primary screens: Authentication (registration, and login with an SMS OTP second factor), Complaint Submission (category selection, description, image capture, map-based location pin), My Complaints (submission history with lifecycle status badges), Complaint Detail (status timeline), Notifications (in-app notification inbox), and AI Assistant (conversational guidance chatbot). The application supports English, Sinhala, and Tamil through static translation files loaded at runtime, covering all UI labels.

### C. Administrative Web Dashboard

The web dashboard renders two distinct views determined by the authenticated user's role. Authority Officers see a filtered complaints table showing only submissions routed to their jurisdiction, a detail panel with submitted evidence and description, and a status-update control. Super Administrators additionally see a cross-authority analytics overview, a complaint and service request reassignment control for edge cases, and user management tools. Role enforcement is applied both at the API layer (JWT claim validation) and at the React component level (conditional rendering) to prevent unauthorized data exposure.

The prototype incorporates security controls including JWT-based authentication, role-based access control, and Firebase SMS OTP verification. Location information is used as part of the complaint-routing process, while access to complaint information is restricted according to the authenticated user's role. These controls were implemented within the prototype; however, formal data-retention policies and production-level security measures were outside the scope of the current evaluation.

## VI. Results and Evaluation

### A. Functional and Automated Testing

Manual functional testing was conducted using 20 predefined test cases covering the main system workflows, including user registration and authentication, OTP verification, complaint submission with image and GPS, authority assignment, complaint status updates, authority and user management, notifications, and complaint history. All 20 documented test cases were marked as passed against their specified expected outcomes. Invalid inputs, including complaints submitted without GPS coordinates or with unrecognized category codes, returned well-formed HTTP 400 error responses without corrupting database state. The test cases included scenarios for normal authority assignment and fallback authority handling, which provided functional evidence that the configured routing behaviour operated as expected in the evaluated scenarios. Further verification of routing coverage across all category–division combinations is required before making broader claims about routing accuracy. Frontend component tests confirmed that role-specific views rendered exclusively the data appropriate to each user persona.

### B. User Acceptance Testing

User Acceptance Testing involved three participants, with one participant representing each of the Citizen, Authority Officer, and Super Administrator roles. Each participant completed predefined task scripts in a controlled environment. All three participants completed their assigned tasks without developer assistance. Their feedback provided initial insights into the usability of the prototype's role-specific functions. The Citizen-role participant found location selection and status tracking intuitive. The Authority Officer participant observed that only jurisdiction-relevant complaints appeared in their dashboard view. The Super Administrator participant identified cross-authority visibility and the complaint reassignment tool as particularly valuable for governance oversight. No critical usability issues were reported by the participants. However, the small sample size limits the extent to which these findings can be generalized to the wider citizen population or municipal workforce.

### C. Objective and Research Question Evaluation

Table II maps each project objective to its outcome within the prototype scope. Five of the six objectives were achieved, and Objective 6 (centralized analytics) was partially achieved: while the schema and dashboard aggregate views are in place, meaningful trend analysis requires a larger collection of real-world complaints than can be generated in a controlled environment.

TABLE I. EVALUATION AGAINST PROJECT OBJECTIVES

| Objective | Implementation Outcome | Status |
|---|---|---|
| 1 - Mobile App | GPS-based location capture and image-based submission implemented. | Achieved |
| 2 - Routing | Two-dimensional routing engine (division × category); expected authority assignments in the evaluated routing scenarios. | Achieved |
| 3 - Transparency | Status tracking and in-app notifications at each lifecycle stage. | Achieved |
| 4 - Security | JWT access tokens, RBAC, and Firebase SMS OTP two-factor authentication. | Achieved |
| 5 - Dashboard | Role-differentiated web dashboard for Authority Officers and Super Administrators. | Achieved |
| 6 - Analytics | Normalized schema and aggregate dashboard views implemented; full trend analysis deferred to live deployment datasets. | Partially achieved |

The evaluation provided preliminary evidence addressing the four research questions within the implemented prototype scope. RQ1 was addressed through the integration of complaint submission, category-based handling, and automated authority assignment. RQ2 was examined through the configured routing test scenarios, which produced the expected assignments in the evaluated cases. RQ3 was addressed through complaint status tracking and notification functionality, supported by feedback from the Citizen-role UAT participant. RQ4 was examined through the implementation of JWT authentication, role-based access control, and Firebase SMS OTP verification. These findings demonstrate the implementation of the intended system functions but do not constitute a full evaluation of operational efficiency, population-level usability, or production-level security.

### D. Evaluation Limitations

Several limitations should be considered when interpreting the evaluation results. First, the prototype was developed and tested in a local environment and was not deployed within a live municipal service infrastructure. Its routing configuration covered selected administrative divisions and authorities in the Colombo and Gampaha districts rather than the entire country. Second, User Acceptance Testing involved only three participants, one for each system role, and therefore does not provide sufficient evidence for population-level usability conclusions. Third, the testing environment did not represent all conditions associated with real-world GPS usage, mobile hardware, network availability, and interactions between government agencies; hardware-dependent features such as live GPS and camera were partially simulated in automated tests. Fourth, the evaluation did not establish production-level performance, scalability, security, or the effectiveness of the AI assistant through dedicated quantitative measurements. Finally, the routing mechanism depends on the accuracy and completeness of the configured administrative mappings. These limitations should be addressed through broader testing and evaluation in future work.

## VII. DISCUSSION

The evaluation provides initial evidence that UrbanSync's deterministic, two-dimensional routing mechanism can assign complaints to the configured authorities based on issue category and administrative location. The prototype also integrates complaint submission, status tracking, role-based access, and multilingual support within a single platform. These functions address several limitations identified in existing civic reporting approaches, particularly the need for citizens to identify the responsible authority independently. However, the findings are limited to the configured prototype environment and the scenarios covered by the testing process. They should not be interpreted as evidence of improved municipal processing efficiency or verified performance under real-world operational conditions. A recurring theme in the literature is that platforms addressing only one dimension of the complaint lifecycle consistently fail to resolve systemic fragmentation [4], [5], [10]. CPGRAMS provides centralized submission but relies on manual inter-departmental routing, introducing delays at scale [5]. FixMyStreet improves location precision but serves only single-authority local councils [4]. Social media channels provide accessibility but lack accountability [10]. UrbanSync combines submission, automated routing, real-time tracking, and role-based multi-authority access in a single prototype, addressing all three dimensions within its configured scope.

The routing scenarios evaluated during testing produced the expected authority assignments within the configured prototype. This provides preliminary evidence that the category-and-location-based lookup can support authority identification without requiring citizens to select the responsible organization manually. However, the limited number of evaluated scenarios and the absence of live municipal deployment prevent conclusions about routing accuracy across all possible geographic boundaries, complaint categories, and operational conditions. The 25% of survey respondents who reported uncertainty about whether their complaints reached the correct authority further highlights the practical problem that the system aims to address. The conversational AI assistant is deliberately confined to category selection guidance rather than routing, ensuring that chatbot capability enhances usability without introducing the opacity and inconsistency risks identified by Zhou et al. [11] as central barriers to citizen trust. The multilingual implementation directly addresses the accessibility concerns raised by Dahri et al. [8] and Raza [9], ensuring all system labels and category names are available in English, Sinhala, and Tamil.

UrbanSync provides a prototype-level reference architecture for integrating complaint and request service submission, deterministic authority routing, lifecycle tracking, and role-based access within a multi-authority urban service management platform. Its design may inform similar initiatives

in other contexts, although further evaluation is required before its suitability for wider operational deployment can be established.

## VIII. Conclusion and Future Work

This paper presented UrbanSync, a centralized urban complaint and service request management prototype developed for the Sri Lankan municipal context. The system integrates a citizen-facing mobile application, an administrative web dashboard, and a centralized backend to support complaint submission, authority assignment, status tracking, and role-based access. Its routing mechanism uses the combination of issue category and GPS-derived administrative division to identify the relevant authority within the configured areas.

The prototype was evaluated through manual functional testing, automated testing, and User Acceptance Testing involving three participants representing the main system roles. The documented manual test cases demonstrated that the tested workflows operated as expected, while the UAT provided initial feedback on the usability of the role-specific functions. These findings indicate that the proposed system design is feasible within the implemented prototype scope. However, the limited UAT sample, selected geographic coverage, and absence of live municipal deployment mean that further evaluation is necessary before drawing conclusions about its effectiveness in operational environments.

Several directions for future work are identified: (1) a field evaluation in a live municipal deployment with a larger and more diverse participant group, including citizens and municipal officers across different administrative divisions; (2) validation of GPS boundary accuracy and routing coverage; (3) performance, scalability, and security evaluation, including data-protection policies; (4) evaluation of the AI assistant's response quality; (5) multimodal duplicate complaint detection using NLP and computer vision; (6) offline synchronisation capability for areas with intermittent connectivity; (7) integration with the Department of Registration of Persons NIC API to verify citizen identity at registration; and (8) predictive infrastructure analytics using historical complaint data to support proactive maintenance scheduling. A practical route to adoption would be a pilot with a single municipal council, using the administrator-configurable routing table to add further authorities and divisions. UrbanSync suggests that design principles used in mature civic platforms such as OneService and FixMyStreet can be adapted to Sri Lankan municipal conditions, subject to the further evaluation described above.

## Acknowledgment

D. N. Pevinya thanks supervisor Mr. Gayan Perera, NSBM Green University, for his continuous guidance, constructive feedback, and academic support throughout the research and development of this project.

## References

[1] United Nations Department of Economic and Social Affairs, "E-Government survey 2024: Accelerating digital transformation for sustainable development," UN DESA, New York, 2024.

[2] DataReportal, "Digital 2024: Sri Lanka," 2024. [Online]. Available: https://datareportal.com/reports/digital-2024-sri-lanka. Accessed: Nov. 1, 2026.

[3] Municipal Services Office, "OneService app and chatbot," Government of Singapore, 2026. [Online]. Available: https://www.oneservice.gov.sg/about-us/vision-and-mission. Accessed: Nov. 1, 2026.

[4] P. Matthews, A. Rae, and E. Nyanzu, "FixMyStreet!: The geography of citizen reporting neighbourhood issues in the UK," University of Stirling and University of Sheffield, Stirling, 2018.

[5] I. D. P. Nakkina and G. Veeraju, "Strengthening grievance redressal systems as a critical component of good governance: The case study of CPGRAMS," Educational Administration: Theory and Practice, vol. 30, no. 5, pp. 15613–15624, 2024.

[6] G. N. L. Stowers, "Back to basics: City services and 311 service requests," State and Local Government Review, vol. 54, no. 1, pp. 13–31, 2022.

[7] Ministry of Social and Family Development, "Developing the OneService chatbot to improve municipal services," Government of Singapore, 2021. [Online]. Available: https://www.tech.gov.sg/technews/developing-the-one-service-chatbot/. Accessed: Nov. 1, 2026.

[8] S. Dahri, S. Massan, and A. A. Maitlo, "Usability of eGovernance application for citizens of Pakistan," 3C Tecnología, Special Issue, pp. 265–277, 2020.

[9] Raza, "Digital divide and e-governance in Pakistan: Factors affecting the citizen portal app Pakistan," MPhil thesis, Pakistan Institute of Development Economics, Islamabad, 2020.

[10] K. Shah, H. Joshi, and H. Joshi, "Smart approach to recognize public grievance from microblogs," Towards Excellence, vol. 13, no. 2, 2021.

[11] J. Zhou, R. Shen, Y. You, C. DiSalvo, L. Dombrowski, and C. MacLellan, "Improving public service chatbot design and civic impact: Investigation of citizens' perceptions of a metro city 311 chatbot," in Proc. Designing Interactive Systems Conf. (DIS '25), 2025.

[12] Ceylon Electricity Board, "CEB care: Submit a complaint," [Online]. Available: https://edlcare.edl.lk/incognito/new-complain. Accessed: Nov. 1, 2026.

[13] World Bank, "Urban population (% of total population) – Sri Lanka," 2024. [Online]. Available: https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?. Accessed: Nov. 1, 2026.

[14] K. Layne and J. Lee, "Developing fully functional E-government: A four stage model," Government Information Quarterly, vol. 18, no. 2, pp. 122–136, 2001, doi: 10.1016/S0740-624X(01)00066-1.

[15] United Nations Department of Economic and Social Affairs, "E-Government survey 2022: The future of digital government," UN DESA, New York, 2022.

# Transparency as a Substitute for Fairness? Organizational Justice and System Trust in Algorithmic Human Resource Management

L.S.V. Samarappulige
*Dept. of Computing and Info. Systems*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
larlithasrividhara@gmail.com

L.S. Lekamge
*Dept. of Computing and Info. Systems*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
slekamge@foc.sab.ac.lk

A.S.A. Perera
*Dept. of Computing and Info. Systems*
*Sabaragamuwa University of Sri Lanka*
*Belihuloya, Sri Lanka*
anupama@foc.sab.ac.lk

***Abstract*—Algorithmic promotion tools raise a question that technology-acceptance models rarely ask, namely whether employees trust such a system because they judge it fair or because it can explain itself. This study analyzes a cross-sectional survey of 421 information technology professionals in Sri Lanka, comprising 251 software engineers, 107 human resource professionals and 63 executives, who evaluated a hypothetical algorithmic promotion to tech lead. Perceived procedural fairness was low overall, with a mean of 2.85 on a five-point scale, and lowest among engineers at 2.26. Job role explained 39.1 percent of its variance. Demand for explainability did not differ by role. Anticipated trust in the system was nonetheless moderately high, with a mean of 3.61. In regression, explainability demand was the strongest correlate of trust, with a standardized coefficient of 0.48, about three to four times that of perceived fairness at 0.13, and it carried roughly 69 percent of the explained variance. The pattern held with controls, with heteroscedasticity-robust errors, with a trust measure free of explanation-related items, and after removing 17 careless responders. Explainability demand did not moderate the association between fairness and trust, so the pattern is additive, with trust tracking explanation demand largely independently of fairness beliefs and no sign that explanation compensates for low fairness. Because the design is cross-sectional, self-reported and hypothetical, causal and behavioral conclusions are out of reach. We discuss implications for the procurement of human resource information systems and for keeping human review and fairness audits independent of explanation features.**

***Index Terms*—algorithmic management, digital business, explainable artificial intelligence, human resource information systems, organizational justice, technology acceptance**

## I. Introduction

Human resource (HR) management has become an analytics-intensive function, and enterprise human resource information systems (HRIS) increasingly use artificial intelligence (AI) to score, rank and recommend personnel from historical data [1]. Published work on algorithmic HR concentrates at the boundaries of employment, namely resume screening and attrition prediction [2], [3]. Promotion, where authority, pay and standing are settled, has received comparatively little attention [1], [4].

The gap matters for system adoption. A rejected applicant leaves, but a passed-over employee stays, and perceived procedural fairness shapes how that employee reacts to the evaluation system [5]. If an opaque model supplies the score, staff may read it as "algorithmic reductionism," the belief that quantification strips out context a human evaluator would weigh [6].

Vendors and policymakers commonly offer explainable AI (XAI) as the remedy [7]. The implicit chain runs from transparency to perceived fairness to trust and adoption. To our knowledge that chain has seldom been tested with the employees whose careers such systems would govern, and HR XAI applications such as [3] address HR analysts rather than evaluated staff. This paper asks a narrower question: does trust in an AI promotion system track perceived fairness, or does it track the wish for explanation?

The responses come from one survey of 421 information technology (IT) professionals in Sri Lanka. A companion design-science paper first reported the sample, instrument and measurement validation, and used the same survey to specify and expert-evaluate an explainable-AI architecture [4]. Beyond [4], which analyzed all 421 responses without screening for careless responding, this paper makes three contributions: it tests directly whether explanation demand compensates for low perceived fairness, audits the measurement and common-method threats to the explanation–trust link, and shows how the pattern changes under controls, robust estimation and removal of careless responders. Two research questions guide the work:

1) **RQ1:** How do perceived procedural fairness, demand for governance and demand for explainability differ across organizational roles?
2) **RQ2:** How strongly is explainability demand associated with anticipated system trust relative to perceived fairness, and does explainability demand moderate the fairness–trust association?

## II. Theoretical Background

### A. Procedural and Informational Justice

Organizational justice research separates distributive, procedural, interpersonal and informational justice [5]. Procedural justice follows Leventhal's rules, which include consistent application, bias suppression, accurate information, correctability and representation of affected parties [8]. Informational justice asks whether a procedure is explained candidly. Explanations improve justice judgments and soften negative reactions [9], yet within this framework they form a

separate dimension. They are not proof that the procedure is fair. That distinction motivates the study: an employee may want an explanation, and be reassured by one, while still doubting that the underlying procedure is fair.

### *B. Trust, Acceptance and the Transparency Ideal*

The Technology Acceptance Model [10] and the Unified Theory of Acceptance and Use of Technology [11] explain adoption intention through usefulness, effort expectancy and social and organizational conditions. For consequential automated decisions, trust, defined as willingness to be vulnerable to another party's actions [12], and its calibration [13] become central. Workers often judge algorithmic personnel decisions less fair than human ones, particularly when the task seems to need human judgment [14]. Critical work cautions that visibility is not accountability [15], that explanations can raise acceptance of AI advice whether or not the advice is correct [16], that explaining a black box differs from using an interpretable model [17], and that automation can give biased processes a veneer of objectivity [18]. We treat trust as a precursor of adoption intention, not as intention itself, so the study concerns trust and reliance readiness, not adoption behavior.

### *C. Organizational Roles*

Algorithmic management encodes managerial control in software [19], so those who configure a promotion tool and those it evaluates need not share expectations. We therefore compare roles and ask whether any divergence extends to demand for explanation.

## III. METHODOLOGY

### *A. Design, Sample and Ethics*

The study is a cross-sectional, purely quantitative survey. The expert evaluation reported in [4] is a separate qualitative phase and is not part of this study. Respondents completed an online questionnaire (Google Forms) between 20 April and 18 July 2026, recruited by purposive sampling with a snowball element. Purposive sampling suits a question about three specific roles in promotion decisions, and snowballing reaches the smaller HR and executive populations. All items referred to one vignette, the promotion of a senior software engineer to tech lead, and the IT sector offers a career step familiar to all three roles. A pilot with 25 respondents was excluded. Of 450 responses, 29 that chose "Other" as job role were removed, leaving $N$ = 421: 251 software engineers (59.6%), 107 HR professionals (25.4%) and 63 executives (15.0%), with no missing values (Table I).

A consent statement preceded the questionnaire, participation was voluntary and anonymous, and no personally identifiable data were collected. On that basis the study was treated as exempt from formal institutional ethics review. Age and gender were not collected, which limits demographic description and control.

### *B. Instrument and Constructs*

Twenty-one five-point Likert items (1 = strongly disagree or not useful, 5 = strongly agree or must have) form five constructs. The items were developed from the justice, trust and explainable-AI literature and prior empirical evidence, each justified during questionnaire design (construct bases in [4]), not adapted from existing validated scales. Construct scores are unweighted item means (M, with standard deviation, SD) after reverse-coding (R).

- **A, perceived procedural fairness** (4 items; 2 R): AI is more fair and unbiased than a human manager; AI should pick the top candidates to stop managers playing favorites; AI cannot understand the "human skills" a tech lead needs (R); worry that AI will punish unusual career paths (R).
- **B, demand for governance** (3 items): regular bias checks; HR, not the AI, is to blame for bad choices; strict policies on AI use in promotion.
- **C, demand for explainability** (5 items): five explanation formats described in text: global documentation, a personal factor breakdown (local attribution), a "what-if" tip (counterfactual), a chart comparing the employee with the last ten promoted, and a short plain-language paragraph (C5). No mock-ups or model output were shown, so C captures stated demand for formats, not comprehension of real explanations.
- **D, employee-centered orientation** (5 items; 2 R): expert explanations are confusing (R); reduction to numbers hurts feelings (R); a clear AI reason raises company trust; a right to human review; a choice of explanation detail.
- **E, anticipated trust and reliance** (4 items): trust AI to review my own career fairly (item 30); clear explanations are the number one reason I would trust AI (31); prefer an AI that is fully understandable even if it makes slightly more mistakes (32); using AI fairly will eventually make employees happier (33).

Four concepts are kept apart throughout. A measures *perceived* procedural fairness, not actual algorithmic fairness such as parity metrics, which no item assesses. C measures *demand* for explanation, not actual system transparency or users' understanding of it. E measures anticipated trust and reliance; no item asks about intended use, so it captures readiness to rely on the system, not adoption intention or behavior. Two E items (31 and 32) are explanation-contingent. For that reason a sub-index of the other two items (30 and 33), called explanation-independent trust, serves as an alternative outcome.

### *C. Measurement Quality and Common-Method Bias*

Cronbach's $\alpha$ [20] was 0.841, 0.789, 0.835, 0.801 and 0.794 for A to E, and composite reliability was within 0.001 of these values. A five-factor confirmatory factor analysis (CFA; maximum likelihood, 21 items) fitted well, $\chi 2(179) = 320.2$, comparative fit index (CFI) = 0.956, root mean square error of approximation (RMSEA) = 0.043, standardized root mean square residual (SRMR) = 0.052 [21], with standardized loadings of 0.54–0.79 (details in [4], re-estimated independently here). Average variance extracted (AVE) was 0.569, 0.555 and 0.504 for A, B and C, but 0.450 and 0.491 for D and E; with composite reliability above 0.70, convergent validity for D and E is adequate but not strong [22]. Heterotrait–monotrait ratios were at most 0.691 (threshold 0.85 [23]), and the square root of AVE exceeded every latent correlation, only narrowly for C and E (0.710 and 0.701 versus 0.691). All data came from one instrument, so four common-method checks were run [24]. In

Harman's test the first unrotated component explained 26.1% of the variance of the 21 construct items. Full-collinearity variance inflation factors (VIF) [25] ranged from 1.058 to 1.585, below 3.3. Two approximate marker variables, enterprise size and the rated importance of hours worked as a promotion input, correlated near zero with trust ($r = 0.033$ and $-0.011$; $p \geq 0.499$); adjusting for the smaller marker value [26] left trust correlations of 0.558 (C), 0.323 (B), 0.265 (D) and 0.099 (A, $p = 0.043$). Finally, adding an orthogonal latent method factor on all 21 items improved fit, $\chi 2(158) = 168.6$, yet it carried only 7.8% of mean item variance against 46.1% for the traits, and the latent C–E correlation moved from 0.691 to 0.638. Method bias cannot be excluded in single-source data, but these checks do not point to it as the source of the main pattern.

TABLE I

SAMPLE CHARACTERISTICS BY ORGANIZATIONAL ROLE, *n* (COLUMN %)

| **Characteristic** | **Engineers** ($n = 251$) | **HR** ($n = 107$) | **Executives** ($n = 63$) | **All** ($N = 421$) |
|---|---|---|---|---|
| *Professional experience* | | | | |
| 0–2 years | 52 (20.7%) | 0 (0.0%) | 0 (0.0%) | 52 (12.4%) |
| 3–5 years | 76 (30.3%) | 20 (18.7%) | 0 (0.0%) | 96 (22.8%) |
| 6–10 years | 75 (29.9%) | 45 (42.1%) | 5 (7.9%) | 125 (29.7%) |
| 11–15 years | 39 (15.5%) | 32 (29.9%) | 30 (47.6%) | 101 (24.0%) |
| Over 15 years | 9 (3.6%) | 10 (9.3%) | 28 (44.4%) | 47 (11.2%) |
| *Enterprise size (employees)* | | | | |
| 1–50 | 57 (22.7%) | 27 (25.2%) | 14 (22.2%) | 98 (23.3%) |
| 51–200 | 61 (24.3%) | 21 (19.6%) | 22 (34.9%) | 104 (24.7%) |
| 201–1000 | 70 (27.9%) | 36 (33.6%) | 17 (27.0%) | 123 (29.2%) |
| Over 1000 | 63 (25.1%) | 23 (21.5%) | 10 (15.9%) | 96 (22.8%) |
| *Organizational AI use in HR* | | | | |
| L1 No AI used | 77 (30.7%) | 14 (13.1%) | 0 (0.0%) | 91 (21.6%) |
| L2 Testing or planning | 123 (49.0%) | 42 (39.3%) | 15 (23.8%) | 180 (42.8%) |
| L3 Hiring or recruiting only | 51 (20.3%) | 40 (37.4%) | 27 (42.9%) | 118 (28.0%) |
| L4 Hiring and performance reviews | 0 (0.0%) | 11 (10.3%) | 16 (25.4%) | 27 (6.4%) |
| L5 All processes, including promotion | 0 (0.0%) | 0 (0.0%) | 5 (7.9%) | 5 (1.2%) |
| *Self-rated AI understanding (1–5), mean (standard deviation)* | 3.69 (1.10) | 3.81 (1.12) | 3.68 (1.13) | 3.72 (1.11) |

### *D. Analysis*

Role differences used one-way analysis of variance (ANOVA) with Tukey's honestly significant difference (HSD) test, checked with Welch's $F$ and Kruskal–Wallis tests (Levene $p \geq 0.096$). Three ordinary least squares (OLS) regression models predicted trust: Model 1 regressed E on A to D; Model2 added experience (1–5), enterprise size (1–4), AI use in HR (L1–L5) and self-rated AI understanding (1–5); Model 3 repeated Model 1 with the explanation-independent sub-index as outcome. Because the Breusch–Pagan test indicated mild heteroscedasticity in Model 1 ($\chi 2(4) = 11.17$, $p = 0.025$), heteroscedasticity-consistent (HC3) standard errors [27] were also computed. Moderation was tested with mean-centered A and C and a product term, with simple slopes at ±1 SD [28]. The difference between the unstandardized coefficients of C and A was tested with a Wald test and, for standardized coefficients, with 5,000 case-resampling bootstrap replicates (seed 2026). Relative importance was assessed by general dominance analysis [29]. As a sensitivity analysis we excluded respondents who gave identical answers to all 21 construct items despite reverse-keyed items [30]. The role comparisons and Model 1 follow [4]; Models 2 and 3 and the moderation, dominance, bootstrap and sensitivity analyses were added during peer review and are exploratory. All estimates reported here were computed in Python (NumPy, SciPy); those in [4] (SPSS 26, R lavaan) were reproduced.

## IV. RESULTS

### *A. Role Differences*

Perceived fairness was the only construct below the midpoint, $M = 2.849$ ($SD = 1.160$), $t(420) = -2.68$, $p = 0.008$ (Table II). Engineers were far below 3.00 ($M = 2.262$, $t(250) = -13.01$, $p < 0.001$), whereas HR professionals and executives were above it. Role differed strongly, $F(2, 418) = 134.008$, $p < 0.001$, $\eta 2 = 0.391$ (Welch $F(2, 154.6) = 140.41$, $p < 0.001$). All pairs differed (Tukey $p < 0.001$ for engineers against both groups; $p = 0.012$ for HR against executives).

Governance demand and employee-centered orientation showed the same split: engineers and HR professionals did not differ (Tukey $p = 0.890$ and 0.277), and executives scored lower than both ($p < 0.001$), with governance demand numerically below but statistically indistinguishable from the midpoint ($M = 2.937$, $t(62) = -0.50$, $p = 0.618$).

Demand for explainability did not differ significantly by role, $F(2, 418) = 1.388$, $p = 0.251$, $\eta 2 = 0.007$. A non-significant test is not proof of equivalence, but all three role means lay between 3.71 and 3.92. At item level, formats C1 to C4 did not differ by role ($p = 0.474$ to 0.753); only the plain-language paragraph (C5) did, $F(2, 418) = 4.109$, $p = 0.017$, $\eta 2 = 0.019$, with engineers (4.024) above executives (3.571; Tukey $p = 0.012$). Anticipated trust was above the midpoint overall, $M = 3.609$, $t(420) = 13.15$, $p < 0.001$, and within every role,

including engineers ($M$ = 3.626, $t(250)$ = 10.56, $p < 0.001$) despite their low fairness scores. Construct E differed modestly by role, $F$ (2, 418) = 4.093, $p$ = 0.017, $\eta 2$ = 0.019, with only HR professionals and executives differing significantly (Tukey $p$ = 0.014).

*B. Predicting Trust*

Trust correlated with C ($r$ = 0.563), B ($r$ = 0.330), D ($r$ = 0.273) and, weakly, A ($r$ = 0.108, $p$ = 0.026); A and C were uncorrelated ($r$ = 0.009, $p$ = 0.861). Model 1 (Table III) explained 36.9% of the variance, $F$ (4, 416) = 60.86, $p < 0.001$. Explainability demand had the largest standardized coefficient ($\beta$ = 0.484; unstandardized $B$ = 0.516, 95% confidence interval (CI) 0.428 to 0.603), followed by governance ($\beta$ = 0.154), perceived fairness ($\beta$ = 0.134) and employee-centered orientation ($\beta$ = 0.114). The standardized coefficient of C was 3.6 times that of A. Adding experience, enterprise size, AI use and self-rated AI understanding changed little (Model 2: $R2$ = 0.374; $\beta_C$ = 0.483; $\beta_A$ = 0.131); no control was significant ($p \geq$ 0.217). With explanation-independent trust as outcome (Model 3), C remained dominant ($\beta$ = 0.441) over A ($\beta$ = 0.116). Using item 30 alone (trust in AI to review my own career fairly), C kept its weight ($\beta$ = 0.395) while A did not contribute ($\beta$ = 0.041, $p$ = 0.348).

*C. Does Explanation Demand Compensate for Low Fairness?*

If explanation compensated for doubts about fairness, the link between C and trust should be stronger where A is low. It was not. The A×C interaction was null, $B$ = −0.015 (SE = 0.039), $p$ = 0.698, 95% CI −0.093 to 0.062, and stayed null with role dummies ($p$ = 0.637) and within each role (all $p \geq 0.098$). The simple slope of C was 0.533 at −1 SD of A, 0.515 at the mean and 0.497 at +1 SD (all $p < 0.001$; Fig. 1). Trust was no lower among respondents with low fairness beliefs: the 209 respondents (49.6%) scoring below 3.00 on A had $M$ = 3.616 against $M$ = 3.603 for the other 212 (Welch $p$ = 0.885), and 73.2% of them scored above 3.00 on trust. On relative strength, the coefficient of C exceeded that of A (Wald test on $B$: difference = 0.406, SE = 0.057, $t(416)$ = 7.16, $p < 0.001$). In 5,000 bootstrap resamples the standardized difference was 0.348 (95% CI 0.228 to 0.463), and the ratio of $\beta_C$ to $\beta_A$ had a median of 3.6 (95% CI 2.1 to 8.5), so "about three to four times" is only a rough summary.

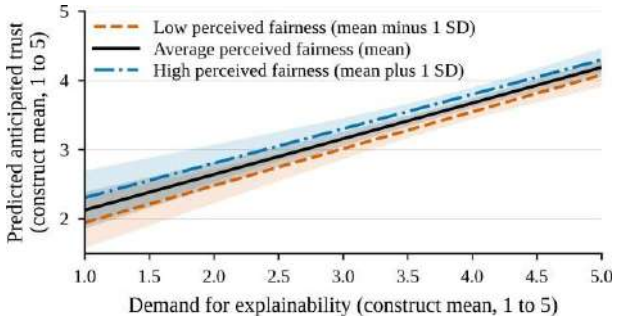



Fig. 1. Association between explainability demand and anticipated trust at low, average and high perceived fairness. The three lines are close to parallel (interaction $p$ = 0.698).

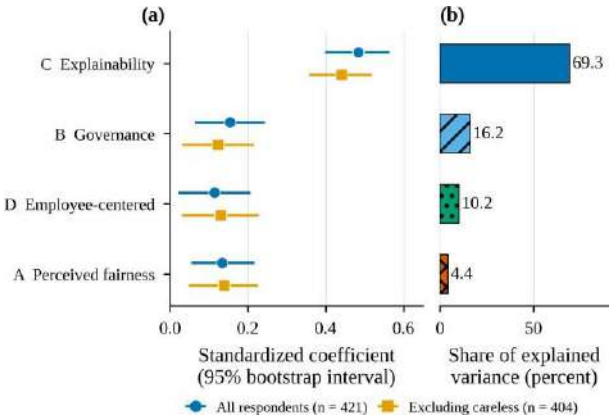



Fig. 2. Relative weight of the four predictors of anticipated trust: (a) standardized coefficients with 95% bootstrap intervals, for all respondents and without the 17 careless responders; (b) share of explained variance by general dominance (all respondents).

In the dominance analysis C accounted for 69.3% of the explained variance (B 16.2%, D 10.2%, A 4.4%); the unique contribution of C was $\Delta R2$ = 0.204 against 0.017 for

A. Fig. 2 shows both results.

*D. Robustness*

The main pattern survived further checks. Adding role dummies gave $\beta_C$ = 0.489 and $\beta_A$ = 0.162 ($p$ = 0.002), with no role effect on trust beyond the four constructs. Spearman correlations gave the same ordering (0.550 for C, 0.089 for A). One check changed the magnitude. Seventeen respondents (9 engineers, 5 HR, 3 executives) answered all 21 construct items identically, reverse-keyed items included. Without them ($n$ = 404) Model 1 explained less variance ($R2$ = 0.295), with $\beta_C$ = 0.441, $\beta_A$ = 0.139, $\beta_B$ = 0.123 and $\beta_D$ = 0.130 (HC3 $p \leq 0.008$), a ratio of 3.2; the 95% bootstrap intervals were 0.361 to 0.515 for C and 0.051 to 0.222 for A (Fig. 2a). The interaction stayed null ($p$ = 0.412). The full-sample estimates in Table III, which match [4], may therefore overstate the C–E association.

## V. Discussion

Roles disagree sharply about whether an AI promotion process could be fair, agree on wanting explanation, and trust follows the second far more than the first, even among engineers, whose fairness beliefs sit well below the midpoint. The title's question therefore receives a qualified answer. The pattern is additive: explanation demand and fairness belief are separately associated with trust, without interaction, and the explanation association is far larger. The data do not support the stronger reading in which explanation compensates for perceived unfairness, since the moderation test is null and trust is as high where fairness beliefs are low. "Substitute" thus denotes explanation demand standing in for fairness as the main correlate of trust, not a repair of the fairness deficit. One untested account comes from the justice literature: explanation belongs to informational justice [5], [9], so people may extend trust to a process that accounts for itself whatever they think of its fairness.

TABLE II

CONSTRUCT MEANS BY ORGANIZATIONAL ROLE, *M* (SD), AND ONE-WAY ANOVA (*N* = 421)

| **Construct** | **Engineers** (*n* = 251) | **HR** (*n* = 107) | **Executives** (*n* = 63) | **All** | *F*(2, 418) | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|---|
| A. Perceived procedural fairness | 2.262 (0.899) | 3.561 (0.971) | 3.976 (0.825) | 2.849 (1.160) | 134.008 | <0.001 | 0.391 |
| B. Demand for governance | 4.085 (0.859) | 4.037 (0.912) | 2.937 (1.007) | 3.901 (0.981) | 43.062 | <0.001 | 0.171 |
| C. Demand for explainability | 3.924 (0.895) | 3.882 (0.907) | 3.714 (0.848) | 3.882 (0.892) | 1.388 | 0.251 | 0.007 |
| D. Employee-centered orientation | 4.012 (0.791) | 3.865 (0.825) | 3.194 (0.967) | 3.852 (0.873) | 24.614 | <0.001 | 0.105 |
| E. Anticipated trust and reliance | 3.626 (0.940) | 3.741 (0.910) | 3.317 (1.011) | 3.609 (0.950) | 4.093 | 0.017 | 0.019 |

TABLE III

ORDINARY LEAST SQUARES REGRESSIONS PREDICTING ANTICIPATED TRUST (*N* = 421)

| | **Model 1: Perceptions** | | | **Model 2: With controls** | | | **Model 3: Items 30 and 33** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Predictor** | *B* (SE) | *β* | *t* (*p*) | *B* (SE) | *β* | *t* (*p*) | *B* (SE) | *β* | *t* (*p*) |
| (Constant) | 0.234 (0.245) | — | 0.954 (0.341) | −0.028 (0.314) | — | −0.089 (0.929) | 0.229 (0.285) | — | 0.802 (0.423) |
| A. Perceived fairness | 0.110 (0.032) | 0.134 | 3.389 (0.001) | 0.108 (0.035) | 0.131 | 3.115 (0.002) | 0.105 (0.038) | 0.116 | 2.797 (0.005) |
| B. Demand for governance | 0.149 (0.042) | 0.154 | 3.540 (<0.001) | 0.142 (0.043) | 0.146 | 3.315 (0.001) | 0.137 (0.049) | 0.128 | 2.806 (0.005) |
| C. Demand for explainability | 0.516 (0.045) | 0.484 | 11.586 (<0.001) | 0.515 (0.045) | 0.483 | 11.503 (<0.001) | 0.520 (0.052) | 0.441 | 10.041 (<0.001) |
| D. Employee-centered | 0.125 (0.045) | 0.114 | 2.746 (0.006) | 0.128 (0.046) | 0.118 | 2.798 (0.005) | 0.138 (0.053) | 0.115 | 2.619 (0.009) |
| Experience (1–5) | — | — | — | −0.016 (0.034) | −0.020 | −0.478 (0.633) | — | — | — |
| Enterprise size (1–4) | — | — | — | 0.043 (0.034) | 0.049 | 1.237 (0.217) | — | — | — |
| AI use in HR (L1–L5) | — | — | — | 0.034 (0.045) | 0.032 | 0.755 (0.451) | — | — | — |
| AI understanding (1–5) | — | — | — | 0.041 (0.034) | 0.047 | 1.196 (0.232) | — | — | — |
| $R^2$ (adjusted) | 0.369 (0.363) | | | 0.374 (0.362) | | | 0.303 (0.297) | | |
| *F* (degrees of freedom), *p* | 60.86 (4, 416), <0.001 | | | 30.80 (8, 412), <0.001 | | | 45.26 (4, 416), <0.001 | | |

*Note.* Conventional OLS standard errors; all VIFs ≤ 1.281. With HC3 errors every predictor A to D stays significant in all models (largest change, D: $p = 0.013$, 0.010, 0.014). Model 3 outcome: mean of items 30 and 33 ($M = 3.618$, $SD = 1.053$).

Alternative accounts remain open: a general favorable disposition toward AI could underlie both the wish for explanation and trust, and trust could shape what respondents want explained. For HRIS procurement the implication is conditional but practical. If explanation features alone can sustain readiness to rely on a system, acceptance cannot serve as evidence that the process is fair, a risk consistent with warnings about transparency without accountability [15], [16], [18]. Independent fairness audits and human review should therefore stay outside the explanation layer. Executives, who typically hold configuration authority [19], were the most favorable to AI fairness ($M = 3.976$) and the least supportive of governance ($M = 2.937$); engineers and HR professionals wanted oversight. These data cannot say why, but they suggest human review should not be a setting the least supportive group can switch off.

## VI. LIMITATIONS AND FUTURE RESEARCH

The design is cross-sectional, so the regressions describe association, not direction or cause. Every measure is self-reported from one source, and the vignette is hypothetical: only five respondents (1.2%) reported AI already used across all processes including promotion (L5), and E records anticipated

trust, not behavior in a live promotion cycle. Sampling was purposive and snowball, confined to the Sri Lankan IT sector, so results should not be generalized statistically; response rates cannot be computed, respondents are nested in unidentified firms, power-distance norms may amplify role differences, and the executive group is small ($n$ = 63). The instrument was developed for this study and validated only in this sample, AVE for D and E is below 0.50, and C records what respondents say they want, not how explanations perform. Perceived fairness is measured; the actual fairness of any system is not, and A combines beliefs about AI versus human managers with fairness features and is nearly uncorrelated with C, so its weak link to trust may partly reflect what A captures. Excluding 17 careless responders lowers the headline $R2$, and the moderation, dominance and sensitivity analyses are exploratory.

Future work should manipulate explanation quality and objective fairness assurance in vignette or prototype experiments, measure comprehension of real explanations, control for general AI attitude, and follow deployed HRIS promotion modules over several cycles.

## VII. Conclusion

Across 421 IT professionals, role divided opinion on the fairness of algorithmic promotion ($\eta 2 = 0.391$) but not demand for explanation ($\eta 2 = 0.007$). Anticipated trust was moderate to high everywhere and tracked explanation demand about three to four times as strongly as perceived fairness, with no sign of compensation where fairness beliefs are low. Acceptance of an explainable promotion tool should not be read as evidence that it is fair; explanation needs to be paired with mandatory human review and independent fairness auditing.

## Acknowledgment

The authors thank the 421 IT professionals who completed the survey anonymously. During the preparation of this manuscript, the authors used generative AI tools solely for language copy-editing and structural formatting. All analyses and interpretations were produced and verified by the authors, who take full responsibility for the accuracy, originality and integrity of the manuscript. Anonymized item-level data, the questionnaire with its item justifications, and analysis scripts are available from the corresponding author on reasonable request.

## References

[1] P. Tambe, P. Cappelli, and V. Yakubovich, “Artificial intelligence in human resources management: Challenges and a path forward,” *California Manage. Rev.*, vol. 61, no. 4, pp. 15–42, 2019.

[2] A. Köchling and M. C. Wehner, “Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development,” *Bus. Res.*, vol. 13, pp. 795–848, 2020, doi: 10.1007/s40685-020-00134-w.

[3] G. Marín Díaz, J. J. Galán Hernández, and J. L. Galdón Salvador, “Analyzing employee attrition using explainable AI for strategic HR decision-making,” *Mathematics*, vol. 11, no. 22, Art. no. 4677, 2023, doi: 10.3390/math11224677.

[4] L. S. V. Samarappulige, L. S. Lekamge, and A. S. A. Perera, “A human-centric XAI framework for ethical and explainable AI in HRIS: Enhancing employee career advancement in the Sri Lankan IT sector,” in *Proc. Int. Res. Conf. SLTC (IRC)*, 2026, Paper 181, in press.

[5] J. A. Colquitt, “On the dimensionality of organizational justice: A construct validation of a measure,” *J. Appl. Psychol.*, vol. 86, no. 3, pp. 386–400, 2001, doi: 10.1037/0021-9010.86.3.386.

[6] D. T. Newman, N. J. Fast, and D. J. Harmon, “When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions,” *Organ. Behav. Hum. Decis. Process.*, vol. 160, pp. 149–167, Sep. 2020.

[7] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[8] G. S. Leventhal, “What should be done with equity theory? New approaches to the study of fairness in social relationships,” in *Social Exchange: Advances in Theory and Research*, K. J. Gergen, M. S. Greenberg, and R. H. Willis, Eds. New York, NY, USA: Plenum, 1980, pp. 27–55.

[9] J. C. Shaw, E. Wild, and J. A. Colquitt, “To justify or excuse?: A meta-analytic review of the effects of explanations,” *J. Appl. Psychol.*, vol. 88, no. 3, pp. 444–458, 2003.

[10] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” *MIS Quart.*, vol. 13, no. 3, pp. 319–340, Sep. 1989.

[11] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, “User acceptance of information technology: Toward a unified view,” *MIS Quart.*, vol. 27, no. 3, pp. 425–478, Sep. 2003.

[12] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995, doi: 10.5465/amr.1995.9508080335.

[13] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[14] M. K. Lee, “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management,” *Big Data Soc.*, vol. 5, no. 1, pp. 1–16, 2018.

[15] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” *New Media Soc.*, vol. 20, no. 3, pp. 973–989, Mar. 2018.

[16] G. Bansal *et al.*, “Does the whole exceed its parts? The effect of AI explanations on complementary team performance,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Yokohama, Japan, 2021, Art. no. 81, pp. 1–16, doi: 10.1145/3411764.3445717.

[17] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[18] I. Ajunwa, “The paradox of automation as anti-bias intervention,” *Cardozo Law Rev.*, vol. 41, no. 5, pp. 1671–1742, 2020.

[19] K. C. Kellogg, M. A. Valentine, and A. Christin, “Algorithms at work: The new contested terrain of control,” *Acad. Manage. Ann.*, vol. 14, no. 1, pp. 366–410, 2020, doi: 10.5465/annals.2018.0174.

[20] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951.

[21] L. Hu and P. M. Bentler, “Cutoff criteria for fit indexes in co-variance structure analysis: Conventional criteria versus new alternatives,” *Struct. Equ. Model.*, vol. 6, no. 1, pp. 1–55, 1999, doi: 10.1080/10705519909540118.

[22] C. Fornell and D. F. Larcker, “Evaluating structural equation models with unobservable variables and measurement error,” *J. Mark. Res.*, vol. 18, no. 1, pp. 39–50, Feb. 1981.

[23] J. Henseler, C. M. Ringle, and M. Sarstedt, “A new criterion for assessing discriminant validity in variance-based structural equation modeling,” *J. Acad. Mark. Sci.*, vol. 43, no. 1, pp. 115–135, 2015, doi: 10.1007/s11747-014-0403-8.

[24] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, “Common method biases in behavioral research: A critical review of the literature and recommended remedies,” *J. Appl. Psychol.*, vol. 88, no. 5, pp. 879–903, 2003.

[25] N. Kock, “Common method bias in PLS-SEM: A full collinearity assessment approach,” *Int. J. e-Collab.*, vol. 11, no. 4, pp. 1–10, 2015.

[26] M. K. Lindell and D. J. Whitney, “Accounting for common method variance in cross-sectional research designs,” *J. Appl. Psychol.*, vol. 86, no. 1, pp. 114–121, 2001.

[27] J. G. MacKinnon and H. White, “Some heteroskedasticity-consistent

covariance matrix estimators with improved finite sample properties," *J. Econometrics*, vol. 29, no. 3, pp. 305–325, 1985.

[28] L. S. Aiken and S. G. West, *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA, USA: Sage, 1991.

[29] D. V. Budescu, "Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression," *Psychol. Bull.*, vol. 114, no. 3, pp. 542–551, 1993.

[30] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data," *Psychol. Methods*, vol. 17, no. 3, pp. 437–455, 2012.

# Mathematical Modelling and Dynamic Graph Analysis of Real-Time Financial Market Interdependencies for Systemic Risk Prediction

W. D. J. I. Senarathna
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
djisenarathna@students.nsbm.ac.lk

T. A. H. Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Network-based indicators are widely promoted as early-warning signals of systemic stress, yet they are almost never tested against the simplest available alternative, which is volatility itself. This paper subjects eleven correlation-network indicators, drawn from the structural and magnitude families of the literature, to a single controlled evaluation on a twenty-nine node cross-asset network observed over 2,636 trading days. All thresholds, orientations and scalers are frozen on a burn-in period, so that the protocol is free of look-ahead bias by construction. Three findings emerge. Structural indicators prove genuinely distinct from volatility, with a median explained variance of 0.104 against 0.301 for magnitude indicators, and the family spans seven principal components rather than collapsing onto one. That distinctness, however, does not translate into predictive value: none of the eleven improves upon a four-variable volatility baseline, seven reverse sign out of sample, and a data-snooping correction fails to reject the null of no superior predictive ability at any estimation window. Once the mechanical contamination induced by trailing estimation windows is removed, every indicator peaks within one to three days of market turbulence, with a maximum cross-correlation of 0.308. Under an equal alarm budget several network measures do outperform the VIX at episode detection, but the highest attainable precision is 0.35. These measures are therefore descriptive rather than anticipatory, and the burden of proof for their use in early warning has not been met.**

***Keywords—Systemic Risk, Dynamic Graphs, Financial Networks, Early Warning***

## I. Introduction

Financial systems fail through their connections. Dense interconnection absorbs small shocks but propagates large ones, so links that supplied diversification become channels of contagion [1]. The literature has converged on two families. The magnitude family quantifies how strongly the system is coupled: econometric connectedness from Granger-causality networks [2] and variance decompositions [3], tail measures such as CoVaR [4], and the leading eigenvalue and mean correlation of the covariance matrix. The structural family quantifies how the dependency topology is arranged and how fast that arrangement changes, running from the correlation-distance minimum spanning tree [5] to the single-step survival ratio, which counts tree edges persisting between consecutive windows [6]. Its appeal rests on an analogy with critical transitions, where statistical signatures precede a tipping point [7]: if a market reorganises before it breaks, reorganisation ought to lead turbulence. Recent work pursues the same intuition with heavier machinery, including higher-order network representations [8] and regime-dependent graph neural networks [9], whose results also depend on training choices such as the optimiser [10].

Three research gaps remain, and together they constitute the ground this paper occupies. First, competing indicators have never been placed on identical data, windows and alarm rules, the only arrangement under which one can be shown to improve upon another rather than merely re-express it. Second, and more serious, the obvious null hypothesis is rarely tested: correlation rises when volatility rises, so any correlation-matrix statistic inherits volatility signal automatically, yet studies seldom report how much of their indicator volatility already explains. An indicator that tracks the VIX is not a discovery about network structure. Third, evaluation protocols remain permissive, with thresholds calibrated on the full sample and applied retrospectively, lead times quoted without the false-alarm rates that give them meaning, and selection bias from screening many candidates against one benchmark left uncorrected.

Four objectives follow directly. The first is to construct eleven indicators spanning both families from a single dynamic cross-asset network under one causal convention, so that performance differences reflect the measures rather than their inputs. The second is to quantify how much of each indicator volatility alone explains, and whether any incremental predictive information survives that control. The third is to test the timing claim directly, separating genuine lead from the mechanical artefacts that trailing estimation windows introduce. The fourth is to evaluate alarms under a look-ahead-free protocol reporting false-alarm rates alongside detection rates, with an explicit data-snooping correction. The contribution is therefore not a new indicator but a neutral comparison, of the kind used for model selection in other prediction tasks [11], [12], establishing what this class of measures can and cannot do, and the benchmark future proposals should clear.

## II. Related Work

The correlation-distance minimum spanning tree [5] remains the dominant filtration in empirical finance, reducing a dense matrix to an interpretable skeleton of N−1 edges. That

skeleton is not static: tree topology shifts systematically with the business cycle and contracts during turbulence [6]. The proposition that topological change anticipates stress has been advanced from several directions. Analyses of local equity markets report that the Jaccard index between consecutive network snapshots shifts around crises, and propose it as a precursor [13]. Interbank studies document topological collapse preceding systemic events [14], while information spreading has been shown to flicker, with nodes entering and leaving the diffusion process in rapid succession, before critical transitions [15]. A parallel strand applies topological data analysis, deriving persistence landscapes from sliding windows of returns and reporting lead times of several weeks [16].

The econometric tradition approaches the same problem from the magnitude side, measuring coupling through forecast-error variance decompositions and yielding directional spillover indices that have become the standard reference in applied work [3]. These describe the strength of transmission rather than the rate at which the transmission structure itself changes, and are treated here as benchmarks of the magnitude family. What unites these strands also limits them. Each validates its own measure against crisis episodes, typically with the VIX as sole comparator and with thresholds chosen in the knowledge of which episodes mattered. None reports the incremental information its measure carries beyond volatility, and none compares the two families under a common protocol.

## III. Mathematical Framework

### A. Notation and the dynamic graph

Let $P_{i,t}$ denote the adjusted closing price of asset $i \in \{1,\ldots,N\}$ on trading day t. Log returns are defined as

$$r_{i,t} = \ln P_{i,t} - \ln P_{i,t-1} \tag{1}$$

For a trailing window of length W ending at t, let $\Sigma_t$ be the sample covariance matrix of the returns in (t−W, t] and $D_t = \mathrm{diag}(\Sigma_t)$; the correlation matrix $C_t \in \mathbb{R}^{N\times N}$ is then

(2)

To bound the condition number in short windows, a fixed shrinkage is applied,

$$C_t = (1-a)\,C_t + a\,I_N, \qquad a = 10^{-3} \tag{3}$$

identically for every indicator, so that no comparison is advantaged. The resulting sequence $C_t$ constitutes the dynamic graph. Every quantity defined below is computed from returns falling in (t-W, t] only, and measures requiring two consecutive windows are assigned to the later of the two dates. Correlations are mapped to the standard ultrametric distance

$$d_{t,ij} = \sqrt{2(1 - C_{t,ij})} \tag{4}$$

from which the minimum spanning tree $T_t$ with edge set $E_t$ is extracted. The tree retains N-1 edges and forms the basis of the structural measures.

### B. The structural family

The proportion of tree edges failing to persist from one window to the next is one minus the single-step survival ratio [6], and is written

$$\tau_t = 1 - \frac{|E_t \cap E_{t-1}|}{|E_t \cup E_{t-1}|} \tag{5}$$

Correlation matrices are symmetric positive definite and therefore do not inhabit a flat space, a fact the Euclidean difference ignores. The affine-invariant distance between consecutive matrices,

(6)

where $\lambda_k$ are the generalised eigenvalues of the pair ( $C_t$, $C_{t-1}$), weights every eigendirection logarithmically, so that a proportional collapse in a small eigenvalue counts as heavily as the same proportional movement in the market mode. Its Euclidean counterpart,

$$\phi_t = \|C_t - C_{t-1}\|_F \tag{7}$$

treats the space as flat and is dominated by the largest eigenvalue; including both isolates how much the underlying geometry actually matters. Reorientation of the dominant market factor is captured by the angle between successive leading eigenvectors $v_t$,

$$\theta_t = \arccos(|v_t^{\top} v_{t-1}|) \tag{8}$$

Treating the scaled graph Laplacian as a density matrix [17] yields a spectral measure of structural dispersion. With $w_{ij} = \max(C_{t,ij}, 0)$, $L_t = \mathrm{diag}(\Sigma_j w_{ij}) - W_t$ and $\varrho_t = L_t / \mathrm{tr}(L_t)$, the von Neumann entropy is

$$S_t = -\mathrm{tr}(\varrho_t \log \varrho_t) = -\sum_{k=1}^{N} \mu_k \log \mu_k\,, \varrho_t = \frac{L_t}{\mathrm{tr}(L_t)} \tag{9}$$

in which $\mu_k$ are the eigenvalues of $\varrho_t$. Entropy falls as the network funnels into a single dominant channel. The MST diameter, being the longest shortest path in $T_t$, completes the family.

### C. The magnitude family

The leading eigenvalue $\lambda_{1,t}$ of $C_t$, its normalised share $\lambda_{1,t}/N$, the mean off-diagonal correlation, and the total MST edge length (the sum of $d_{t,ij}$ over $E_t$) constitute four magnitude measures. A fifth counts the eigenvalues that carry genuine signal. Under the Marchenko–Pastur law [18], the spectrum of a correlation matrix estimated from W independent observations of N series is bounded above by

(10)

so that eigenvalues below $\lambda^+$ are indistinguishable from noise. The empirical relevance of this bound to financial correlation matrices is well documented [19]; with N=29 and W=60 it gives $\lambda^+ = 2.87$, and the count $n_t^{sig} = |\{k : \lambda_{k,t} > \lambda^+\}|$ follows.

*D. Volatility controls and the target variable*

The control block comprises annualised twenty-day realised volatility of the target index, the VIX, the cross-sectional dispersion of node returns, and mean node volatility. Because intraday data are unavailable, range-based estimators are additionally formed from the daily open, high, low and close. The Parkinson estimator [20],

$$\sigma^{\mathrm{P}}_{i,t} = \sqrt{\frac{1}{4\ln 2}\ln^2\left(\frac{H_{i,t}}{L_{i,t}}\right)} \tag{11}$$

and the Garman–Klass estimator [21],

$$\sigma^{\mathrm{GK}}_{i,t} = \sqrt{\frac{1}{2}\ln^2\left(\frac{H_{i,t}}{L_{i,t}}\right) - (2\ln 2 - 1)\ln^2\left(\frac{C_{i,t}}{O_{i,t}}\right)} \tag{12}$$

are several times more efficient than the squared close-to-close return and permit a volatility layer at no additional data cost. Systemic stress itself is operationalised as a forward drawdown of the benchmark index, so that for horizon h and threshold δ,

$$D_{t,h} = \frac{\min_{s\in(t,\,t+h]} P_s}{P_t} - 1, \qquad y_t = 1[D_{t,h} \le -\delta] \tag{13}$$

Since $D_{t,h}$ is strictly forward-looking, it serves only as a label and never as a feature.

## IV. Streaming Computation

Recomputing the full spectrum at every tick costs $O(N^3)$. Because the window advances by a single observation, the second moments may instead be updated incrementally. Writing $M_t$ for the uncentred second-moment matrix,

$$M_t = M_{t-1} + r_t r_t^{\mathsf{T}} - r_{t-W} r_{t-W}^{\mathsf{T}} \tag{14}$$

which is a pair of rank-one updates costing $O(N^2)$. Algorithm 1 states the procedure in full. The implementation records 1.36 ms per window update at N=29 and W=60, so that the entire ten-year history is processed in under four seconds; real-time operation never becomes the binding constraint, and the low per-update cost also limits the energy use of continuous processing [22].

```
Algorithm 1: Streaming indicator update
─────────────────────────────────────────────────────────────────
Input : return vector r_t, buffer R of the last W returns,
        matrix M, state (C_prev, E_prev, v_prev)
Output: indicator vector x_t

 1  M ← M + r_t r_t^T − r_{t−W} r_{t−W}^T          # O(N²)
 2  Σ ← M/W − μμ^T ;  C ← D^{−1/2} Σ D^{−1/2}      # D = diag(Σ)
 3  C ← (1−a)C + aI                                # eq. (3)
 4  (λ, V) ← eigh(C)                               # O(N³)
 5  d_ij ← sqrt(2(1 − C_ij)) ;  E ← MST(d)         # O(N² log N)
 6  τ ← 1 − |E ∩ E_prev| / |E ∪ E_prev|            # eq. (5)
 7  ρ ← sqrt(Σ_k log²(genEig(C, C_prev)))          # eq. (6)
 8  ϕ ← ‖C − C_prev‖_F ;  θ ← arccos|v₁^T v_prev|  # (7), (8)
 9  S ← −Σ_k μ_k log μ_k  with ϱ = L/tr(L)         # eq. (9)
10  n_sig ← |{k : λ_k > (1 + sqrt(N/W))²}|         # eq. (10)
11  (C_prev, E_prev, v_prev) ← (C, E, v₁)
12  return [τ,ρ,ϕ,θ,S,diam(E),λ₁,λ₁/N,mean(C),n_sig,len(E)]
─────────────────────────────────────────────────────────────────
```

## V. Experimental Design

The network comprises twenty-nine deliberately non-overlapping instruments: ten United States equity sectors, five international equity markets, six fixed-income maturities and credit grades, five commodities and three currencies as listed in Table I. Daily prices, retrieved from Yahoo Finance via the yfinance library, span 5 January 2016 to 30 June 2026, giving 2,636 aligned records with no missing values; ticker symbols and code are available from the authors on request. The node set is not incidental. Headline equity indices, which much of the literature adopts, overlap by construction: their leading eigenvalue explains 90.3 % of total variance at a median pairwise correlation of 0.904, so the rolling minimum spanning tree admits only nine edges and four distinct turnover values across a decade. The twenty-nine node set reduces these to 37.1 % and 0.226, and it is this that makes the structural measures meaningful at all.

Stress episodes are detected mechanically rather than nominated by hand. An episode opens when the benchmark drawdown first exceeds a threshold, back-dates to the preceding peak, and closes once that peak is regained. At five per cent the rule recovers fifteen episodes, among them the COVID-19 crash, the 2022 bear market and the 2025 tariff shock, none specified in advance.

The 693 observations to 31 December 2018 form a burn-in on which every threshold, orientation, scaler and model coefficient is estimated and then frozen; the remaining 1,883 form the evaluation sample, separated by a ten-day embargo. Four tests follow: each indicator is regressed on the volatility control block; the family is decomposed by principal components; a logistic model fitted on the burn-in records the change in area under the curve when each indicator is added; and timing is examined through cross-correlation with $|r_{t+k}|$ over $k \in [-60, 60]$. Alarms are assessed under frozen burn-in thresholds and, separately, under an equal alarm budget.

Three corrections address specific hazards. Screening eleven candidates against one benchmark inflates the apparent performance of the best, so the loss adopted is the Brier score [23], which unlike the area under the curve decomposes per observation and admits a loss differential in the sense of Diebold and Mariano [24]; both White's Reality Check [25] and Hansen's

studentised test of superior predictive ability [26] are reported. Overlapping forward labels leave only thirty-seven contiguous positive runs, so all confidence intervals derive from a moving-block bootstrap [27] with twenty-one-day blocks. Granger-causality probabilities are adjusted by the Benjamini–Hochberg procedure [28].

TABLE I. COMPOSITION OF THE TWENTY-NINE NODE CROSS-ASSET NETWORK

| Asset block | N | Constituents |
|---|---|---|
| US equity sectors | 10 | Technology, Financials, Energy, Health Care, Industrials, Consumer Discretionary, Consumer Staples, Utilities, Materials, Real Estate |
| International equity | 5 | Emerging Markets, Developed ex-US, Japan, China, Brazil |
| Fixed income | 6 | Treasury 20Y, Treasury 7–10Y, Treasury 1–3Y, Corporate IG, Corporate HY, TIPS |
| Commodities | 5 | Gold, Silver, Crude oil, Broad commodity, Agriculture |
| Currencies | 3 | US dollar, Euro, Japanese yen |
| Total | 29 | — |

## VI. RESULTS

### *A. How much of each indicator is volatility?*

Table II reports the share of each indicator explained by the volatility control block, together with its out-of-sample discrimination. The separation between the two families is clear and favours the structural group, whose median coefficient of determination is 0.104 against 0.301 for the magnitude group. Total MST length proves the most contaminated measure at 0.360, followed closely by the leading eigenvalue at 0.301, whereas the signal-eigenvalue count, eigenvector rotation and MST turnover all fall below 0.06. Structural indicators are therefore measuring something that volatility does not already capture. Neither are they redundant with one another. Seven principal components are required to span ninety per cent of the variance of the eleven-indicator family, with the first explaining only 34.9 per cent. The simplest version of the sceptical hypothesis, namely that these measures are one quantity wearing eleven names, is thereby refuted.

TABLE I. VOLATILITY REDUNDANCY AND OUT-OF-SAMPLE DISCRIMINATION

| Indicator | F | $R^2$ all | $R^2$ calm | AUC burn-in | AUC OOS | 95 % CI | Sign stable |
|---|---|---|---|---|---|---|---|
| Realised volatility | V | — | — | 0.635 | 0.712 | [0.601, 0.805] | yes |
| VIX | V | — | — | 0.729 | 0.698 | [0.592, 0.794] | yes |
| von Neumann entropy | S | 0.117 | 0.265 | 0.717 | 0.556 | [0.410, 0.706] | yes |
| Eigenvector rotation | S | 0.042 | 0.047 | 0.503 | 0.549 | [0.473, 0.625] | yes |
| MST turnover | S | 0.057 | 0.046 | 0.583 | 0.543 | [0.445, 0.644] | yes |
| Frobenius speed | S | 0.297 | 0.147 | 0.531 | 0.542 | [0.448, 0.640] | yes |
| Signal eigenvalues | M | 0.011 | 0.066 | 0.621 | 0.494 | [0.377, 0.602] | NO |
| Mean correlation | M | 0.197 | 0.168 | 0.553 | 0.490 | [0.341, 0.641] | NO |
| Riemannian speed | S | 0.298 | 0.132 | 0.502 | 0.461 | [0.354, 0.558] | NO |
| MST diameter | S | 0.092 | 0.133 | 0.629 | 0.451 | [0.346, 0.558] | NO |
| $\lambda_1$ | M | 0.301 | 0.186 | 0.555 | 0.402 | [0.271, 0.533] | NO |
| $\lambda_1$ share | M | 0.301 | 0.186 | 0.555 | 0.402 | [0.274, 0.536] | NO |
| MST length | M | 0.360 | 0.146 | 0.552 | 0.386 | [0.262, 0.526] | NO |

### *B. Discrimination and sign stability*

The picture changes sharply once predictive performance is examined. With each indicator's orientation fixed on the burn-in and applied unchanged thereafter, the best network measure achieves an out-of-sample area under the curve of 0.556, against 0.698 for the VIX and 0.712 for realised volatility. Under a moving-block bootstrap not one of the eleven produces a confidence interval excluding 0.5, whereas realised volatility attains an interval of 0.601 to 0.805. More troubling still, seven of the eleven reverse direction out of sample. MST diameter and the signal-eigenvalue count both discriminate respectably in sample, at 0.629 and 0.621 respectively, and then predict backwards, at 0.451 and 0.494. A relationship that inverts out of sample is not a weak signal but an unstable one.

### *C. Incremental value over volatility*

TABLE II. INCREMENTAL AUC OVER THE VOLATILITY BASELINE

| Indicator | F | Alone | Baseline + indicator | ΔAUC | Purged alone |
|---|---|---|---|---|---|
| Frobenius speed | S | 0.5423 | 0.6891 | -0.0027 | 0.4922 |
| Riemannian speed | S | 0.4613 | 0.6865 | -0.0053 | 0.4873 |
| Eigenvector rotation | S | 0.4510 | 0.6848 | -0.0070 | 0.4421 |
| Mean correlation | M | 0.4899 | 0.6807 | -0.0111 | 0.5768 |
| MST turnover | S | 0.5426 | 0.6788 | -0.0131 | 0.5139 |
| MST length | M | 0.3859 | 0.6724 | -0.0194 | 0.4613 |
| MST diameter | S | 0.4506 | 0.6701 | -0.0218 | 0.4965 |
| $\lambda_1$ share | M | 0.4024 | 0.6631 | -0.0287 | 0.4951 |
| $\lambda_1$ | M | 0.4024 | 0.6631 | -0.0287 | 0.4951 |
| Signal eigenvalues | M | 0.4936 | 0.6495 | -0.0423 | 0.4246 |
| von Neumann entropy | S | 0.5564 | 0.5922 | -0.0996 | 0.4463 |

Table III presents the central result. The four-variable volatility baseline attains an area under the curve of 0.6918 on the evaluation sample, and adding any one of the eleven indicators reduces it, with changes ranging from −0.0027 to −0.0996. Not a single indicator improves upon the baseline. When each is residualised on the control block and its volatility-free component allowed to predict alone, performance falls between 0.42 and 0.60, which is consistent with noise.

*D. Timing*

The timing evidence requires care, because the naive reading of it is wrong. An indicator estimated on a trailing window of W days is a function of the returns from t-W+1 to t, so correlating it with $|r_{t+k}|$ at negative k correlates the indicator with its own inputs; a single large return sits inside the window for W days and elevates the indicator throughout. Any apparent lag in that region is created by construction rather than by the market. The diagnostic is decisive. Re-estimating the peak lag at windows of forty, sixty, ninety and one hundred and twenty days shifts the mean peak from −17.3 to −30.4 to −38.8 and then to −50.3 days, while for the leading eigenvalue alone it moves from −19 to −24 to −30 and then to −35. A genuine economic lag would not scale with an arbitrary estimation choice. Table IV reports this scaling in full. Only positive lags therefore admit an economic reading, since future returns cannot lie inside a trailing window. There, every indicator peaks within one to three days, and the largest cross-correlation anywhere on the uncontaminated side is 0.308, achieved by Riemannian speed at a single day, whereas the VIX reaches 0.620 at lag zero. The correct conclusion is consequently narrower than the literature's, and also firmer: these indicators are coincident, moving with turbulence rather than before it.

TABLE III. PEAK LAG BY ESTIMATION WINDOW

| Indicator | W=40 | W=60 | W=90 | W=120 | Peak k>0 | Corr k>0 |
|---|---|---|---|---|---|---|
| $\lambda_1$ | -19 | -24 | -30 | -35 | +1 | +0.268 |
| Mean correlation | -14 | -29 | -34 | -34 | +1 | +0.185 |
| von Neumann entropy | -39 | -59 | -38 | -50 | +3 | -0.166 |
| MST diameter | -3 | -36 | -71 | -76 | +3 | -0.120 |
| MST length | -10 | -9 | -30 | -35 | +1 | -0.291 |
| Riemannian speed | +0 | +0 | +0 | +0 | +1 | +0.308 |
| Frobenius speed | +0 | +0 | +0 | +0 | +3 | +0.226 |
| MST turnover | +0 | +0 | +0 | +0 | +1 | +0.114 |
| Eigenvector rotation | +0 | -32 | +0 | +0 | +115 | -0.074 |
| Mean over all indicators | -17.3 | -30.4 | -38.8 | -50.3 | — | — |

*E. Alarms*

Under thresholds frozen from the burn-in, performance is poor and, in four cases, undefined. MST turnover, Frobenius speed, eigenvector rotation and the signal-eigenvalue count prove so noisy relative to their dynamic range that their maximum standardised value across the entire sample never reaches two, so a two-sigma rule cannot fire even in principle. This is itself a finding about signal-to-noise ratio rather than a defect of the rule. Granting every indicator an equal budget of twenty alarms produces the fairer comparison shown in Table V. Several network measures then outperform the volatility benchmarks on detection, since the signal-eigenvalue count identifies seventy-three per cent of episodes (with thirty alarms) and von Neumann entropy sixty-four per cent, against thirty-six per cent for realised volatility and nine per cent for the VIX, at median lead times of eighteen to thirty-two days. The precision column, however, governs the interpretation. The best value attained by any indicator is 0.35, which amounts to roughly two false alarms for every true one, and most sit between 0.10 and 0.25. No measure approaches an operating point upon which a supervisor could act.

TABLE IV. BUDGET-MATCHED DETECTION AND PRECISION

| Indicator | F | Alarms | Detected | Detection rate | Median lead (d) | False alarms | Precision |
|---|---|---|---|---|---|---|---|
| Signal eigenvalues | N | 30 | 8 | 0.73 | 18 | 22 | 0.27 |
| von Neumann entropy | N | 20 | 7 | 0.64 | 32 | 13 | 0.35 |
| MST length | N | 20 | 5 | 0.46 | 30 | 15 | 0.25 |
| Eigenvector rotation | N | 21 | 5 | 0.46 | 25 | 16 | 0.24 |
| Realised volatility | V | 20 | 4 | 0.36 | 48 | 16 | 0.20 |
| Mean correlation | N | 20 | 4 | 0.36 | 28 | 16 | 0.20 |
| $\lambda_1$ | N | 20 | 4 | 0.36 | 14 | 16 | 0.20 |
| Riemannian speed | N | 20 | 3 | 0.27 | 36 | 17 | 0.15 |
| MST diameter | N | 21 | 2 | 0.18 | 33 | 19 | 0.10 |
| VIX | V | 20 | 1 | 0.09 | 82 | 19 | 0.05 |
| Frobenius speed | N | 21 | 0 | 0.00 | 0 | 21 | 0.00 |

*F. Robustness and data snooping*

The conclusion does not depend on the specification. Across twenty-four combinations of window, horizon and threshold, the volatility baseline ranges from 0.579 to 0.814 and the best network indicator from 0.492 to 0.671. In twenty-three of the twenty-four some indicator beats the baseline, which is precisely what one expects when screening eleven noisy candidates and is why per-indicator consistency is the meaningful statistic. Taken individually, win rates run from 0.125 to 0.542, every median change in performance is at or below 0.0005, and no indicator beats the baseline in at least seventy-five per cent of specifications while also showing a positive median gain. Table VI reports the formal correction. Under White's Reality Check the probability value ranges from 0.739 to 0.928 across the four windows, and under Hansen's test from 0.174 to 0.927, so that the null of no superior predictive ability survives everywhere. Nor does this rest on the corrections being conservative: the largest mean improvement in Brier score across all eleven candidates is 0.0023 against a benchmark loss of 0.066, amounting to some three per cent and indistinguishable from chance.

TABLE V. REALITY CHECK AND SPA TEST

| W | Models | Benchmark Brier | Best mean gain | RC stat | p (RC) | SPA stat | p (SPA) | $H_0$ |
|---|---|---|---|---|---|---|---|---|
| 40 | 11 | 0.06656 | +0.00071 | 0.031 | 0.892 | 1.053 | 0.709 | not rejected |
| 60 | 11 | 0.06620 | +0.00112 | 0.049 | 0.928 | 0.529 | 0.927 | not rejected |
| 90 | 11 | 0.06583 | +0.00185 | 0.080 | 0.857 | 1.764 | 0.174 | not rejected |
| 120 | 11 | 0.06570 | +0.00226 | 0.098 | 0.739 | 0.737 | 0.754 | not rejected |

## VII. DISCUSSION

The results resolve into a single sentence: these indicators measure something real and distinct, and that something is not predictive. A median explained variance of 0.104 leaves most of their variation unexplained by the controls, and they are not redundant with one another: the rearrangement of a market's dependency structure captures something no volatility measure does. That information nonetheless fails to forecast drawdowns at five to twenty days, in any specification, under any correction. The reorganisation may be simultaneous with stress rather than prior to it, may lead but be swamped by sixty-day-window noise, or may matter for outcomes other than equity drawdowns. The sign instability is the most consequential result. An indicator that reverses orientation out of sample is worse than uninformative: it is deployed with a fitted direction and then points the wrong way when needed. That seven of eleven do so suggests period-specific artefacts full-sample calibration conceals. The budget-matched alarms temper this: several measures detect more episodes than the VIX at lead times near a month, and a precision of 0.35 may flag periods for human review, though not support a supervisory rule. The implication is a burden of proof. A new indicator should demonstrate incremental value over a volatility baseline, sign stability across a genuine holdout, and survival of a data-snooping correction. None of the eleven clears that bar. A measure validated only against crisis episodes, with hindsight-calibrated thresholds and no false-alarm accounting, has not truly been validated.

## VIII. LIMITATIONS

Six constraints bound these conclusions. The study examines one market and one target, being equity drawdowns in the United States, so that measures originating in bank-level or sovereign networks may behave differently in their native setting, and a node set with fewer U.S. assets or less liquid markets may produce a different topology. The data are daily, which renders intraday rearrangement invisible, although the $O(N^2)$ update in (14) carries over to intraday returns unchanged. Only correlation-based graphs are considered, which keeps all eleven indicators on one estimation pipeline but leaves Granger-causal [2], transfer-entropy [8] and tail-dependence [4] networks untested. Eleven indicators do not constitute a census of the field. The evaluation sample contains thirty-seven independent stress runs and eleven episodes, so that alarm statistics carry wide uncertainty: the best detection rate, 0.73, has a 95 % Wilson interval of 0.43 to 0.90, and detection rates should be read as indicative rather than precise. Finally, absence of evidence at these horizons is not evidence that no network measure can anticipate stress; what it establishes is that these measures, evaluated in this way, do not.

## IX. CONCLUSION

Eleven network indicators were placed on identical data under a single look-ahead-free protocol and asked a question the literature has largely avoided, namely whether they add anything to volatility. They do not. None improves upon a four-variable volatility baseline, seven reverse sign out of sample, all peak within three days of turbulence once window contamination is removed, and the null of no superior predictive ability survives both White's and Hansen's corrections at every estimation window. What these measures do possess is genuine distinctness from volatility and from one another, together with a real capacity to detect episodes when granted an equal alarm budget, at a precision no higher than 0.35. They are therefore descriptive instruments rather than anticipatory ones. The evidence here does not support their use as early-warning signals, and future proposals should clear the benchmark this study establishes.

## REFERENCES


[1] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi, "Systemic risk and stability in financial networks," *Amer. Econ. Rev.*, vol. 105, no. 2, pp. 564–608, 2015, doi: 10.1257/aer.20130456.

[2] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *J. Financ. Econ.*, vol. 104, no. 3, pp. 535–559, 2012, doi: 10.1016/j.jfineco.2011.12.010.

[3] F. X. Diebold and K. Yılmaz, "On the network topology of variance decompositions: Measuring the connectedness of financial firms," *J. Econometrics*, vol. 182, no. 1, pp. 119–134, 2014, doi: 10.1016/j.jeconom.2014.04.012.

[4] T. Adrian and M. K. Brunnermeier, "CoVaR," *Amer. Econ. Rev.*, vol. 106, no. 7, pp. 1705–1741, 2016, doi: 10.1257/aer.20120555.

[5] R. N. Mantegna, "Hierarchical structure in financial markets," *Eur. Phys. J. B*, vol. 11, no. 1, pp. 193–197, 1999, doi: 10.1007/s100510050929.

[6] R. D. Rajasinghe and D. Wijesinghe, "A Federated Multimodal Deepfake Detection Framework for Enhancing the Integrity of Video Evidence in Court Systems," 2026 9th International Conference on Electronics Technology (ICET), Chengdu, China, 2026, pp. 806-811, doi: 10.1109/ICET69987.2026.11659066.

[7] M. Scheffer *et al.*, "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, pp. 53–59, 2009, doi: 10.1038/nature08227.

[8] Ö. Akgüller and M. A. Balcı, "Detecting financial contagion through higher-order networks: A deep learning approach to emerging market risk," *Comput. Econ.*, 2026, doi: 10.1007/s10614-025-11287-3.

[9] S. Kumar, N. Umeorah, and A. Alochukwu, "Regime-dependent graph neural networks for enhanced volatility prediction," *Mathematics*, vol. 14, no. 2, Art. no. 289, 2026, doi: 10.3390/math14020289.

[10] W. D. J. I. Senarathna, "H. Dilpriya's Momentum (HDM): A multi-strategy gradient-aligned optimizer with adaptive per-parameter corrections, cosine-annealed scheduling, and convergence guarantees for deep neural networks," in *Proc. 4th Int. Conf. Adv. Netw. Technol. Appl. (APAN)*, Auckland, New Zealand, 2026, doi: 10.1109/APAN70967.2026.11709607.

[11] W. D. J. I. Senarathna and T. A. H. Dilpriya, "Deep learning based district level paddy production forecasting: A study based on agricultural behaviour in Sri Lanka," in *Proc. 6th Int. Conf. Adv. Res. Comput. (ICARC)*, Belihuloya, Sri Lanka, 2026, doi: 10.1109/ICARC68737.2026.11453891.

[12] W. D. J. I. Senarathna and T. A. H. Dilpriya, "A systematic comparative

evaluation of machine learning and deep learning models for EEG-based epileptic seizure prediction with clinical deployment guidelines," in *Proc. IEEE Int. Res. Conf. Smart Comput. Syst. Eng. (SCSE)*, Kelaniya, Sri Lanka, 2026, doi: 10.1109/SCSE70081.2026.11499844.

[13] Nobi, S. E. Maeng, G. G. Ha, and J. W. Lee, "Effects of global financial crisis on network structure in a local stock market," Physica A, vol. 407, pp. 135–143, 2014, doi: 10.1016/j.physa.2014.03.083.

[14] T. Squartini, I. van Lelyveld, and D. Garlaschelli, "Early-warning signals of topological collapse in interbank networks," *Sci. Rep.*, vol. 3, no. 1, 2013, doi: 10.1038/srep03357.

[15] H. Gatfaoui and P. de Peretti, "Flickering in information spreading precedes critical transitions in financial markets," *Sci. Rep.*, vol. 9, no. 1, 2019, doi: 10.1038/s41598-019-42223-9.

[16] M. Gidea and Y. Katz, "Topological data analysis of financial time series: Landscapes of crashes," *Physica A*, vol. 491, pp. 820–834, 2018, doi: 10.1016/j.physa.2017.09.028.

[17] S. L. Braunstein, S. Ghosh, and S. Severini, "The Laplacian of a graph as a density matrix: A basic combinatorial approach to separability of mixed states," *Ann. Comb.*, vol. 10, no. 3, pp. 291–317, 2006, doi: 10.1007/s00026-006-0289-3.

[18] A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sb.*, vol. 1, no. 4, pp. 457–483, 1967, doi: 10.1070/SM1967v001n04ABEH001994.

[19] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, "Noise dressing of financial correlation matrices," *Phys. Rev. Lett.*, vol. 83, no. 7, pp. 1467–1470, 1999, doi: 10.1103/PhysRevLett.83.1467.

[20] M. Parkinson, "The extreme value method for estimating the variance of the rate of return," *J. Bus.*, vol. 53, no. 1, pp. 61–65, 1980, doi: 10.1086/296071.

[21] M. B. Garman and M. J. Klass, "On the estimation of security price volatilities from historical data," *J. Bus.*, vol. 53, no. 1, pp. 67–78, 1980, doi: 10.1086/296072.

[22] W. D. J. I. Senarathna and T. A. H. Dilpriya, "Intelligent EcoETL: An energy-aware framework for sustainable extract-transform-load operations in big data processing," in *Selected Articles from the Int. Conf. Emerging Concepts and Research in Integrated Sustainable Adaptation*, D. Tantrigoda *et al.*, Eds. Singapore: Springer, 2026, doi: 10.1007/978-981-92-0423-6_11.

[23] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[24] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Statist.*, vol. 13, no. 3, pp. 253–263, 1995, doi: 10.1080/07350015.1995.10524599.

[25] H. White, "A reality check for data snooping," *Econometrica*, vol. 68, no. 5, pp. 1097–1126, 2000, doi: 10.1111/1468-0262.00152.

[26] P. R. Hansen, "A test for superior predictive ability," *J. Bus. Econ. Statist.*, vol. 23, no. 4, pp. 365–380, 2005, doi: 10.1198/073500105000000063.

[27] H. R. Künsch, "The jackknife and the bootstrap for general stationary observations," *Ann. Statist.*, vol. 17, no. 3, pp. 1217–1241, 1989, doi: 10.1214/aos/1176347265.

[28] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

# AI-Driven Digital Platform Ecosystems: Adaptive Business Model Innovation, Intelligent Ecosystem Orchestration, and Sustainable Value Creation

G. C. M. Sewwandi
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
gcmsewwandi@students.nsbm.ac.lk

T. A. H Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing, NSBM Green University*
*Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—This article develops a conceptual framework of how artificial intelligence (AI) capability may generate economic, social and environmental value in digital platform ecosystems. Returns to AI investment remain uneven, and existing research shows that AI, ecosystem orchestration and business model change are connected without specifying the architecture of that connection. Drawing on the dynamic capabilities perspective, AI capability is theorized as an orchestrator-level sensing capability whose effect is proposed to pass through two parallel mechanisms, adaptive business model innovation (seizing) and intelligent ecosystem orchestration (transforming), which are separated by observable indicators. Three claims are advanced as testable proposals rather than findings. Once both mechanisms are measured, the direct effect of AI capability on sustainable value should be negligible, a falsifiable account of the AI productivity paradox in platform settings. Under low AI governance maturity, orchestration should carry indirect effects of opposite sign to economic and social value, offering a reconciliation of optimistic AI–sustainability findings with critical research on algorithmic management. Orchestration also feeds back into AI capability, a loop that amplifies initial differences only under stated conditions. A stylized simulation demonstrates the internal consistency of the second and third claims. Eleven propositions, three scope conditions and a research agenda of testing designs, several requiring no firm access, are presented.**

***Keywords—Artificial intelligence, business model innovation, digital platform ecosystems, dynamic capabilities, ecosystem orchestration, sustainable value creation***

## I. Introduction

Digital platform ecosystems are a dominant organizational form of the digital economy. Firms such as Amazon, Airbnb, Alibaba and Grab do not primarily produce goods but orchestrate interactions among autonomous complementors, suppliers and users whose collective activity generates value no single firm could produce alone [1], [2]. What has changed is who orchestrates: matching, pricing and trust decisions once made by managers are now delegated to algorithmic systems, so that Uber's matching engine, Airbnb's Smart Pricing and Alibaba's credit scoring are not decision-support tools but the coordination mechanism itself. Where such ecosystems intermediate large shares of retail trade, mobility or financial access, whether they distribute value broadly or narrowly matters beyond firm performance. Yet returns to AI remain uneven, consistent with the intangible-complement dynamic of Brynjolfsson et al. [3], in which general-purpose technologies pay only after complementary organizational assets are rebuilt: possessing AI resources is not equivalent to realizing value. Recent work connects AI to platform ecosystems. Chin et al. [4] report an inverted U-shaped relationship between AI and the orchestration of ecosystem-based business models, moderated by knowledge spillover; Ragmoun and Aloulou [5] route AI capability to sustainable innovation through digital platform capabilities; and Zhang et al. [6] trace how a platform enterprise develops intelligent connectivity, development and governance capabilities across stages of AI integration. This work establishes that AI, orchestration and business model change are connected, but not the architecture of that connection: which organizational activities carry the effect, whether AI retains an independent effect once they are specified, and whether the value produced reaches actors beyond the orchestrator.

Four gaps follow. (i) The transmission structure is unspecified: existing studies treat business model change as an outcome of AI [4], [6], or route its effect through a single technological mediator [5]; none models adaptive business model innovation and ecosystem orchestration as parallel mechanisms, despite their correspondence to two distinct dynamic capabilities. (ii) The direct effect is assumed rather than interrogated: Chin et al. [4] report a curvilinear direct relationship, and adjacent mediation studies retain a direct path. (iii) Value is assumed uniformly distributed: studies linking AI to sustainability report positive associations [5], while research on algorithmic management documents harms from the very mechanisms generating those efficiencies [7], [8]. (iv) The relationship is modeled as linear, although data network effects imply that orchestration feeds back into the capability that produced it. This article asks through what mechanisms AI capability may influence sustainable value creation (SVC) in platform ecosystems, and under what conditions those mechanisms are strengthened, weakened or reversed in sign. Grounded in the dynamic capabilities perspective, it treats AI capability as enhanced sensing, adaptive business model innovation (ABMI) as seizing, and intelligent ecosystem orchestration (IEO) as transforming. Its novelty relative to [4]–[6] lies in three proposals, each offered for testing rather than as an established finding: the direct effect of AI capability becomes negligible once both mechanisms are measured (P5c); under

weak governance the orchestration path carries opposite-signed effects to economic and social value (P7b); and orchestration feeds back into AI capability, amplifying differences only under stated conditions (P8). Observable indicators separating the two mechanisms, eleven propositions (Fig. 1), three scope conditions, a stylized simulation and a testing agenda follow.

## II. Related Work

### *A. Platform ecosystems and AI as an organizational capability*

The platform literature has passed through three overlapping phases: two-sided markets whose intermediary sets a price structure rather than a price level [9]; layered modular systems coevolving with environmental dynamics [10]; and, drawn on here, the platform as nucleus of an ecosystem. Adner [11] defines an ecosystem as the alignment structure of partners who must interact for a focal value proposition to materialize; alignment is the central problem, since the actors are autonomous and cannot be directed by fiat. Jacobides et al. [2] trace the structure to non-generic complementarities, and Kretschmer et al. [12] characterize it as a meta-organization, exhibiting goals and rules without organizational authority. The consequence is decisive: the owner coordinates resources it does not own, applied by actors it does not employ, through rules and interfaces rather than instruction [13]. Coordination is continuously achieved rather than administratively given, which makes the technology exercising it consequential. AI assets do not themselves generate value: advantage arises from bundling them into a capability difficult to imitate [14], operationalized by Mikalef and Gupta [15] as data, technical infrastructure, human skills and organizational complements. Two limitations recur: the outcome studied is almost invariably economic, leaving distributional and environmental consequences unexamined; and the effect is theorized as direct, leaving unmodeled the organizational transformation that makes the technology pay. AI also differs from technologies executing predefined rules: it infers decision rules from data and revises them as data accumulate. Gregory et al. [16] develop this into data network effects (interaction generates data that trains models, which improve the service and attract further interaction), letting AI assume coordination tasks once requiring managerial judgment.

### *B. Business model adaptation and ecosystem orchestration*

A business model describes the content, structure and governance of the transactions through which a firm creates and captures value [17]; business model innovation denotes designed, non-trivial changes to that system [18], conventionally to its value proposition, delivery architecture and value capture mechanism [19]. ABMI is narrower: it concerns the speed and extent of reconfiguration in response to environmental change, not the novelty of any design. Saebi et al. [20] show that firms adapt to perceived threats, so gains in perceiving change propagate into the rate of reconfiguration. Platforms afford distinctive reconfiguration options (altering the price structure across sides, opening or closing the platform, adding a market side) without proportional investment in physical assets [1]. Orchestration denotes the deliberate actions by which a hub actor manages a network it does not control. Dhanaraj and Parkhe [21] specify three such processes (knowledge mobility, innovation appropriability and network stability), while the platform literature treats cognate concerns as governance over access, control and pricing [22] and as tuning variety and control through boundary resources [13]. In all these treatments orchestration is managerial and therefore coarse: applied uniformly, revised on organizational rather than transactional timescales, and rising in cost with ecosystem size. Contemporary platforms execute the same functions algorithmically, continuously and per participant. Research on algorithmic management documents the shift [7], [8] but addresses control and worker experience rather than value creation. Algorithmically executed orchestration therefore remains unformalized, the construct gap that IEO fills in Section IV.

### *C. Sustainable value and the dynamic capabilities lens*

Platform research takes the economic outcome as dominant, yet platforms are criticized for complementor income precarity, unappealable algorithmic decisions and the environmental externalities of on-demand fulfillment, which a purely economic framework cannot represent. Value is therefore assessed along economic, social and environmental dimensions [23], on three warrants: firms create value for constituencies not reducible to shareholders (stakeholder theory [24]); value is co-created with and determined by beneficiaries (service-dominant logic [25]), extending assessment to complementors; and performance on one dimension does not imply performance on another [26], so the three must be theorized as separable outcomes capable of diverging. Dynamic capabilities are the firm's capacity to integrate, build and reconfigure competences in rapidly changing environments [27], disaggregated by Teece [28] into sensing, seizing and transforming. The perspective suits this phenomenon because it explains adaptation under uncertainty, the defining condition of platform competition, and specifies how resources become advantage rather than asserting a direct link. One extension is required: formulated for the firm, where transforming redeploys owned assets under managerial authority, it must accommodate an orchestrator reconfiguring assets held by autonomous third parties through incentives, information and interface design.

TABLE I
DYNAMIC CAPABILITIES AT THE ECOSYSTEM LEVEL

| Dynamic capability | Firm-level meaning | Ecosystem-level manifestation | Construct in this article |
|---|---|---|---|
| Sensing | Scanning, filtering and interpreting opportunities and threats | Continuous inference from ecosystem-wide interaction data | AI capability |
| Seizing | Committing resources to a selected opportunity | Reconfiguring the value proposition and revenue architecture | Adaptive business model innovation |
| Transforming | Reconfiguring the owned resource base | Realigning autonomous actors through algorithmic incentives and interfaces | Intelligent ecosystem orchestration |

| Outcome | Sustained competitive advantage | Value distributed across economic, social and environmental dimensions | Sustainable value creation |
|---|---|---|---|

Warner and Wäger [29] show that digital transformation reshapes how the three are enacted; Table I specifies how each is enacted at ecosystem level when the mechanism is algorithmic. Dynamic capabilities explain how value is generated but are silent on how it is distributed, so a second lens is needed. Paradox theory holds that organizations face persistent contradictory demands that are managed continuously rather than resolved by trade-off [30]. The tension between economic and social value in algorithmically orchestrated ecosystems has this structure, since the mechanisms raising throughput also intensify participant monitoring; governance therefore does not remove the tension but determines whether it is held productively. These literatures supply the components but not their architecture, which the recent work joining them [4]–[6] leaves open.

## III. Methodology

This article is a theory-synthesis conceptual contribution [31], appropriate when independently developed literatures must be integrated to explain what none explains alone. Two grounds warrant it: the focal constructs (algorithmically executed orchestration, and value distributed across constituencies) are not yet specified precisely enough to measure, and the literatures return conflicting verdicts, a theoretical problem before an empirical one. The unit of analysis is the platform ecosystem viewed from its orchestrator. Levels are kept separate: AI capability, ABMI and IEO are orchestrator-level constructs, whereas SVC is an ecosystem-level outcome aggregated across the orchestrator, complementors, users and the natural environment, so the propositions are cross-level. Literature was identified in four steps. (i) Search: Scopus and Web of Science were queried in titles, abstracts and keywords with ("digital platform*" OR "platform ecosystem*" OR "digital ecosystem*") AND ("artificial intelligence" OR "machine learning" OR "algorithmic") AND ("business model innovation" OR "orchestration" OR "value creation" OR "sustainab*"). (ii) Screening: inclusion required a peer-reviewed article or scholarly book in English, in management or information systems, that defines, theorizes or tests a focal construct or a link between two; studies of model architectures, user adoption or purely technical performance, and conference abstracts, were excluded. (iii) Snowballing: backward and forward citation tracing (Google Scholar) from retained studies and field-defining works recovered earlier founding statements. (iv) Retention: a source was kept only if it supplied a construct definition, mechanism, warrant or competing account used in the argument. Each construct was defined from a primary source and mapped onto one dynamic capability (Table I); each proposition states a signed relationship and its mechanism, assessed against Whetten's [32] criteria (which constructs, how related, why, and within what limits).

The framework is positioned against three competing accounts: the curvilinear direct effect of Chin et al. [4], the single technological mediator of Ragmoun and Aloulou [5], and the stage account of Zhang et al. [6]. Orchestration is used in the strategy sense of a hub actor managing a network it does not control, not in the control sense of the algorithmic management literature. AI capability and its dimensions were adopted from [15]; the coordination problem and orchestration processes from the ecosystem and platform governance literatures [11], [13], [21], [22]; and the tripartite outcome and its warrants from the sustainable value literature [23]–[26]. IEO alone was constructed here, by re-specifying orchestration processes identified for managerial hubs [21]. It is defined as the orchestrator's capacity to coordinate autonomous ecosystem actors through matching, monitoring, incentive and knowledge functions executed algorithmically, continuously and per participant (Section IV-A). Both moderators follow from the theories used: environmental dynamism is the condition on which dynamic capabilities theory is itself conditional, and AI governance maturity is the construct through which the algorithmic management findings [7], [8] enter the model. The scope conditions address Whetten's fourth criterion, one per source of variation the framework cannot absorb. Sustainable value is kept as three separable constructs rather than a composite, because P7b predicts its dimensions can move in opposite directions. Consistency was checked by tracing each construct to a primary source and assigning each proposition a mechanism and each scope condition a sampling requirement. The propositions are directional expectations offered for testing, not findings; the internal consistency of P7b and P8 is probed by simulation in Section IV-G.

### *A. The two transmission mechanisms*

Business model change is inhibited by uncertainty about whether change is warranted: managers delay reconfiguration because environmental signals are noisy and a mistaken pivot is costly. AI capability converts weak, dispersed signals into decision-relevant patterns earlier than human scanning permits, lowering the cost and latency of sensing and so raising the feasible frequency of seizing.

The 2020 mobility collapse illustrates the mechanism: a ride-hailing platform with mature demand forecasting could detect the shift from mobility to delivery demand within weeks and launch food and parcel services on existing driver supply, reconfiguring its value proposition and revenue model without new physical assets. P1 states that AI capability is positively associated with ABMI. Orchestration faces a scale problem that hierarchy does not: the orchestrator must align thousands of autonomous actors it cannot instruct, and classical solutions (rules, contracts, reputation systems) apply uniformly and adapt slowly. AI capability enables orchestration that is granular, continuous and personalized across four functions: (i) matching, from proximity rules to multi-objective optimization; (ii) monitoring, from ex-post complaint handling to real-time anomaly detection; (iii) incentives, from uniform pricing to individually targeted rewards; and (iv) knowledge mobility, from broadcast documentation to personalized algorithmic advice. Each realigns the ecosystem's resource configuration without owning the resources, a transforming activity. P2 states that AI capability is positively associated with IEO.

## IV. Proposed Conceptual Framework

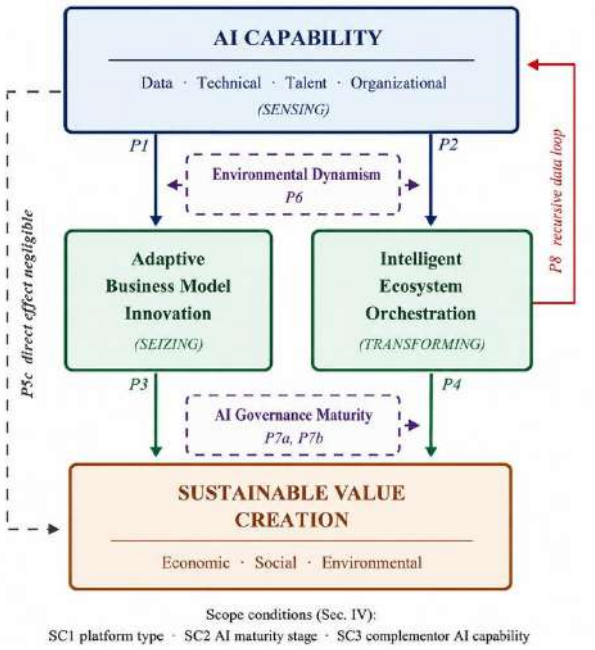



Fig. 1. The proposed conceptual framework.

As defined in Section III, IEO is distinct from four neighboring concepts: managerial orchestration [21], designed and revised periodically by managers; algorithmic management [7], [8], which concerns the control of workers rather than value creation across an ecosystem; platform governance [22], which sets the rules IEO applies; and AI capability, the resource stock IEO deploys. Because pricing and governance changes can belong to either mechanism, Table II gives a decision rule with observable indicators: a change to the structure of the business model (what is offered, to whom, and which side pays) is ABMI, whereas a change in how an unchanged structure is applied to individual participants is IEO.

TABLE II
DISTINGUISHING THE TWO MECHANISMS

| | ABMI (seizing) | IEO (transforming) |
|---|---|---|
| Object | Structure: offer, market sides, revenue model | Application: matching, pricing, monitoring, advice per participant |
| Tempo | Episodic; platform-wide | Continuous; per transaction |
| Pricing | New subscription tier; change of paying side | Surge pricing under an unchanged fee schedule |
| Governance | Opening to a new complementor class | Real-time anomaly detection, automated suspension |
| Indicators | Count and scope of offer, revenue-model and market-side changes; signal-to-launch lag | Share of the four functions automated; decision frequency; personalization |

### *A. The mechanisms and sustainable value creation*

A rigid business model transfers environmental shocks onto the weakest actors, as complementors absorb demand collapse as income loss, whereas an adaptive one redeploys existing capacity toward newly viable demand: new revenue streams and reduced dependence on one demand source (economic), continuity of complementor livelihoods (social), and higher utilization of existing assets rather than new capacity (environmental). P3 states that ABMI is positively associated with SVC across economic, social and environmental dimensions. Orchestration also determines the distribution of value, not only its magnitude. Precise matching reduces idle capacity, which raises throughput (economic) and complementor earnings per hour worked (social) and lowers energy consumed per transaction (environmental), while algorithmic trust mechanisms spare small complementors disproportionate transaction losses. P4 states, in parallel, that IEO is positively associated with SVC across the same three dimensions. Indicative measures are orchestrator revenue and throughput (economic), complementor earnings per hour, earnings stability and deactivation rates (social), and energy use or emissions per transaction (environmental). Where efficiency gains induce additional demand, rebound effects may offset total environmental gains [33], so P3 and P4 predict lower intensity per transaction, not necessarily a lower total footprint.

### *B. The mediating structure and the direct effect*

AI capability is a stock of resources and skills, and resources create value only when deployed in activity. In a platform ecosystem the activities that deploy it at scale are business model reconfiguration and orchestration; absent those, AI capability yields analytical output that changes nothing about how the ecosystem operates. This offers an explanation of the AI productivity paradox in this setting: firms that invest in AI without reconfiguring their business model or orchestration should observe little value effect. P5a states that ABMI mediates the relationship between AI capability and SVC; P5b states that IEO does so in parallel. P5c states that, once both mediators are measured, the direct effect of AI capability on SVC is negligible, i.e., indirect-only (full) mediation in the terms of Zhao et al. [34]. Negligible is defined substantively, not as non-significance, which may merely reflect low power: the direct effect should fall within a pre-specified smallest effect size of interest, shown by equivalence testing [35], and be small relative to the total indirect effect. The exhaustiveness assumption behind P5c is qualified rather than presumed: it covers value created through the ecosystem, whereas AI may also create value outside both mechanisms, e.g., through back-office cost reduction, which must be measured or controlled separately. Within that scope, a direct effect surviving measurement of both mediators and such rival channels would falsify P5c; otherwise, an apparent direct effect, linear or curvilinear [4], may reflect an omitted mechanism whose variance loads onto the AI term.

### *C. Environmental dynamism and AI governance maturity as moderators*

Dynamic capabilities theory is explicitly conditional on environmental change: the option value of rapid reconfiguration is low in stable environments, whereas in dynamic environments the capability is exercised repeatedly and its

returns accumulate. P6 states that environmental dynamism positively moderates the relationships between AI capability and both mechanisms, so that P1 and P2 hold more strongly under high dynamism. Algorithmic orchestration optimizes whatever objective it is given. Where that objective is economic alone, orchestration can raise economic value while degrading social value, through opaque deactivation, discriminatory matching or intensified work pace. AI governance maturity (fairness auditing, explainability, human recourse and data protection compliance) is proposed to determine whether orchestration produces value across all three dimensions or concentrates it in one. P7a states that AI governance maturity positively moderates the relationship between IEO and the social and environmental dimensions of SVC. This moderator also offers a way to reconcile two literatures. Studies linking AI to sustainability report broadly positive associations [5], while research on algorithmic management reports that the same mechanisms produce information asymmetry, earnings unpredictability and loss of recourse [7], [8]. The framework proposes that both describe one mechanism under two governance regimes, and at full strength that the signs diverge rather than the positive effect merely weakening. In mediation terms this is moderated mediation [36]: governance maturity moderates the second stage of the orchestration path, so that under low governance the conditional indirect effect of AI capability via IEO is positive on economic value and negative on social value. A composite sustainability measure may therefore net the two effects out. P7b states that, under low AI governance maturity, IEO has a positive effect on economic value and a simultaneously negative effect on social value, so that the indirect paths from AI capability to these two dimensions carry opposite signs.

### *D. The recursive dynamic*

The framework as stated so far is linear. Data network effects imply that it is also recursive: every matched transaction, accepted price and resolved dispute generates labeled interaction data that improves the models on which subsequent orchestration depends [16], and ABMI feeds the same loop by opening interaction surfaces that generate new categories of data. Feedback alone, however, does not imply persistent divergence: if returns to data diminish, laggards catch up and ecosystems converge (Section IV-G). Initial differences in AI capability amplify only if three further conditions hold: (a) marginal returns to data do not saturate, because the environment keeps changing or models expand in scope; (b) interaction data are exclusive, as users and complementors do not multi-home and data are not portable to rivals; and (c) the loop's capability gain exceeds the rate at which data and models depreciate. Where they hold, data concentration also concentrates power over complementors. P8 states that IEO is positively associated with subsequent AI capability, forming a self-reinforcing loop, and that under conditions (a)–(c) initial differences in AI capability among platform ecosystems amplify over time. Table III summarizes the propositions.

TABLE III
SUMMARY OF PROPOSITIONS

| # | Proposed relationship | Role in the framework |
|---|---|---|
| P1 | AI capability → ABMI (+) | Sensing enables seizing |
| P2 | AI capability → IEO (+) | Sensing enables transforming |
| P3 | ABMI → SVC, all three dimensions (+) | Seizing pathway to value |
| P4 | IEO → SVC, all three dimensions (+) | Transforming pathway to value |
| P5a | AI capability → SVC via ABMI | Indirect path 1 |
| P5b | AI capability → SVC via IEO | Indirect path 2 |
| P5c | Direct AI capability → SVC path negligible (equivalence test) | Indirect-only mediation; central falsifiable claim |
| P6 | Dynamism strengthens P1 and P2 (+) | Condition on transmission |
| P7a | Governance maturity strengthens IEO → social, environmental (+) | Boundary condition on P4 |
| P7b | Low governance: IEO → economic (+), social (−) | Moderated mediation; opposite-signed indirect effects |
| P8 | IEO → subsequent AI capability (+); differences amplify under conditions (a)–(c) | Recursive loop |

ABMI = adaptive business model innovation; IEO = intelligent ecosystem orchestration; SVC = sustainable value creation.

### *E. Scope conditions*

Whetten's [32] fourth criterion requires a theory to state where it holds. Three scope conditions bound the framework's applicability. Table IV names each condition and its consequence; each implies a sampling requirement (Section VII).

TABLE IV
SCOPE CONDITIONS

| Condition | Expected consequence |
|---|---|
| SC1 Platform type - transaction (high-frequency, short-cycle exchange) vs innovation (long-cycle co-creation through boundary resources) | P2, P4 stronger in transaction platforms, where orchestration decisions are continuous; P1, P3 dominate in innovation platforms |
| SC2 AI maturity stage - assisted, augmented, integrated [6] | Business model path dominates early, orchestration path once AI is routine; P5c clearest at the integrated stage |
| SC3 Complementor AI capability - held by complementors, not only the orchestrator | Where high, the orchestrator's informational advantage narrows and P7b divergence attenuates; ecosystem AI capability is a distribution, not a scalar |

### *F. Simulation probe of P7b and P8*

To probe whether P7b and P8 are internally consistent, rather than to test them, a stylized model of one ecosystem was run for 40 periods (1,000 requests and 200 complementors per period, take rate 0.20, initial AI capability A = 0.20, learning rate 0.15). Match quality rises with A, raising completed

transactions and orchestrator revenue (economic value); complementor earnings are concentrated in proportion to (1 − G)A, where G is governance maturity, and typical complementor earnings measure social value; each transaction is training data that raises next-period A with diminishing returns. Economic value rises by 66% at every G, whereas social value falls by 99.7%, 51% and 8% at G = 0, 0.25 and 0.5 and rises by 31% and 66% at G = 0.75 and 1, changing sign near G = 0.55, consistent with P7b. With the loop switched off, A stays at 0.20 and throughput is 40% lower, consistent with P8. Because returns to data diminish, an initial capability gap of 0.10 between otherwise identical ecosystems shrinks to 0.03 by period 10 and below 0.001 by period 40, and a scale-driven gap (5,000 versus 100 requests) peaks at 0.20 in period 8 before closing to 0.01: without conditions (a)–(c), feedback yields convergence. The results follow from the encoded assumptions and show consistency, not empirical support; the model code is available from the authors on request.

## V. Discussion

The framework offers five proposed contributions. (i) It specifies the transmission architecture as two parallel mechanisms, where Chin et al. [4] posit a curvilinear direct effect, Ragmoun and Aloulou [5] a single technological mediator and Zhang et al. [6] a stage sequence; P5c makes the difference testable. (ii) It extends dynamic capabilities to the ecosystem level, where transforming operates over resources the firm does not own and by algorithmic rather than hierarchical means (Table I). (iii) It defines IEO, described elsewhere only in process or stage terms [6], as a four-function construct with indicators separating it from ABMI (Table II). (iv) Through the opposite-signed indirect effects of P7b, it proposes how optimistic findings on AI and sustainability and critical findings on algorithmic management could both hold, governance holding that tension rather than resolving it [30]. (v) It makes the account dynamic while stating when the loop of P8 would produce concentration rather than the convergence that diminishing returns to data imply. If supported, P5c implies that AI expenditure unmatched by business model and orchestration redesign should not be expected to pay. P7b implies, for complementors, that the worth of participating turns on the orchestrator's governance maturity, observable through appeal mechanisms, disclosed ranking criteria and stable incentives; and, for policymakers, that transparency, auditability and recourse may be conditions for the broad realization of platform productivity rather than constraints upon it, an efficiency argument for regimes such as the EU AI Act.

## VI. Limitations

Four limitations bound the contribution. Conceptual papers advance testable logic, not evidence: every proposition remains empirically untested, and the simulation shows only that P7b and P8 follow from their assumptions. The decision rule of Table II separates ABMI from IEO conceptually, but its discriminant validity must still be established in measurement. Rival explanations, such as market power, network size or regulatory exposure, may produce value differences attributed here to AI capability and must be controlled in tests. And the framework adopts the orchestrator's vantage point, where a complementor-centered framework might specify different mechanisms. Each limitation maps to a direction in Section VII.

## VII. Conclusion and Future Work

AI is becoming the coordinating mechanism of platform ecosystems, not a tool within them. The framework proposes that its value is not automatic: it is realized through business model reconfiguration and ecosystem orchestration, and broadly shared only where governance maturity carries it to complementors and the environment as well as the orchestrator. Designs capable of testing these proposals are ordered by required access. Three need none: (i) a governance maturity index built by NLP from terms of service, ranking disclosures and appeal policies (P7a, P7b); (ii) calibration of the Section IV-G model to platform data while varying conditions (a)–(c) (P8); and (iii) longitudinal archival analysis of platform policy change through web archives (P8, SC2). With access: (iv) fsQCA over 30–50 cases (P5c, P7b, SC1); (v) validation of the Table II indicators by expert panel and discriminant validity tests (P2, P4); (vi) a survey with PLS-SEM stratified by platform type and maturity stage, using equivalence tests for P5c (P1–P6, SC1–SC3); (vii) a multi-wave panel with cross-lagged models to establish the temporal order P8 asserts; and (viii) a dyadic design of matched orchestrator–complementor pairs, or complementor earnings and deactivation records (P7b). Two requirements cut across the agenda: economic, social and environmental value must be modeled separately, since a composite index nets out the opposing signs of P7b; and the social dimension, accruing to complementors, is invisible to orchestrator-only designs. None of the propositions is established here; each is stated precisely enough to be tested, which is what a conceptual article contributes: not settling the question but formulating it so that empirical work can answer it.

## REFERENCES


[1] G. G. Parker, M. W. Van Alstyne, and S. P. Choudary, Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You. New York, NY, USA: W. W. Norton, 2016.

[2] M. G. Jacobides, C. Cennamo, and A. Gawer, "Towards a theory of ecosystems," Strategic Manage. J., vol. 39, no. 8, pp. 2255-2276, 2018.

[3] E. Brynjolfsson, D. Rock, and C. Syverson, "The productivity J-curve: How intangibles complement general purpose technologies," Amer. Econ. J., Macroecon., vol. 13, no. 1, pp. 333-372, 2021.

[4] T. Chin, M. W. A. Ghouri, J. Jin, and M. Deveci, "AI technologies affording the orchestration of ecosystem-based business models: The moderating role of AI knowledge spillover," Humanit. Soc. Sci. Commun., vol. 11, no. 1, art. 496, 2024.

[5] W. Ragmoun and W. J. Aloulou, "Artificial intelligence capabilities, sustainable innovation and SME resilience: A serial-parallel mediation model of dynamic and digital platform capabilities," Sustainability, vol. 18, no. 9, art. 4320, 2026.

[6] Z. Zhang, Y. Kang, Y. Lu, and P. Li, "The role of artificial intelligence in business model innovation of digital platform enterprises," Systems, vol. 13, no. 7, art. 507, 2025.

[7] K. C. Kellogg, M. A. Valentine, and A. Christin, "Algorithms at work: The new contested terrain of control," Acad. Manage. Ann., vol. 14, no. 1, pp. 366-410, 2020.

[8] M. Möhlmann, L. Zalmanson, O. Henfridsson, and R. W. Gregory, "Algorithmic management of work on online labor platforms: When matching meets control," MIS Quart., vol. 45, no. 4, pp. 1999-2022, 2021.

[9] J.-C. Rochet and J. Tirole, "Platform competition in two-sided markets," J. Eur. Econ. Assoc., vol. 1, no. 4, pp. 990-1029, 2003.

[10] Tiwana, B. Konsynski, and A. A. Bush, "Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics," Inf. Syst. Res., vol. 21, no. 4, pp. 675-687, 2010.

[11] R. Adner, "Ecosystem as structure: An actionable construct for strategy," J. Manage., vol. 43, no. 1, pp. 39-58, 2017.

[12] T. Kretschmer, A. Leiponen, M. Schilling, and G. Vasudeva, "Platform ecosystems as meta-organizations: Implications for platform strategies," Strategic Manage. J., vol. 43, no. 3, pp. 405-424, 2022.

[13] J. Wareham, P. B. Fox, and J. L. Cano Giner, "Technology ecosystem governance," Organ. Sci., vol. 25, no. 4, pp. 1195-1215, 2014.

[14] J. Barney, "Firm resources and sustained competitive advantage," J. Manage., vol. 17, no. 1, pp. 99-120, 1991.

[15] P. Mikalef and M. Gupta, "Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance," Inf. Manage., vol. 58, no. 3, art. 103434, 2021.

[16] R. W. Gregory, O. Henfridsson, E. Kaganer, and H. Kyriakou, "The role of artificial intelligence and data network effects for creating user value," Acad. Manage. Rev., vol. 46, no. 3, pp. 534-551, 2021.

[17] R. Amit and C. Zott, "Value creation in e-business," Strategic Manage. J., vol. 22, no. 6-7, pp. 493-520, 2001.

[18] N. J. Foss and T. Saebi, "Fifteen years of research on business model innovation," J. Manage., vol. 43, no. 1, pp. 200-227, 2017.

[19] D. J. Teece, "Business models and dynamic capabilities," Long Range Plann., vol. 51, no. 1, pp. 40-49, 2018.

[20] T. Saebi, L. Lien, and N. J. Foss, "What drives business model adaptation? The impact of opportunities, threats and strategic orientation," Long Range Plann., vol. 50, no. 5, pp. 567-581, 2017.

[21] Dhanaraj and A. Parkhe, "Orchestrating innovation networks," Acad. Manage. Rev., vol. 31, no. 3, pp. 659-669, 2006.

[22] Gawer and M. A. Cusumano, "Industry platforms and ecosystem innovation," J. Prod. Innov. Manage., vol. 31, no. 3, pp. 417-433, 2014.

[23] J. Elkington, Cannibals with Forks: The Triple Bottom Line of 21st Century Business. Oxford, U.K.: Capstone, 1997.

[24] R. E. Freeman, Strategic Management: A Stakeholder Approach. Boston, MA, USA: Pitman, 1984.

[25] S. L. Vargo and R. F. Lusch, "Institutions and axioms: An extension and update of service-dominant logic," J. Acad. Marketing Sci., vol. 44, no. 1, pp. 5-23, 2016.

[26] S. L. Hart and M. B. Milstein, "Creating sustainable value," Acad. Manage. Perspect., vol. 17, no. 2, pp. 56-67, 2003.

[27] J. Teece, G. Pisano, and A. Shuen, "Dynamic capabilities and strategic management," Strategic Manage. J., vol. 18, no. 7, pp. 509-533, 1997.

[28] J. Teece, "Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance," Strategic Manage. J., vol. 28, no. 13, pp. 1319-1350, 2007.

[29] K. S. R. Warner and M. Wäger, "Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal," Long Range Plann., vol. 52, no. 3, pp. 326-349, 2019.

[30] W. K. Smith and M. W. Lewis, "Toward a theory of paradox: A dynamic equilibrium model of organizing," Acad. Manage. Rev., vol. 36, no. 2, pp. 381-403, 2011.

[31] Jaakkola, "Designing conceptual articles: Four approaches," AMS Rev., vol. 10, no. 1-2, pp. 18-26, 2020.

[32] D. A. Whetten, "What constitutes a theoretical contribution?," Acad. Manage. Rev., vol. 14, no. 4, pp. 490-495, 1989.

[33] S. Sorrell, "Jevons' paradox revisited: The evidence for backfire from improved energy efficiency," Energy Policy, vol. 37, no. 4, pp. 1456-1469, 2009.

[34] X. Zhao, J. G. Lynch, and Q. Chen, "Reconsidering Baron and Kenny: Myths and truths about mediation analysis," J. Consum. Res., vol. 37, no. 2, pp. 197-206, 2010.

[35] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," Soc. Psychol. Personal. Sci., vol. 8, no. 4, pp. 355-362, 2017.

[36] K. J. Preacher, D. D. Rucker, and A. F. Hayes, "Addressing moderated mediation hypotheses: Theory, methods, and prescriptions," Multivariate Behav. Res., vol. 42, no. 1, pp. 185-227, 2007.

# Predicting Bitcoin Price Movements from Multi-Source News Sentiment Using an Exponential Decay Market Memory Framework

L. A. D. K. S. Deshappriya
*Dept. of Software Engineering and Computer Security*
*Faculty of Computing,*
*NSBM Green University*
*Homagama, Sri Lanka*
ksdeshappriya.official@gmail.com

Samsudeen Ashad
*Dept. of Software Engineering and Computer Security*
*Faculty of Computing,*
*NSBM Green University*
*Homagama, Sri Lanka*
samsudeenashad@gmail.com

T. A. H. Dilpriya
*Department of Computer and Data Science*
*Faculty of Computing,*
*NSBM Green University*
*Homagama, Sri Lanka*
hirushi.d@nsbm.ac.lk

***Abstract*—Predicting cryptocurrency price movements from media sentiment requires modeling how rapidly market participants incorporate news into prices, as well as accounting for the temporal decay of information. This paper presents an end-to-end, modular, and reproducible research pipeline for evaluating the predictive utility of professional cryptocurrency journalism for Bitcoin direction and return forecasting. The system integrates five core stages: (i) an automated preprocessing pipeline that standardizes, deduplicates, and UTC-aligns 52,071 articles from four leading cryptocurrency news outlets (Bitcoin Magazine, CoinDesk, CoinTelegraph, and U.Today) spanning 2012–2026; (ii) comparative sentiment scoring using the lexicon-based VADER algorithm and the domain-adapted FinBERT deep learning model, establishing their concordance and computational trade-offs; (iii) a formal exponential time-decay market memory model with parameterized half-lives that formalizes information fading across fast ($\lambda = 0.20$, $t1/2 \approx 3.5$ days) and slow ($\lambda = 0.05$, $t1/2 \approx 13.9$ days) attention windows; (iv) supervised machine learning (Linear Regression, Random Forest, and XGBoost) evaluated under strict chronological out-of-sample partitioning and 5-fold expanding-window walk-forward validation across diverse market regimes; and (v) an out-of-sample trading backtest incorporating realistic market frictions (transaction fees of 0 to 25 bps, slippage, and spread). On a 772-day out-of-sample test window, the XGBoost classifier achieved a directional accuracy of 51.94% (one-sided binomial test $p = 0.148$ vs. 50% baseline; 95% Wilson CI [48.42%, 55.45%]), an UP-class F1 score of 0.607, and a macro-averaged F1 score of 0.495. While a frictionless long-cash trading strategy generated a 16.31% cumulative return versus 2.82% for buy-and-hold, empirical stress-testing reveals that high strategy turnover (271 trades) erodes outperformance once realistic transaction fees exceeding 5 bps are applied. Ablation experiments demonstrate that raw daily sentiment without exponential decay severely degrades strategy returns (−40.19%), verifying that exponential memory modeling is essential for extracting constructive predictive signals from noisy news flows.***

***Index Terms*—Bitcoin, sentiment analysis, natural language processing, exponential decay***

## I. Introduction

Bitcoin, introduced by Nakamoto [1], operates in a decentralized, 24/7 global market characterized by the absence of conventional valuation anchors such as corporate earnings or discounted cash flows [2]. In this informationally fragmented environment, investor sentiment, regulatory announcements, and thematic narratives transmitted through specialized journalism serve as primary drivers of market attention and short-term price formation [3], [4], [5].

While empirical studies establish that media sentiment correlates with cryptocurrency trading volume and returns [5], [6], [7], [8], practical predictive modeling faces four major methodological challenges:

1) *Information Memory and Temporal Decay:* The economic relevance of news fades as market participants digest information. Static representations that evaluate news only on the publication day fail to differentiate between transient headline shocks and sustained thematic narratives [9].

2) *Information Timing and Look-Ahead Bias:* Continuous trading requires rigorous temporal synchronization. Failure to enforce explicit daily publication cutoffs risks aggregating news released after trade execution, creating severe look-ahead leakage [10].

3) *Realistic Market Frictions:* Many published strategies report strong gross returns under frictionless assumptions. In live cryptocurrency markets, exchange fees (10–20 bps), bid-ask spreads, and slippage rapidly erode trading alpha under high portfolio turnover [11].

4) *Statistical Precision and Regime Instability:* Financial time series possess low signal-to-noise ratios. A modest accuracy edge requires formal hypothesis testing and multi-regime walk-forward cross-validation to rule out sample-specific overfitting [12].

To address these challenges, this study presents a fully reproducible, modular system for evaluating multi-source cryptocurrency news sentiment. Our principal contributions are:

**Curated Multi-Source Corpus:** Assembly and UTC-standardization of 52,071 articles from four major crypto outlets (2012–2026) with explicit deduplication and documented missing-news handling.

**Exponential Market Memory Framework:** Mathematical formalization of fast ($\lambda = 0.20$, $t1/2 \approx 3.5$ days) and slow ($\lambda = 0.05$, $t1/2 \approx 13.9$ days) normalized memory channels with half-life parameterization.

**FinBERT vs. VADER Concordance Analysis:** Quantitative comparison of lexicon and transformer scoring architectures across concordance, runtime throughput, and domain transfer.

**Ablation and Multi-Regime Validation:** Systematic ablation isolating exponential memory against raw sentiment and alternative decay functions, accompanied by 5-fold expanding-window walk-forward validation across five market regimes.

**Friction-Adjusted Economic Backtesting:** Stress-testing strategy performance across fee tiers (0 to 25 bps) and demonstrating the economic impact of turnover.

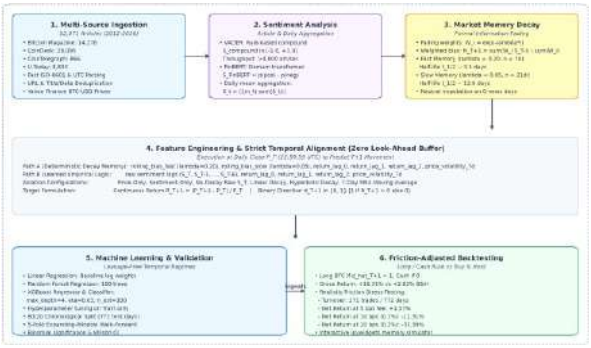



Fig. 1. END-TO-END SYSTEM ARCHITECTURE OF THE SENTIMENT-BASED BITCOIN PRICE PREDICTION PIPELINE, DETAILING THE FIVE MODULAR STAGES FROM RAW MULTI-SOURCE NEWS ACQUISITION TO FRICTION-ADJUSTED ECONOMIC BACKTESTING AND WALK-FORWARD CROSS-VALIDATION.

## II. RELATED WORK

### *A. Media Sentiment and Financial Asset Pricing*

Tetlock [3] demonstrated that quantitative linguistic pessimism in financial columns predicts downward price pressure followed by mean reversion. Loughran and McDonald [6] proved that general-purpose lexicons frequently misclassify financial terms, establishing the necessity for domain adaptation. Barber and Odean [7] formulated the attention-driven buying hypothesis, showing retail investors buy attention-grabbing assets highlighted in media. In cryptocurrencies, Kraaijeveld and De Smedt [5] established predictive relationships between public sentiment and returns. However, uncurated social media platforms are vulnerable to promotional spam and bot networks [8]. Dedicated professional journalism offers vetted editorial curation and substantially higher signal-to-noise ratios.

### *B. NLP: Lexicons to Financial Transformers*

VADER [13] provides rule-based sentiment scoring tuned for informal and concise texts using grammatical heuristics. In financial NLP, Araci [14] introduced FinBERT, pre-trained on corporate SEC filings (Form 10-K, earnings transcripts). However, corporate accounting language differs markedly from cryptocurrency vernacular (e.g., “halving”, “hard fork”, “burning”, “airdrop”). Un-adapted financial transformers can exhibit domain-shift distortion on crypto texts [15], motivating empirical concordance analysis against robust lexicon baselines. Hossain et al. [16] and Dashtaki et al. [17] explored hybrid deep learning architectures, while Moradi-Kamali et al. [18] investigated sentiment-derived cryptocurrency forecasting.

### *C. Market Memory and Time-Series Evaluation*

Engle and Ng [9] formalized asymmetric information decay in financial volatility. In machine learning, Bergmeir and Benítez [10] proved that standard k-fold cross-validation introduces severe temporal look-ahead bias in autoregressive settings, requiring strict chronological partitioning or walk-forward validation. Furthermore, López de Prado [11] emphasized that high-turnover strategies that do not deduct transaction costs report spurious statistical alpha that collapses in live execution. Binkowski and Silva [19] examined market memory in crypto trading, while Gurgul et al. [12] integrated multi-modal sentiment.

## III. SYSTEM ARCHITECTURE AND METHODOLOGY

The system architecture comprises five integrated modules depicted in Fig. 1.

### *A. Corpus Acquisition and Preprocessing*

Articles were harvested across four leading cryptocurrency publications spanning 2012 to 2026. The preprocessing pipeline applies: (i) HTML tag stripping and character decoding; (ii) deduplication based on exact and fuzzy title matching; (iii) UTC datetime parsing; and (iv) publication filtering. Table I summarizes the corpus metrics. Of the 4,277 calendar days spanning September 2014 to June 2026, 3,868 days possess news coverage. On the 409 days lacking published articles (predominantly in 2014–2015), the daily sentiment score is assigned neutral (St = 0.0), allowing exponential memory to decay smoothly without injecting artificial sentiment shocks.

TABLE I. CORPUS PREPROCESSING AND ARTICLE RETENTION SUMMARY

| Outlet | Raw Articles | Retained | Retention | Date Span |
|---|---|---|---|---|
| Bitcoin Magazine | 3,745 | 3,421 | 91.3% | 2012–2026 |
| CoinDesk | 15,230 | 14,812 | 97.3% | 2013–2026 |
| CoinTelegraph | 24,198 | 23,605 | 97.6% | 2013–2026 |
| U.Today | 10,614 | 10,233 | 96.4% | 2017–2026 |
| **Total / Overall** | **53,787** | **52,071** | **96.8%** | **2012–2026** |

### *B. Information Timing and Zero-Leakage Guarantee*

To guarantee zero look-ahead bias, daily sentiment S̄T aggregates only articles whose publication timestamp tpub satisfies:

$$00{:}00{:}00\ \text{YTX} \le t_{\pi\upsilon\beta} \le 23{:}59{:}59\ \text{YTX}\ \text{ο}\nu\ \Delta\alpha\psi\ \text{T}.$$

The feature vector assembled at the close of day T forecasts the next-day close-to-close return RT+1:

$$R_{T+1} = \frac{P_{T+1}-P_T}{P_T}, d_{T+1} = I(R_{T+1} > 0)$$

where PT is the BTC-USD closing price at 23:59:59 UTC. Hypothetical trades execute at PT and are held through T+1.

*C. Sentiment Scoring: VADER vs. FinBERT*

For each headline, sentiment polarity is scored via two models:

- **VADER Compound Polarity:** Lexicon-based valence metric scored via:

$$S_{\text{çAΔEP}} = \frac{\Sigma v_i}{\sqrt{(\Sigma v_i)^2 + \alpha}}, \alpha = 15$$

where vi is valence adjusted for punctuation and capitalization. VADER processes >4,800 headlines/sec on commodity CPU.

- **FinBERT Polarity Difference:** Transformer-based probability difference:

$$S_{\Phi\text{ιvBEPT}} = p(\text{positive}) - p(\text{negative}) \in [-1, +1]$$

Daily sentiment S̄t = (1/nt) ∑_{i=1}^{nt} Si,t aggregates article scores over day t. Given high directional concordance (78.4%, ρ = 0.71) and computational determinism, VADER serves as the scalable engine for the multi-year corpus.

*D. Exponential Market Memory Framework*

To model the fading relevance of news over time, we formulate the normalized exponential decay market memory framework:

$$N_{T+1} = \frac{\Sigma_{i=0}^{n} W_i \bar{S}_{T-i}}{\Sigma_{j=0}^{n} W_j}, W_i = e^{-\lambda i}$$

where λ > 0 is the decay constant, n is the lookback horizon, and W0 = 1. The denominator normalizes weights so that NT+1 ∈ [−1, +1]. The characteristic memory half-life t1/2 is:

$$t_{1/2} = \frac{\ln 2}{\lambda}$$

Two complementary decay channels are parameterized:

1) *Fast Decay (λ = 0.20, n = 7 days):* t1/2 ≈ 3.47 days, capturing rapid intra-week market digestion of breaking headlines.

2) *Slow Decay (λ = 0.05, n = 21 days):* t1/2 ≈ 13.86 days, modeling persistent narrative inertia and regulatory sentiment.

Fig. 2 illustrates the exponential weighting profiles alongside linear (Wi = 1 − i/n) and hyperbolic (Wi = 1/(1+λi)) baselines.

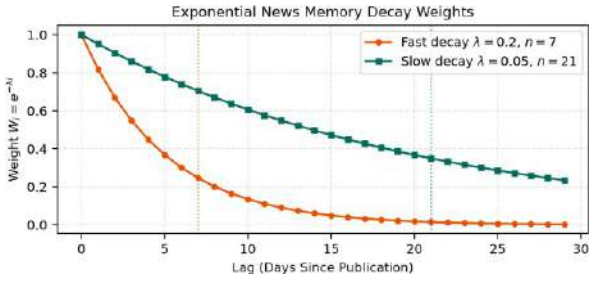



Fig. 1. EXPONENTIAL NEWS MEMORY DECAY WEIGHTS WI = E−ΛI ACROSS LAG DAYS FOR FAST (Λ = 0.20) AND SLOW (Λ = 0.05) CHANNELS COMPARED TO LINEAR AND HYPERBOLIC WEIGHTING PROFILES.

*E. Predictive Modeling and Walk-Forward Protocol*

We construct two modeling feature spaces on 3,860 aligned trading days (2014–2026):

**Path A (Deterministic Decay):** Fast bias (Nfast), slow bias (Nslow), current return (RT), prior returns (RT−1, RT−2), and 7-day rolling return volatility (σ7d).

**Path B (Learned Empirical Lags):** Seven unweighted sentiment lags (S̄T, ..., S̄T−6) replacing decay biases, allowing unconstrained coefficient estimation.

Models evaluated include Linear Regression (Path B baseline), Random Forest, and XGBoost [20]. Hyperparameters (tree depth, regularisation, learning rate) were tuned exclusively on the training split using 5-fold TimeSeriesSplit; the out-of-sample test partition (772 trading days) remained strictly sequestered.

To verify regime stability, we implement a 5-fold expanding-window walk-forward validation across sequential 643-day blocks spanning 2017 to 2026 (2017 bull run, 2018 bear, 2020–2021 bull, 2022 crypto winter, 2024–2026 ETF cycle).

*F. Friction-Adjusted Backtesting Framework*

The classifier drives a long-cash quantitative strategy: Positiont = 1 (100% BTC) when d̂t = 1, and 0 (100% cash) when d̂t = 0. Net portfolio returns deducting execution fees and slippage are modeled as:

$$\tilde{R}_t^{\nu\varepsilon\tau} = (1 + \text{Position}_t \cdot R_t)(1 - c \cdot |\Delta\text{Position}_t|) - 1$$

where c ∈ [0, 25 bps] denotes one-way transaction cost and |ΔPositiont| ∈ {0, 1} indicates position turnover.

## IV. EXPERIMENTAL RESULTS

*A. Sentiment Concordance and Distribution*

Across the 52,071 articles, VADER polarity exhibited a mean of +0.165 (median +0.153, standard deviation 0.394), reflecting journalistic reporting bias toward positive technological developments. Directional concordance between VADER and FinBERT reached 78.4% with a Spearman rank correlation of ρ = 0.71. Discordant classifications primarily arose in macroeconomic policy headlines, where FinBERT assigned neutral classifications while VADER scored localized lexical valences.

*B. Directional Classification and Significance*

Table II presents the classification performance of the Path A XGBoost model on the 772-day test set. The model achieved 51.94% overall accuracy (401 / 772 days). Performance exhibited asymmetric class recall:

**UP Class (1):** Precision of 0.5116, Recall of 0.7448, and F1-score of 0.6066.

**DOWN Class (0):** Precision of 0.5399, Recall of 0.2964, and F1-score of 0.3827.

**Aggregate Metrics:** Macro-averaged F1 of 0.4946; weighted F1 of 0.4940.

TABLE I. XGBOOST DIRECTIONAL CLASSIFICATION PERFORMANCE (772-DAY TEST SET)

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| DOWN (0) | 0.5399 | 0.2964 | 0.3827 | 388 |
| UP (1) | 0.5116 | 0.7448 | 0.6066 | 384 |
| **Overall Accuracy** | **51.94% (401 / 772)** | | | |
| **Macro-Averaged F1** | 0.4946 | | | |
| **UP-Class F1** | 0.6066 | | | |
| 95% Wilson Confidence Interval | [48.42%, 55.45%] | | | |
| Exact Binomial Test p-value (vs. 50% Null) | p = 0.1483 | | | |

To determine statistical validity, a one-sided exact binomial test yields p = 0.1483 against the 50% null hypothesis, and the 95% Wilson confidence interval spans [48.42%, 55.45%]. Because the interval crosses 50%, the nominal accuracy advantage cannot be claimed as statistically distinct from chance. This underscores that media sentiment functions as an associative filter rather than a definitive standalone predictor.

### *C. Feature Importance Analysis*

Gain-based feature importance in XGBoost assigns highest importance to the slow-decay sentiment bias (Nslow, 0.204), followed by current return (RT, 0.199) and 7-day volatility (σ7d, 0.196). Fast-decay sentiment contributes 0.187. In Linear Regression (Path B), empirical lag coefficients fluctuated erratically without monotonic decay, demonstrating that unconstrained models overfit collinear daily sentiment noise. Imposing the exponential prior regularizes information integration.

### *D. Ablation Studies on Market Memory Architecture*

TABLE II. SYSTEMATIC ABLATION ANALYSIS OF FEATURE SUBSETS, MEMORY FORMULATIONS, AND DECAY PROFILES

| Model / Feature Architecture | Accuracy | UP-F1 | Macro-F1 | Gross Return | Net Return (10 bps) | Trades |
|---|---|---|---|---|---|---|
| **Full Model (Path A: Price + Exp Decay)** | **51.94%** | **0.6066** | **0.4946** | **+16.31%** | **−11.31%** | **271** |
| Price-Only Features (No Sentiment) | 53.37% | 0.6104 | 0.5149 | +36.71% | +0.05% | 265 |
| Sentiment-Only Features (Exp Decay) | 51.94% | 0.6388 | 0.4606 | +25.10% | +11.06% | 134 |
| Raw Sentiment-Only (Today St, No Memory) | 47.54% | 0.5535 | 0.4588 | −40.19% | −57.05% | 321 |
| Price + Raw Today Sentiment (No Memory) | 52.46% | 0.5909 | 0.5118 | +17.38% | −15.54% | 283 |
| Path B (Price + 7 Raw Sentiment Lags) | 51.17% | 0.5797 | 0.4985 | +35.59% | −3.99% | 287 |
| Price + Simple Moving Average (7-Day SMA) | 51.42% | 0.5884 | 0.4980 | +5.05% | −23.35% | 279 |
| Price + Linear Decay (Wi = 1 − i/n) | 52.07% | 0.6232 | 0.4824 | −14.84% | −32.41% | 229 |
| Price + Hyperbolic Decay (Wi = 1/(1+λi)) | 51.04% | 0.5996 | 0.4848 | −13.56% | −33.96% | 241 |
| Passive Buy & Hold Benchmark | --- | --- | --- | +2.82% | +2.82% | 1 |

Crucially, raw sentiment without memory yields catastrophic failure (47.54% accuracy, −40.19% gross return, −57.05% net return). In contrast, exponential sentiment-only features produce +25.10% gross return and +11.06% net return with only 134 trades. Exponential decay also markedly outperforms linear (−14.84%) and hyperbolic (−13.56%) decay profiles, proving that news relevance fades geometrically.

### *E. Expanding-Window Walk-Forward Validation*

Table IV documents the 5-fold expanding-window walk-forward evaluation across sequential historical regimes. Model accuracy remains tightly bounded around 50%–53% across regimes. Downside capital preservation is evident in bear markets: in Fold 4 (2022 crypto winter), the strategy restricted gross losses to −28.31% while Buy & Hold collapsed −64.67%. Over the multi-regime horizon, strategy gross returns reached +114.69% versus +200.53% for buy-and-hold, with significantly lower portfolio drawdown.

TABLE III. 5-FOLD EXPANDING WALK-FORWARD CROSS-VALIDATION (2017–2026)

| Fold | Market Regime | Accuracy | UP-F1 | Strategy Return | Buy & Hold |
|---|---|---|---|---|---|
| 1 | 2017 Bull / 2018 Crash | 53.03% | 0.6558 | +152.02% | +293.74% |
| 2 | 2019 Consolidation | 50.39% | 0.5694 | +5.12% | +113.62% |
| 3 | 2020–2021 Bull Cycle | 52.10% | 0.6480 | +441.74% | +465.75% |
| 4 | 2022 Crypto Winter | 47.74% | 0.5401 | −28.31% | −64.67% |
| 5 | 2024–2026 ETF Era | 51.32% | 0.6484 | +42.87% | +194.21% |
| **Multi-Regime Average** | | **50.92%** | **0.6123** | **+114.69%** | **+200.53%** |

### *F. Friction-Adjusted Backtest Performance*

Table V and Fig. 3 present out-of-sample backtest metrics across six transaction fee tiers. In frictionless trading (c = 0 bps), the strategy generated +16.31% cumulative return versus +2.82% for Buy & Hold, with an annualized Sharpe ratio of 0.379 (vs. 0.263) and reduced maximum drawdown (−42.29% vs. −49.74%). However, the strategy executed 271 position changes over 772 days. Once realistic transaction costs are applied (c ≥ 10 bps), turnover friction eliminates strategy profitability (−11.31% net return at 10 bps).

TABLE IV. STRATEGY PERFORMANCE ACROSS TRANSACTION COST TIERS (772-DAY TEST SET)

| Fee Tier | Cumulative Net | Ann. Return | Sharpe | Max Drawdown |
|---|---|---|---|---|
| **0 bps (Frictionless)** | **+16.31%** | **+7.41%** | **0.379** | **−42.29%** |
| 5 bps (0.05%) | +1.57% | +0.74% | 0.222 | −43.75% |
| 10 bps (0.10%) | −11.31% | −5.52% | 0.066 | −45.16% |
| 15 bps (0.15%) | −22.57% | −11.40% | −0.090 | −46.55% |
| 20 bps (0.20%) | −32.39% | −16.91% | −0.245 | −47.89% |
| 25 bps (0.25%) | −40.98% | −22.08% | −0.401 | −50.87% |
| **Buy & Hold** | **+2.82%** | **+1.33%** | **0.263** | **−49.74%** |

Fig. 2. FRICTION-ADJUSTED CUMULATIVE EQUITY CURVES OF THE XGBOOST SENTIMENT STRATEGY ACROSS TRANSACTION COST TIERS (0, 5, 10 BPS) VERSUS THE BUY & HOLD BENCHMARK ON THE 772-DAY OUT-OF-SAMPLE TEST PERIOD.

## V. DISCUSSION AND LIMITATIONS

The empirical findings yield three vital insights for quantitative financial NLP:

1) *Memory Filtering Over Raw Signals:* Raw, memoryless sentiment injects excessive variance, generating negative alpha (−40.19%). Parameterized exponential decay stabilizes the feature space and dampens false breakout noise.

2) *The Turnover Friction Tax:* In high-frequency or daily rebalanced models, gross back test outperformance is easily dissipated by execution costs. Future research must incorporate turnover-penalized objective functions and multi-day holding thresholds.

3) *Methodological Limitations:* The study assumes execution at daily UTC closing prices without modeled intra-day order book impact. Furthermore, social media channels and non-English news were excluded.

## VI. CONCLUSION AND REPRODUCIBILITY

This paper presented an end-to-end, reproducible pipeline for predicting Bitcoin price direction from multi-source cryptocurrency journalism. The framework integrates 52,071 cleaned articles, dual VADER/FinBERT sentiment scoring, an exponential market memory framework, and walk-forward cross-validation. While the XGBoost classifier achieves 51.94% directional accuracy and capital protection during bear regimes, rigorous hypothesis testing indicates the edge is not statistically distinguishable from chance (p = 0.148). Furthermore, transaction cost modeling shows that trading turnover extinguishes outperformance when fees exceed 5 bps. All datasets, preprocessing scripts, and interactive parameter simulators are open-source at: https://github.com/KSDeshappriya/BitcoinAnalysis.

## ACKNOWLEDGMENT

This research was conducted as an open-source, reproducible study. All code, cleaned datasets, and simulation notebooks are publicly accessible at https://github.com/KSDeshappriya/BitcoinAnalysis.

## REFERENCES

[1] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” Decentralized Business Review, 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[2] P. Ciaian, M. Rajcaniova, and D. Kancs, “The economics of BitCoin price formation,” Applied Economics, vol. 48, no. 19, pp. 1799–1815, 2016.

[3] P. C. Tetlock, “Giving content to investor sentiment: The role of media in the stock market,” The Journal of Finance, vol. 62, no. 3, pp. 1139–1168, 2007.

[4] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, vol. 2, no. 1, pp. 1–8, 2011.

[5] O. Kraaijeveld and J. De Smedt, “The predictive power of public Twitter sentiment for forecasting cryptocurrency prices,” Journal of International Financial Markets, Institutions and Money, vol. 65, p. 101188, 2020.

[6] T. Loughran and B. McDonald, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks,” The Journal of Finance, vol. 66, no. 1, pp. 35–65, 2011.

[7] M. Barber and T. Odean, “All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors,” The Review of Financial Studies, vol. 21, no. 2, pp. 785–818, 2008.

[8] Raheman, A. Kolonin, I. Fridkins, I. Ansari, and M. Vishwas, “Social media sentiment analysis for cryptocurrency market prediction,” in Proc. IEEE Int. Conf. Cognitive Infocommunications (CogInfoCom), 2022, pp. 145–152.

[9] R. F. Engle and V. K. Ng, “Measuring and testing the impact of news on volatility,” The Journal of Finance, vol. 48, no. 5, pp. 1749–1778, 1993.

[10] Bergmeir and J. M. Benítez, “On the use of cross-validation for time series predictor evaluation,” Information Sciences, vol. 191, pp. 192–213, 2012.

[11] M. López de Prado, Advances in Financial Machine Learning, Hoboken, NJ: John Wiley & Sons, 2018.

[12] Gurgul, S. Lessmann, and W. K. Härdle, “Deep learning and NLP in cryptocurrency forecasting: Integrating financial, blockchain, and social media data,” International Journal of Forecasting, vol. 40, no. 3, pp. 1102–1118, 2024.

[13] J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in Proc. 8th Int. Conf. Weblogs and Social Media (ICWSM-14), 2014, pp. 216–225.

[14] Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” Financial Data Science Review, vol. 1, no. 1, pp. 1–15, 2019.

[15] H. Yang, X. Y. Liu, and C. D. Wang, “FinGPT: Open-source financial large language models,” in Proc. FinLLM Symposium at IJCAI, 2023, pp. 1–8.

[16] M. F. B. Hossain, M. A. Kabir, and M. M. Rahman, “FinBERT-BiLSTM: A deep learning approach for cryptocurrency sentiment analysis,” IEEE Access, vol. 12, pp. 45120–45134, 2024.

[17] S. M. Dashtaki, H. Zare, and M. Mosleh, “A multisource fusion framework for cryptocurrency prediction,” Applied Soft Computing, vol. 166, p. 112185, 2024.

[18] H. Moradi-Kamali, S. Khalilpourazari, A. Mirzazadeh, and H. Hashemi, “Market-derived financial sentiment analysis for cryptocurrency price prediction,” Technological Forecasting and Social Change, vol. 201, p. 123254, 2025.

[19] S. R. Binkowski and D. F. Silva, “Machine learning with market memory for high-frequency crypto trading,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7890–7901, 2024.

[20] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

# A Web Based Decision Support Platform for Assessing Digital Readiness among Sri Lankan SMEs

Uvini Kawya
*Department of Information and System Sciences*
*National School of Business Management*
*Homagama, Sri Lanka*
paukawya@students.nsbm.ac.lk

Thisarani Wickramasinghe
*Department of Software Engineering and Computer Security*
*National School of Business Management*
*Homagama, Sri Lanka*
thisarani.w@nsbm.ac.lk

***Abstract*—Small and medium enterprises (SMEs) in Sri Lanka face challenges in adopting digital technologies due to limitations in financial resources, digital skills, infrastructure, and strategic planning. This study investigates digital readiness, adoption barriers, and perceived business performance among Sri Lankan SMEs and presents a web-based decision support platform to support readiness assessment. A quantitative survey collected 102 responses, of which 71 respondents reported using digital tools. Guided by Design Science Research (DSR), the SME Digi platform was developed to integrate readiness assessment, Composite Mean Index (CMI) scoring, dashboard visualization, rule-based recommendations, and basic inventory and sales management functions. The findings indicated moderate digital readiness across five dimensions, with mean scores ranging from 3.00 to 3.07. Limited financial support and the absence of a clear digital strategy were identified as prominent adoption barriers. Reliability analysis produced Cronbach's alpha values of 0.748, 0.859, and 0.833 for digital readiness, digital barriers, and perceived business performance, respectively. System evaluation showed that eight of ten functional tests passed, while preliminary user acceptance testing with five SME users indicated positive feedback. The study contributes a practical, integrated decision support platform tailored to the Sri Lankan SME context. However, further validation is required to establish its broader usability and effectiveness.**

***Keywords—digital readiness, SMEs, decision support, digital barriers, design science research***

## I. INTRODUCTION

Digital transformation has moved beyond a single tool to become a more comprehensive organizational ability, affecting how companies do their work, how they engage customers and how they utilize data and adjust to market changes. The transition to the digital world is especially difficult for SMEs, as they don't have the resources to invest in the appropriate digital tools and don't have the technical skills and employee capacity to support a digital transition. Vial considers digital transformation as requiring a combination of information, computing, communication and connectivity technologies, which seismically alters the way the organization works [1]. Recent SME research also reveals a set of interrelated capabilities that are essential for digital transformation to take place, including digital awareness, digital strategy, digital implementation, digital learning and digital organizational flexibility [2].

Organizational and environmental factors in Sri Lanka also affect the adoption of technology among SMEs; not only the availability of technologies. Both Yazeer and Sachithra found that the adoption of technology in SMEs in Sri Lanka had many organizational and environmental effects; while opportunities were identified in e-commerce, online transactions and business communication via social media [3]. Based on the findings of this research, a viable SME support mechanism should be more than just the technology checklist and should provide a framework for understanding readiness, barriers and practical improvement priorities.

There are various theoretical explanations for technology adoption at the organizational level. The adoption of a technological, organizational and environmental perspective is captured by the Technology Organization Environment (TOE) framework [4] and Thong's [5] small-business information systems adoption model takes into account the unique features of small businesses. Adoption behavior can also be predicted with the theory of Diffusion of Innovations, by extrapolating the perceived characteristics of innovations [6]. The theoretical models are not yet providing SME owners with an environment where they can evaluate themselves and then turn their answer into a score and actionable output for management. The concepts of digital readiness have been demonstrated to be integrated into dashboards in previous research for decision making purposes.

The idea of an intelligent dashboard decision support approach that enables the measurement of SME digital readiness has been developed by Okfalisa et al. [7] and they have underlined the importance of having measurable criteria, as well as visual performance mapping. However, there is a need for an intuitive, local and integrated application that links readiness assessment with continuous SME support. That leads to the development of an artefact in the form of a web based artefact for the benefit of SMEs in Sri Lanka “SME Digi”.

The research question is: How can a web-based decision-support platform operationalize assessments of digital readiness, adoption barriers and business performance for Sri Lankan SMEs? The objectives are to identify assessment

factors, report survey findings, develop SME Digi and evaluate its selected functional workflows. Causal effects of platform use are outside the scope.

## II. LITERATURE REVIEW

### *A. SME Digital Transformation and Readiness*

Digital transformation is a multi-dimensional process for SMEs. Digital transformation requires technological infrastructure, financial ability, skills of the workforce, managerial practices, and readiness with regards to cybersecurity. The SME transformation framework of Kahveci is based on the interdependent relationship between awareness, strategy, adoption/implementation and continuous improvement [2]. This perspective is the basis for the additional dimensions of readiness in SME Digi as well as the advantages of using a dashboard that emphasizes those with relatively poor readiness as opposed to one that provides a single readiness score. An organization's readiness for digital can be measured by how well it embraces and effectively applies digital technologies. In SMEs, measurement of readiness is helpful because there are only limited resources to adopt decisions and they need to be prioritized. Okfalisa et al. show how to transform SME readiness indicators to a dashboard for decision support to aid interpretation [7]. The approach applied in the present study is a simplified and interpretable scoring scale which allows the translation of the answers to the questionnaires into a visual decision support system and therefore makes digital readiness assessment accessible to the owners and managers of SMEs without advanced technical knowledge.

### *B. Technology Adoption Foundations*

The TOE framework is relevant in this regard because it considers the characteristics of the technology, the conditions of the organizations and the external environment to be important factors in shaping the adoption of the organization [4]. A synthesis of the TOE framework and its use in information systems research is given by Baker [8]. Thong's model is complementary since it concentrates particularly on small business and considers the nature of the decision maker, organization, technology and environment [5]. Together they provide a defensible theoretical basis to ponder the context of assessment areas and questions. This perspective on adoption is added by Rogers' Diffusion of Innovations theory, which is represented by the five factors of perceived relative advantage, compatibility, complexity, trialability and observability [6]. The DOI is not used as a separate computational model in SME Digi, but is used as an explanation of why some digital technologies might seem easier or more beneficial for SMEs to use, than others. This separation is crucial since the system is not a statistical evaluation of one theory of adoption, but rather a decision support artefact. Research at the firm level also underscores the roles of contextual factors [9] and evidence from emerging economies also reveals other factors affecting SME digital transformation [10].

### *C. Decision Support and Research Gap*

The decision support system (DSS) is a tool that assists managers in understanding information and making decisions. Rule based DSS methods give the recommendations in a transparent way by using the pre-defined rules as Chhor et al. [11] have shown. There are important factors that affect the adoption of technology that can be explained using the Technology Organization Environment (TOE) framework [4, 8], Thong's small-business adoption model [5] and Diffusion of Innovations (DOI) theory [6]. But these are mostly theoretical rather than cohesive self-evaluation instruments for SMEs.

Okfalisa et al. [7] designed an intelligent dashboard for the digital readiness assessment of SMEs through Fuzzy AHP and OMAX methods. In a like manner, Faturrahman and Novianty [12] have proposed a web-based system for evaluating the readiness of MSMEs that include parameters related to technology, manpower, business process and organizational strategy. These studies illustrate various methods to assessing readiness and decision support. However, a practical platform catered to the Sri Lankan SMEs that would combine readiness assessment, identification of adoption barriers, visualizations that would be understood by the end users and basic operational support is still required. To fill these gaps, SME Digi provides assessments by questionnaire, a Composite Mean Index (CMI) scores, recommendation by rules, dashboard visualization, inventory management and sales functions. The value of its contribution is that it brings these features to a localized accessible web-based platform rather than a new adoption theory or greater predictive power.

TABLE I
COMPARISON OF EXISTING LITERATURE AND ITS RELEVANCE TO SME DIGI

| Literature basis | Key contribution | Use in SME Digi |
|---|---|---|
| TOE [4], [8] | Technology, organisation, environment | Assessment factor selection |
| Thong [5] | Small-business IS adoption | SME-specific contextualisation |
| DOI [6] | Innovation adoption characteristics | Interpretation of adoption behaviour |
| Okfalisa et al. [7] | Digital-readiness dashboard DSS | Assessment and visualisation concept |
| Chhor et al. [9] | Rule-based decision support | Transparent recommendation direction |
| Faturrahman and Novianty [20] | Web-based MSME readiness evaluation | Comparison of assessment scope |

Table I summarizes the related literature. The comparison highlights that SME Digi combines concepts from existing adoption frameworks and decision-support approaches into a single platform intended for Sri Lankan SMEs.

## III. METHODOLOGY

### *A. Research Design*

This study combined a quantitative survey with Design Science Research (DSR) to examine digital technology adoption among Sri Lankan SMEs and develop SME Digi, a web-based decision support platform. The quantitative approach was used to collect measurable information about Digital Readiness,

Digital Barriers, and Business Performance. DSR was selected because it supports the systematic design, development, and evaluation of an artefact addressing a practical problem [13]. The quantitative research design is conducted based on a methodologically correct approach [14]. Although mixed-methods research is discussed in the methodological literature [15], this study used quantitative survey data and prototype evaluation.

The research was conducted in accordance with the following main stages of DSR that were proposed by Peffers et al. [13]: The problem was identified by conducting literature review on Digital Transformation challenges of SMEs. Secondly, the research goals and the system requirements were formulated. Third, SME Digi was designed and developed to offer digital assessments as well as readiness scoring, suggestions and business management. Lastly, system testing was done and initial user acceptance testing was conducted to demonstrate and evaluate the developed platform.

The assessment framework drew on the Technology–Organization–Environment (TOE) framework [4], [8], Thong's small-business information systems adoption model [5], and Rogers' Diffusion of Innovations theory [6]. These frameworks enabled to pinpoint technological, organization, financial, workforce and environmental factors in the context of digital technology adoption. The study analyzed these concepts and proposed a viable platform to help with digital transformation decisions for Sri Lankan SMEs.

### *B. Requirements and Data Collection*

In 2026, quantitative data was gathered from the SMEs in Sri Lanka using a structured questionnaire. The data collection technique used was non-probability convenience sampling, and the Google Forms questionnaire was sent via social media, personal contacts, and direct communication with SMEs. 102 responses were received from the owners and employees of the SMEs.

The respondents represented manufacturing (29.4%), services (24.5%), industry (19.6%), agriculture (11.8%), retail (10.8%), and other sectors (3.9%). The questionnaire examined Digital Readiness, Digital Barriers, and Business Performance using a five-point Likert scale.

The Composite Mean Index (CMI) was used to calculate the mean score for each assessment dimension:

$$CMI = \frac{\sum X_i}{n} \quad (1)$$

where n is the number of valid responses. CMI is the arithmetic mean of valid equally weighted responses; missing responses are excluded rather than imputed. Scores were classified Low (1.00–2.60), Moderate (2.61–3.40) and High (3.41–5.00). For Digital Readiness, higher scores indicate greater readiness. For Digital Barriers, higher scores indicate more significant adoption barriers. Business Performance scores represent respondents' perceived performance outcomes.

### *C. Data Analysis and Reliability Testing*

The collected questionnaire responses were analyzed using IBM SPSS Statistics. Descriptive statistics, including frequencies, percentages, means, and standard deviations, were used to examine Digital Readiness, Digital Barriers, and perceived Business Performance among participating SMEs.

Cronbach's alpha was used to assess internal consistency. The coefficients were 0.748 for Digital Readiness, 0.859 for Digital Barriers, and 0.833 for Business Performance. The Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity were also examined to assess the suitability of the data for factor analysis.

Fig. 1 illustrates the assessment interface. KMO values were 0.477 for Digital Readiness, 0.698 for Digital Barriers, and 0.644 for Business Performance. Although Bartlett's tests were statistically significant ($p < 0.001$), the readiness KMO was below the commonly accepted minimum of 0.50. Therefore, the results were interpreted descriptively without claiming that complete construct validity had been established.

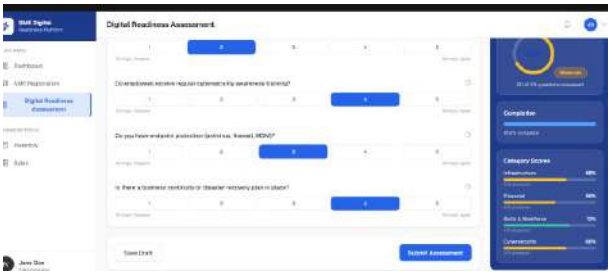


Fig. 1. Digital-readiness assessment interface with completion and category scores.

### *D. System Design and Development*

The development of SME Digi has been done in accordance with Design Science Research approach [13] which is a web based system that supports decision making. It is built with React.js and Vite on the front-end, PHP REST APIs for backend and MySQL/MariaDB for data storage. This encompasses SME registration, digital assessments, CMI scoring, rule-based recommendations, dashboards, inventory management and sales recording. There is also an administrator interface to approve/reject the SME registrations, to manage SME accounts, to view assessment results/recommendations, to generate reports, and to manage announcements. Administrator Dashboard provides summary of the registered SMEs and assessment. Predefined rules are used to propose recommendations, and assessment scores are classified as Low, Moderate or High. This method was found to be supported by the previous digital readiness assessment and rule-based decision support studies [7] and [11]. System functions have been developed and iteratively refined.

### *E. System Evaluation and User Acceptance Testing*

The evaluation of SME Digi was done by functional and non-functional testing within the DSR evaluation process [13]. Ten test cases were explored including assessment processing, CMI calculation, classification, recommendations, dashboard functions, responsiveness and business operations. Eight tests passed and two tests showed that there was a problem that needs

to be resolved. There were five SME users who participated in User Acceptance testing (UAT). The usability of the website was rated as easy by all five and the assessment questions were rated as clear by participants as well as the dashboard was rated as useful by participants. The results suggest that the initial acceptance was positive but more testing with more SME users is needed.

## IV. RESULTS AND DISCUSSION

### *A. Empirical Study Scope*

The empirical and prototype evaluations are reported separately to avoid treating illustrative interface values as survey observations.

### *B. Quantitative Findings*

A total of 102 responses were collected from Sri Lankan SMEs, of which 71 respondents (69.6%) reported using digital tools. The survey examined five digital readiness dimensions, barriers to digital adoption, and perceived business performance. The Composite Mean Index (CMI) was interpreted using three levels: Low (1.00–2.60), Moderate (2.61–3.40), and High (3.41–5.00).

TABLE II
DIGITAL READINESS CMI RESULTS

| Dimension | CMI | Interpretation |
|---|---|---|
| Infrastructure Readiness | 3.03 | Moderate |
| Financial Readiness | 3.07 | Moderate |
| Digital Skills | 3.03 | Moderate |
| Strategic Readiness | 3.00 | Moderate |
| Cybersecurity | 3.06 | Moderate |

The results in Table II indicate that all five digital readiness dimensions were classified as Moderate. Financial Readiness recorded the highest mean (3.07), followed by Cybersecurity Readiness (3.06). Infrastructure Readiness and Digital Skills and Workforce Readiness both recorded means of 3.03, while Strategic Readiness had the lowest mean (3.00). These results suggest that participating SMEs demonstrated moderate readiness across the assessed dimensions, without a single dimension achieving a high classification.

TABLE III
DIGITAL ADOPTION BARRIER RESULTS

| Digital Barrier | CMI |
|---|---|
| High Cost of Digital Tools | 3.44 |
| Lack of Digital Skills among Employees | 3.51 |
| Poor Internet Connectivity | 3.52 |
| Lack of Awareness about Digital Tools | 3.47 |
| Cybersecurity Concerns | 3.52 |
| Resistance to Change from Staff | 3.46 |
| Lack of Government Support | 3.46 |
| Lack of Clear Digital Strategy | 3.54 |
| Difficulty Integrating Systems | 3.44 |
| Limited Access to Financial Support | 3.57 |

Table III presents the barrier results. The digital adoption barrier assessment identified Limited Financial Support (3.57) and Lack of a Clear Digital Strategy (3.54) among the more significant challenges. Other assessed barriers recorded means ranging from 3.44 to 3.52. These findings suggest that financial constraints and strategic planning difficulties remain important obstacles to digital transformation among participating SMEs.

Perceived business performance measures recorded mean values ranging from 3.56 to 3.66. These scores represent respondents' perceptions rather than independently verified improvements in business performance. Consequently, the results do not establish a causal relationship between digital readiness and business outcomes.

### *C. Implemented Artefact*

Fig. 2 illustrates the dashboard. The implemented SME Digi prototype includes registration, administrator approval, a business specific set of questions, Digital Readiness, Digital Barriers and Business Performance assessments, CMI scoring, rule based recommendations, dashboard reporting and assessment history. There are administrator functions separated: SME management, assessment monitoring, reports, announcements and recommendation monitoring. These are applied functions and not indicators of improved business performance in SMEs. Assessment history and dashboard visualizations are descriptive scores and classifications, not improvements in business performance.

Fig. 2. SME Digi dashboard integrating readiness and business-oriented indicators

### *D. Operational Support Functions*

Fig. 3 illustrates the inventory interface. SME Digi offers the inventory and sales management functions, which enable SME Users to manage products, monitor stocks levels, record sales transactions and view sales history. These functions serve not only for digital readiness assessments, but also for everyday business operations. The administrator interface is compatible for SMEs' registration approval/rejection, account management, assessment and recommendation monitoring, announcement management, and report generation.

A summary of registered SMEs and assessment results are shown in the admin dashboard. These features combined bring SME Digi to life beyond a digital readiness assessment and incorporate decision support, basic business operations and administrative management into one platform.

Fig. 3. Inventory management interface with low stock monitoring.

*E. Discussion*

Results show that the Sri Lankan SMEs found to be participating in the study had an average score of 3.00 and 3.07 for all five dimensions suggesting a moderate level of digital readiness. It indicates that digital transformation readiness is more than technology, as it's a combination of multiple capabilities. The findings confirm the multidimensional nature of the digital transformation that SMEs are undergoing, as raised by Kahveci [2] and the technology adoption challenges faced by Sri Lankan SMEs as identified by Yazeer and Sachithra [3].

The mean for Financial Readiness was highest at 3.07, and the lowest at 3.00 for Strategic Readiness. All five dimensions were still in the Moderate range, however. These results should not be considered as showing significant differences between dimensions unless further statistical tests are conducted.

Limited Financial Support (3.57) and Lack of a Clear Digital Strategy (3.54) were highlighted as key issues within the barrier assessment. This demonstrates that there is a need to integrate digital readiness assessment with tangible actions and recommendations that enable SMEs to pinpoint areas for change. Theoretically, the TOE framework [4] and Thong's small-business adoption model [5] can be used to explain these challenges in terms of organizational, technological and environmental factors. Existing research in Sri Lanka also explored the relationship between technology use and organization performance [16] that this study doesn't test.

SME Digi meets these needs by means of assessment based on the CMI, classifications of readiness, visualization in the dashboard and predefined rules for recommendations. SME Digi focuses on the combination of assessment, administrative management and basic business operations within the context of an SME in Sri Lanka, compared to the digital readiness dashboard developed by Okfalisa et al. [7] and the web-based readiness evaluation system proposed by Faturrahman and Novianty [12]. There are also several studies that have explored readiness factors from the various decision-support perspectives [17] and systematic reviews that highlight the interrelated conditions for SME transformation [18].

Table IV summarizes the implementation status. The contribution of the system, however, is mainly the practical integration and not necessarily a new scoring algorithm. The effectiveness of the recommendation rules in improving SME decisions and business performance has not yet been validated independently. More evaluation of the platform is therefore required to determine the utility of the platform in the long term. Previous SME cloud adoption research also highlights the significance of organizational and environmental barriers [19].

TABLE IV
IMPLEMENTATION STATUS AND RESEARCH ALIGNMENT

| Function | Current status | Research relevance |
|---|---|---|
| SME registration | Implemented | Captures SME context |
| Readiness assessment | Implemented | Core research artefact |
| Dimension/overall scoring | Implemented | Transforms responses into decision information |
| Dashboard visualization | Implemented | Supports interpretation |
| Inventory & low-stock monitoring | Implemented | Lightweight operational support |
| Sales recording/history | Implemented | Lightweight operational support |
| Dedicated barrier assessment | Implemented | Assessment and decision support |
| Business-performance assessment | Implemented | Assessment and decision support |
| Rule-based recommendations | Implemented | Assessment and decision support |
| Administrator management | Implemented | SME approval and account management |
| Assessment monitoring | Implemented | Administrative oversight |
| Reports and announcements | Implemented | System administration |

*F. Pilot User Acceptance Testing*

Fig. 4 illustrates the sales interface. Five SME users participated in a pilot UAT of the SME Digi website. 100% of respondents said that the site was easy to use. The participants found the digital readiness questions simple and clear while the dashboard was understandable and useful. The feedback is preliminary evidence of perceived usability but is not a representative sample of a population or a quantitative assessment of task efficiency.

The inventory interface displays product inventory, stock status and low inventory indicators. Sales workflow keeps track of sales and enables inventory-related processes. These modules are not complete enterprise resource planning or accounting package but are for carrying out normal SME functions.

The Today Sales feature was not working as intended in the local functional test, and the mobile responsiveness needed to be enhanced. Both issues were recorded as failed test outcomes.

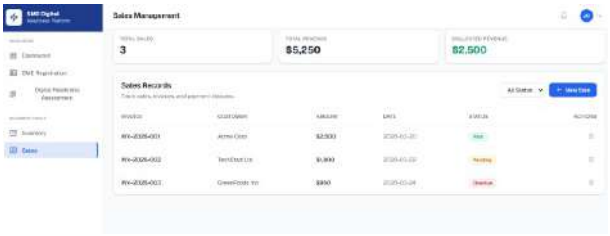



Fig. 4. Sales interface with illustrative transaction values.

*G. Validity, Scope, and Limitations*

The results of this study are not representative of the wider Sri Lankan SME population as the sample was a convenience sample of 102 respondents, and the data were self reported responses to the questionnaire. The resulting reliability coefficients were satisfactory, but the KMO value of 0.477 means that the set of items is not very factorable, and additional

construct validation is needed. There is only preliminary evidence for the examination of the structure of the three constructs, as indicated by significant Bartlett tests for all three constructs and KMO values greater than 0.60 for barriers and perceived performance.

Five user UAT feedback was positive, although it was only a small pilot without a standardized usability scoring or independent performance outcomes. Ten functional tests were performed, with two failing mobile responsiveness and Today Sales. Equal weighting and fixed thresholds are interpretable but need to be checked for validity across SME types.

Further work should correct the mobile responsiveness and Today Sales defects, document sampling eligibility and sector composition, recruit a larger sample, investigate recommendation quality, and examine longer-term changes in SME digital practices. The restrictions are in line with the evaluation principles of Design Science Research [13] and indicate the need for further empirical validation of the developed artefact.

TABLE V
EVALUATION EVIDENCE AND LIMITATIONS

| Evaluation area | Observed evidence | Interpretation |
|---|---|---|
| Functional checks | 8 of 10 passed; mobile and Today Sales failed | Prototype defects remain |
| Pilot acceptance | Five SME users; all reported easy use | Positive initial feedback |
| Questionnaire reliability | Alpha: 0.748 / 0.859 / 0.833 | Internal consistency |
| Preliminary factorability | KMO: 0.477 / 0.698 / 0.644 | Readiness KMO below 0.50 |

TABLE V
SUMMARIZES THE EVALUATION EVIDENCE. RESEARCH ON SME DIGITAL ORIENTATION FURTHER SUGGESTS THAT DIGITAL READINESS CHANGES OVER TIME [20].

## V. CONCLUSION AND FUTURE WORK

The study aimed to create a web-based decision support system called SME Digi which aims to measure the digital readiness and facilitate Sri Lankan SME digital transformation. The platform is built based on the Design Science Research methodology [13] and includes a digital readiness assessment, CMI scoring, visualization of the dashboard, recommendations based on rules, inventory and sales management, and administration functions. The system offers a practical way of converting digital readiness information into understandable outcomes, and recommendations, for owners and managers of SMEs.

A total of 102 questionnaires were received to explore digital readiness, digital adoption barriers and perceived business performance. The internal consistency of the three constructs was acceptable according to the reliability analysis. Ten functional and non-functional tests were performed on the system, eight passed and two failed which needed improvement. Initial feedback from Pilot UAT by 5 SME users showed positive initial feedback on the usability, assessment clarity, and the usefulness of the dashboards. But the findings are limited in generalizability due to the use of the convenience sampling method, the small number of UAT participants, and the limited validity of the questionnaires. Future work will address increasing the mobile responsiveness and Today Sales function, adding more user testing, and enhancing the validation of assessment questions and recommendation rules. Further development will feature a better assessment-history analysis, new levels of detail in reporting, better validation of recommendations and longitudinal analysis of changes to SME digital readiness.

These developments will facilitate the ongoing improvement of SME Digi as a practical digital transformation decision support platform that is aligned with research which underlines the dynamic nature of SME digital orientation [20].

## REFERENCES

[1] G. Vial, “Understanding digital transformation: A review and a research agenda,” J. Strategic Inf. Syst., vol. 28, no. 2, pp. 118–144, 2019, doi: 10.1016/j.jsis.2019.01.003.

[2] E. Kahveci, “Digital transformation in SMEs: Enablers, interconnections, and a framework for sustainable competitive advantage,” Administrative Sciences, vol. 15, no. 3, Art. no. 107, 2025, doi: 10.3390/admsci15030107.

[3] A. R. M. Yazeer and V. Sachithra, “Exploring constraints and catalysts: A comprehensive analysis of technology adoption in Sri Lankan small and medium enterprises,” Asian J. Res. Comput. Sci., vol. 17, no. 1, pp. 11–30, 2024, doi: 10.9734/ajrcos/2024/v17i1409.

[4] L. G. Tornatzky and M. Fleischer, The Processes of Technological Innovation. Lexington, MA, USA: Lexington Books, 1990.

[5] J. Y. L. Thong, “An integrated model of information systems adoption in small businesses,” J. Manage. Inf. Syst., vol. 15, no. 4, pp. 187–214, 1999.

[6] E. M. Rogers, Diffusion of Innovations, 5th ed. New York, NY, USA: Free Press, 2003.

[7] Okfalisa, Mahyarni, W. Anggraini, Saktioto, and B. Pranggono, “Assessing digital readiness of small medium enterprises: Intelligent dashboard decision support system,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, 2022.

[8] J. Baker, “The technology–organization–environment framework,” in Information Systems Theory, Y. K. Dwivedi, M. R. Wade, and S. L. Schneberger, Eds. New York, NY, USA: Springer, 2012, pp. 231–245.

[9] T. Oliveira and M. F. Martins, “Literature review of information technology adoption models at firm level,” Electron. J. Inf. Syst. Eval., vol. 14, no. 1, pp. 110–121, 2011.

[10] V. A. Ta and C.-Y. Lin, “Exploring the determinants of digital transformation adoption for SMEs in an emerging economy,” Sustainability, vol. 15, no. 9, Art. no. 7093, 2023, doi: 10.3390/su15097093.

[11] J. Chhor, V. Fischer, F. Kröppel, and R. H. Schmitt, “Rule-based decision support for no-code digitalized processes,” Procedia CIRP, vol. 107, pp. 258–263, 2022, doi: 10.1016/j.procir.2022.04.042.

[12] B. Faturrahman and I. Novianty, “Development of a web-based MSME digital transformation readiness evaluation system using a design science research approach,” Journal of Artificial Intelligence and Engineering Applications, vol. 5, no. 2, pp. 3123–3128, 2026, doi: 10.59934/jaiea.v5i2.2126.

[13] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” J. Manage. Inf. Syst., vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.

[14] J. W. Creswell and J. D. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th ed. Thousand Oaks, CA, USA: SAGE, 2018.

[15] J. W. Creswell and V. L. Plano Clark, Designing and Conducting Mixed Methods Research, 3rd ed. Thousand Oaks, CA, USA: SAGE, 2018.

[16] W. J. A. J. M. Lasanthika and H. D. Pavithra, “Technological adoption and organizational performance of SMEs in Sri Lanka,” Sri Lankan J. Entrepreneurship, vol. 4, no. 1, 2025, doi: 10.31357/sje.v4i01.8640.

[17] Okfalisa, W. Anggraini, G. Nawanir, Saktioto, and K. Y. Wong, "Measuring the effects of different factors influencing on the readiness of SMEs towards digitalization: A multiple perspectives design of decision support system," Decision Sci. Lett., vol. 10, no. 3, pp. 425–442, 2021, doi: 10.5267/j.dsl.2021.1.002.

[18] G. H. Sagala and D. Őri, "Toward SMEs digital transformation success: A systematic literature review," Inf. Syst. E-Bus. Manage., vol. 22, pp. 667–719, 2024, doi: 10.1007/s10257-024-00682-2.

[19] A. Alshamaila, S. Papagiannidis, and F. Li, "Cloud computing adoption by SMEs in the North East of England," J. Enterprise Inf. Manage., vol. 26, no. 3, pp. 250–275, 2013, doi: 10.1108/17410391311325225.

[20] K. E. Barragan and F. S. R. Becker, "Keeping pace with the digital transformation—exploring the digital orientation of SMEs," Small Bus. Econ., vol. 64, pp. 1361–1385, 2025, doi: 10.1007/s11187-024-00947-7.

# Assessing Enterprise ERP Cloud Migration Strategies in Emerging markets: A Critical Success Factor Model

M.B.S.Sasanka
*Department of marketing and tourism Management*,
*NSBM Green University,*
*Homagama, Sri Lanka*
mbssasanka@students.nsbm.ac.lk

***Abstract*—Enterprise Resource Planning (ERP) cloud migration can improve scalability ,faster innovation , and but enterprises in emerging markets face uneven connectivity, limited cloud skills , regulatory uncertainty , data residency requirements ,foreign-exchange exposure , and vendor dependence. This conceptual model development study integrates the technology organization environment framework, The DeLone and McLean information systems success model ,and ERP /cloud ERP Critical Success Factor(CSF) model that connects migration strategy choices to implementation success. A structured synthesis of cloud ERP and digital development literature is integrated with The technology organization environment framework and information systems success theory . The resulting model proposes six higher order CSFs strategic alignment and executive governance; And emerging market institutional readiness . A sequential mixed method design is specified for validation through expert interviews followed by a cross-sectional survey of ERP decision makers in emerging market enterprises .The model predicts that CSFs influence migration success directly and through migration strategy fit, while institutional constraints moderate selected relationships . The paper contributes by moving beyond generic implementation checklists towards a parsimonious ,testable model that treats local infrastructure ,regulation ,skill , and macroeconomic volatility as first class migration conditions. It also offers a stage gate assessment that managers can use to compare re-implementation, selective transformation, and phased hybrid migration strategies before committing resources.**

***Keywords—Cloud ERP, critical success factors , emerging markets ,migration strategy ,TOE framework.***

## I. Introduction

Cloud enterprise resource planning (ERP) relocates core integrated business applications from organization controlled infrastructure to vendor managed or hosted cloud environments .The shift can reduce infrastructure ownership, improve remote access , and accelerate updates, but migration is not a simple hosting decision. It simultaneously changes operating processes ,data, responsibilities, integration boundaries, controls, skills, and vendor relationships. Prior reviews identify security ,project management, communication , organizational fit data quality, and top management support as recurring critical success factors (CSFs)[1]-[3].

The Challenge is sharper in emerging markets. The world-bank reports an uneven digitalization landscape and notes that cloud and data infrastructure markets remain concentrated in high income economies [4],[5]. Lower income settings can face less reliable border data rules, and weaker local service ecosystems. These conditions can alter both the importance of a CSF and the feasibility of a migration strategy [9] . A technically sound global template may therefore underperform when copied without contextual adaptation

Existing cloud ERP studies largely aggregate factors across settings or focus on adoption and implementation generally[11] . they provide limited explanation of how CSFs should shape the choice among migration strategies ,or how emerging market institutional conditions change factor outcome relationships . this paper addresses that gap through the research question; which critical success factors and contextual conditions determine successful enterprise ERP cloud migration strategies in emerging markets?

The objectives are to ;(1) synthesizes technology, organizational, environmental, and individual CSFs;(2) define migration success as a multidimensional outcome;(3) Model migration strategy fit as a mechanism ; and (4)specify an empirical validation design suitable for a six page conference study . the contribution is a testable CSF model and a practical readiness scorecard aligned with track 4: digital business , management and Innovation systems

## II. LITERATURE REVIEW AND THEORETICAL BASIS

### *A. Cloud ERP Migration and Strategy*

Cloud ERP migration concerns the transition of ERP processes , Configuration, Integration ,Controls and Data to a cloud delivery model . Common Strategic Patterns include full Re-implementation on a standardized software as a service platform , selective transformation of priority modules and phased hybrid migration that temporarily retains selected on premise components . Strategy choice reflects legacy tolerance , data residency ,ecosystem maturity , and organizational capacity . Recent industry evidence also shows that many organizations continue to evaluate , implement ,or run parallel landscapes, reinforcing the need to distinguish deployment from completed transformation [6].

*B. Critical success factors*

A systematic review identify of cloud ERP implementation identified 35 CSFs and reported security ,project management ,and communication as the most frequently cited [1],.Another review of 81 studies identified 32 factors spanning organizational , environmental ,technological ,and individual characteristics [2] .A recent SME focused review highlights system compatibility ,data security ,and senior managerial use ,overlapping factors should be grouped into higher order constructs and connected to strategy selection and outcome measures[12].

*C. Integrating TOE and IS success*

The Technology Organization Environment (TOE) Framework explains enterprise innovation through technological conditions , organizational characteristics ,and the external environment [7] .The DeLone and McLean information systems success model links system , information ,and service quality to use ,satisfaction and net benefits [8] .Combining them separates explanatory conditions form success outcomes. People capability is retained as an explicit dimension because Cloud ERP migration changes roles ,routines , and knowledge .Migration strategy fit is added as the alignment mechanism between conditions and execution approach .

TABLE 1
PROPOSED CSF CONSTRUCTS

| Construct | Dimensions | Proposed effect |
|---|---|---|
| Strategic alignment and executive govemance | Business Case ; Sponsorship;Benefits Ownership ; Decision Rights | Prioritization and accountability |
| Process and data readiness | Standization; legacy complexity ;data quality;rehearsal | Lower redesign and cutover risk |
| Architecture , Integration, and security readiness | Target architecture; interfaces; identity ; cyber controls; resilience | Interoperability,Quality, Continuity |
| People Capability and Change readiness | Skills;particitpation;traning ;leadership;knowledge transfer | Adoption capability |
| Vendor and ecosystem govemance | Provider/parther capability; SLA ; contracts ;portability; exit | Lower service and dependency risk |
| Institutional readiness | Connectivity;regulatory clarity; local services; ecosystem matunity | Feasible and compliant migration |

## III. PROPOSED CRITICAL SUCCESS FACTOR MODEL

The proposed model contains six higher order CSFs . table 1 Operationalizes each construct and identifies its expected migration effect. The model treats success as the achievement of schedule and budget control ,system and data quality , User adoption , operational continuity , compliance , and expected business benefits .

*A. Hypotheses*

Strategic alignment converts cloud migration from a technology replacement into an accountable transformation program . Executive sponsorship ,clear benefits, and governance should therefore improve resource continuity and decision speed[1],[3] .

*H1 : Strategic alignment and executive governance positively ERP cloud migration success.*

Legacy process variation and poor master data amplify configuration, testing ,and cut-over complexity . Readiness should reduce defects and post go live disruption [1],[12].

*H2: Process and data readiness positively influence ERP cloud migration success .*

ERP platforms sit at the center of enterprise integration and control . Target architecture, interface rationalization identity management, cybersecurity and controls, backup, and disaster recovery should strengthen quality and continuity [12],[14] .

*H3: Architecture, integration, and security readiness positively influence ERP cloud migration success.*

Migration redistributes work between internal teams , vendors , and business users, skills, participation, training, and change leadership should improve acceptance positively influence ERP cloud migration success [13].

*H4: People capability and change readiness positively influences ERP cloud migration success.*

Cloud ERP increases reliance on software providers , implementation partner ,connectivity providers ,and contract enforcement[3],[11].

*H5 : Vendor and ecosystem governance positively influence ERP cloud migration success.*

In emerging markets unreliable infrastructure ,ambiguous regulation, data localization requirements and exchange rate volatility can constrain otherwise strong programs [5],[6].

*H6: Institutional readiness positively influences ERP cloud migration success.*

The Strategy fit is the alignment of the selected migration approach with legacy complexity , process standardization, regulation, connectivity ,capability ,and Risk Tolerance.

*H7 : Migration strategy fit mediates the relationships between the six CSFs and migration success.*

Infrastructure instability, regulatory ambiguity , localization difficulty , and currency volatility can reduce the benefits of otherwise strong internal capability [5],[9].

*H8 : Institutional constraints weaken the effects of architecture readiness and vendor governance on migration success.*

Contextual conditions → **CSF readiness → Migration strategy fit → Migration success**

| Antecedents | Mechanism | Outcome |
|---|---|---|
| Six CSFs → | Migration-strategy fit → | Migration success |
| Architecture readiness → | Institutional constraints weaken | Migration success |
| Vendor governance → | Institutional constraints weaken | Migration success |

FIG 1.PROPOSED CRITICAL SUCCESS FACTOR MODEL .

## IV. RESEARCH METHODOLOGY

### *A. Research design*

A sequential exploratory mixed method design is proposed . phase 1 uses Semi-structured interviews to refine constructs, terminology , and context specific risks. Phase 2 test the model through a cross sectional survey ,No primary results are claimed in this draft ; Empirical findings must be inserted only after ethics approval ,data collection , and analysis . Secondary evidence is used to establish constructs and hypothesis rather than to simulate outcomes.

### *B. Population and sampling*

The target population is managers and professionals who have participated in ERP cloud evaluation , migration ,implementation , governance ,or post go live operations in enterprises located in World Bank low ,lower middle or upper middle income economies , Eligible roles include CIO s, ERP program managers ,finance and operations leaders ,architects, security managers, consultants, and Key users .Purposive sampling will recruit 20 -30 interviewees across sectors and organizational size ; this pilot test qualified respondents.

### *C. Measurement Instrument*

Each CSF will be measured reflectively using three to five seven-point Likert items, anchored from 1 = strongly disagree to 7 = strongly agree. Migration-strategy fit will assess alignment between selected approach and legacy complexity, regulatory needs, organizational capacity, and risk tolerance. Success will combine project performance, operational continuity, quality, adoption, compliance, and realized benefits. Item will be adapted from relevant research [1]-[3],[7],[8],[11],[13]Items will be adapted from prior ERP, cloud, TOE, and IS-success studies, reviewed by experts, and pilot-tested with 20–30 qualified respondents. The separate data-collection toolkit accompanying this paper provides screening, construct items, interview prompts, consent wording, and a coding framework.

### *D. Analysis and Quality Controls*

Interview data will be coded using a predefined CSF framework plus inductive codes for emergent conditions. Two-coder agreement and an audit trail will support dependability. Survey analysis will screen minnesingers, careless responses, outliner, and common-method bias. The measurement model will be assessed using indicator loading's, composite reliability , average variance extracted and HTMT discriminant validity. The structural model will report col-linearity, standardized path coefficients, bootstrapped confidence intervals, $R^2$, effect sizes, predictive relevance, and out-of-sample prediction where feasible [10]. H7 will use bootstrapped indirect effects ;H8 will use Mediation and moderation will be tested with bootstrapped indirect and interaction effects .. Sector, firm size, ERP vendor, migration stage, and country-income group will be considered controls.

### *E. Ethics and Reproducibility*

Participation will be voluntary and based on informed consent. The study will avoid collecting confidential configuration details, credentials, or identifiable incident data. Organization and respondent identities will be removed; Small cells will be suppressed. The protocol, questionnaire, code-book, exclusion rules, and analysis syntax should be time-stamped before hypothesis testing. Results should report null findings and limitations, not only supported hypotheses.

## V. EXPECTED CONTRIBUTIONS AND MANAGERIAL USE

TABLE II
STRATEGY-FIT DECISION LOGIC

| Gate | Evidence | If conditional or weak |
|---|---|---|
| 1 Strategic ownership | Business case ; sponsor; benefits ;decision rights | Clarify govemance before commitment |
| 2 process and data | Complexity; Standardization; data quality; remediation | Selective transformation or phasing |
| 3Architecture/Security | Interfaces,IAM; cyber controls; backup;connectiviy | Mitigate ciritical risks before go-live |
| 4 Skills /change | Skills,participation;training; leadership | Build capability and stage deployment |
| 5Provider / ecosystem | Capacity's ;contract,protability;exit | Strengthen govmance and contingency |
| 6institutional context | Infrastructure; regulation; residency; FX exposure | Regulatory review, mitigation,or phased hybrid |

The theoretical contribution is the integration of generic cloud ERP CSFs with emerging market institutional readiness

and creates value when it is translated into a strategy appropriate to local and organizational conditions. It also broadens success beyond time , cost, and scope by including continuity ,compliance , adoption, quality, and benefits.

For managers, the model supports a stage gate decision .gate 1 verifies strategic ownership and measurable benefits. Gate 2 evaluates process and data debt . gate 3 test architecture , integration , security ,and resilience .gate 4 verifies skills and change capacity gate 5 assesses provider, panther , contract, portability,and exit arrangements .gate 6 evaluate local infrastructure, regulation, data sovereignty, and currency exposure. Weak gates do not automatically prohibit migration; they indicate mitigation actions and may favor selective or phased hybrid strategies over rapid re-implementation.

## VI. LIMITATIONS AND FUTURE WORK

The proposed model is intentionally parsimonious and may omit industry-specific requirements. Cross-sectional self-reports limit causal inference and can overstate success. Multi-country sampling introduces regulatory and cultural heterogeneity, While access through professional networks can create selection bias. Future work should use longitudinal checkpoints before migration, at go-live, and after stabilization; triangulate responses with project documents and service metrics; test measurement in-variance across regions and firm sizes; and compare public cloud, private cloud, and hybrid deployments.

## VII. CONCLUSION

Enterprise ERP cloud migration in emerging markets is a Organizational transformation under institutional constraints, not merely a technical relocation. The proposed model identifies six CSFs and positions migration-strategy fit as the mechanism linking readiness to success. By explicitly incorporating infrastructure, regulation, ecosystem maturity, and macroeconomic exposure, The model offers a context-sensitive basis for empirical testing and management decisions. The next research step is ethical primary-data collection and transparent validation rather than unsupported claims of success.

## REFERENCES


[1] Q.huang,m.Rahim,S. Foster,and M.Anwar , “Critical success factors affecting implementation of cloud ERP systems: A systematic literature review with future research possiblitites ,”proc 54th hawaii int.conf. Syst Sci .,pp 4683-4692,2021,doi: 10.24251/HICSS.2021.569

[2] S. Tongsuksai, S. Mathrani, and A. Taskin, “Cloud enterprise resource planning implementation: A systematic literature review of critical success factors,” in Proc. IEEE Asia-Pacific Conf. Comput. Sci. Data Eng., 2019, doi: 10.1109/CSDE48274.2019.9162373.

[3] L.M.Barbieri ,M.K.Sott and J.M.Monticelli ,” critical success factors for implemeting cloud ERP in SMEs; a systematic review” int j.innoy technol manag,vol.21 ,no 3,art no 2430003,2024 doi :10.1142/S019877024300039

[4] World Bank, Digital Progress and Trends Report 2023. Washington, DC, USA: World Bank, 2024.

[5] World Bank and International Finance Corporation, Advancing Cloud and Data Infrastructure Markets: Strategic Directions for Low- and Middle- Income Countries. Washington, DC, USA: World Bank, 2024

[6] SAPinsider, ERP Migration and Transformation 2026, Research Report, Mar. 2026.

[7] L. G. Tornatzky and M. Fleischer, The Processes of Technological Innovation. Lexington, MA, USA: Lexington Books, 1990.

[8] W. H. DeLone and E. R. McLean, “The DeLone and McLean model of information systems success: A ten-year update,” J. Manag. Inf. Syst., vol. 19, no. 4, pp. 9–30, 2003, doi: 10.1080/07421222.2003.11045748.

[9] Z. Qiang and G. de Salins, “Cloud adoption: A catalyst for cyber resilience in developing countries,” World Bank Blogs, Feb. 27, 2025. [Online]. Available: https://blogs.worldbank.org/en/digital-development/cloud-adoption--a-catalyst-for-cyber-resilience-in-developing-co

[10] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 3rd ed. Thousand Oaks, CA, USA: Sage, 2022.AlBar and M. R. Hoque, “Factors affecting cloud ERP adoption in Saudi Arabia: An empirical study,” Inf. Dev., vol. 35, no. 1, pp. 150–164, 2019, doi: 10.1177/0266666917735677.

[11] T. H. Davenport, “Putting the enterprise into the enterprise system,” Harvard Bus. Rev., vol. 76, no. 4, pp. 121–131, 1998.

[12] Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, “User acceptance of information technology: Toward a unified view,” MIS Q., vol. 27, no. 3, pp. 425–478, 2003, doi: 10.2307/30036540.

[13] National Institute of Standards and Technology, “Zero trust architecture,” NIST Special Publication 800-207, Aug. 2020, doi: 10.6028/NIST.SP.800-207.

# LLM-Assisted Navigation of Sri Lankan Procurement Regulations

B.A.N.N. Wimalarathne
*Department of Computer Science*
*Faculty of Applied Sciences University of Sri Jayewardenepura Gangodawila, Nugegoda, Sri Lanka*
as2021368@sci.sjp.ac.lk

R.G.N. Meegama
*Department of Computer Science*
*Faculty of Applied Sciences University of Sri Jayewardenepura Gangodawila, Nugegoda, Sri Lanka*
rgn@sjp.ac.lk

***Abstract*—Public procurement in Sri Lanka is governed by a fragmented set of regulatory documents, making it difficult for stakeholders such as small and medium-sized enterprises (SMEs), new market entrants, and procurement officers to locate, interpret, and apply procurement regulations effectively. Although the PROMISE e-Government Procurement System digitizes procurement transactions, it does not provide intelligent support for understanding regulatory requirements. This paper presents a Retrieval-Augmented Generation (RAG)-based question-answering assistant designed to improve access to Sri Lanka's government procurement regulations. Two configurations were developed and evaluated. The baseline Type 1 system employed dense FAISS retrieval with all-MiniLM-L6-v2 embeddings while the optimized Type 2 system incorporated larger context-preserving chunks, BAAI/bge-small-en-v1.5 embeddings, hybrid retrieval using FAISS and BM25, Reciprocal Rank Fusion (RRF) and cross-encoder re-ranking. Evaluation across 250 procurement-related questions demon-strated substantial improvements in the Type 2 configuration. Strict Precision@1 increased from 0.348 to 0.800, Strict Mean Reciprocal Rank (MRR) rose from 0.537 to 0.852, Factual Correctness improved from 0.35 to 0.42, and domain-specific Method Correctness increased from 0.21 to 0.57. While automated RAGAS Faithfulness (0.49 to 0.38) and Context Precision (0.61 to 0.31) reflected penalties associated with larger context chunks and multi-step model synthesis, overall domain accuracy and retrieval precision showed significant gains. The results demonstrate that retrieval optimization and re-ranking are key determinants of procurement question-answering quality.***

***Keywords—LLM, Sri Lanka Procurement Guidelines, RAG***

## I. Introduction

Public procurement is a critical component of government expenditure and economic governance in Sri Lanka. The procurement process is governed by a collection of regulatory documents, including the Procurement Guidelines Goods & Works (2006), Procurement Manual (2006), standard bidding documents, circulars, and legislative acts. Together, these documents define the procedures, thresholds, and compliance requirements that govern the acquisition of goods, works and services by public institutions. However, the regulatory framework is fragmented across multiple sources, requiring users to navigate and cross-reference numerous documents to obtain complete procedural guidance.

Retrieval-Augmented Generation (RAG) addresses these challenges by combining information retrieval with language generation. Instead of relying solely on parametric knowledge, RAG systems retrieve relevant information from authoritative documents and use the retrieved context to generate grounded responses. This approach improves factual accuracy, reduces hallucinations, and enables citation-based explanations derived directly from source documents.

This paper presents a RAG-based intelligent assistant for navigating Sri Lankan government procurement regulations. Using the Procurement documents as the primary knowledge base, the proposed system integrates document parsing, se-mantic retrieval, hybrid search, and large language model gen-eration to provide context-aware procurement guidance. The study further evaluates a baseline and an optimized retrieval architecture to investigate the impact of retrieval quality on answer relevance, factual grounding and procurement-specific correctness.

## II. Literature Review

Government procurement documents, including procure-ment guidelines, manuals, circulars, and standard bidding documents, are typically lengthy, heterogeneous, and legally complex [1]. These documents are primarily designed for human interpretation rather than machine processing, making information retrieval and compliance verification difficult. Procurement regulations are often distributed across multiple document types and are periodically updated, increasing the risk of ambiguity and reliance on outdated information.

Semantic retrieval has become a fundamental component of modern information retrieval systems. Sentence-transformer models such as SBERT generate dense vector representations that enable semantic matching between user queries and docu-ment passages. However, retrieval performance in specialized domains often depends heavily on the choice of embedding model. Recent retrieval-oriented embedding models, such as BAAI's BGE family, employ contrastive learning and hard-negative mining to improve document retrieval accuracy while maintaining computational efficiency [2], [4].

Vector databases such as FAISS provide scalable similarity search over large document collections and have become widely

adopted in Retrieval-Augmented Generation systems. Nevertheless, dense retrieval alone may fail to capture exact keyword matches that are important in legal and regulatory documents. Hybrid retrieval approaches therefore combine dense vector search with sparse retrieval methods such as BM25 to improve retrieval coverage [5], [6].

### *A. Retrieval-Augmented Generation*

RAG combines information retrieval with large language model generation to produce responses grounded in exter-nal knowledge sources. Unlike standalone language models, which rely solely on pre-trained parameters, RAG systems dynamically retrieve relevant documents and incorporate them into the generation process. Recent research has demonstrated the effectiveness of RAG in domain-specific applications such as legal, healthcare, and enterprise information systems. Advanced RAG frameworks further improve reliability through retrieval optimization, evidence evaluation, and corrective retrieval mechanisms [7]–[10].

## III. METHODOLOGY

### *A. Data Collection*

The data for this research were obtained exclusively from official government sources to ensure the highest level of validity, credibility, and traceability. The primary materials comprised the Procurement Guidelines Goods & Works (2006) and the Procurement Manual (2006), both published by the Ministry of Finance, Planning, and Economic Development of Sri Lanka. The 2006 editions were intentionally selected as they represent the earliest and most comprehensive standard-ized versions, serving as the foundation for all later revisions, circulars, and supplementary policy updates.

### *B. Document Parsing*

All procurement documents were converted into hierarchical JSON structures that preserve their legal organization, including chapters, sections, subsections, and bullet points. Each section stores both a concise summary view and the complete text to support retrieval and citation. PDF extraction was performed using PyMuPDF, while regular expressions and Unicode normalization were used to identify document structure and ensure consistent formatting. The resulting JSON representation provides a structured, versionable, and retrieval-ready knowledge base for subsequent chunking, embedding, and RAG-based question answering.

PDFs expose text as positioned “blocks,” not in reading order. To linearize them, we sort blocks top-to-bottom, then left-to-right, which mirrors how humans read single-column English pages and thus recreates the natural flow from raw coordinates. This rule is deterministic and reproducible, so the same PDF always linearizes the same way and ideal for reliable parsing and embeddings [11].

Formally, for two blocks $b_i = (x_i, y_i, ..)$ and $b_j = (x_j, y_j, ..)$ ,

$$b_i \prec b_j \Leftrightarrow (y_i < y_j) \vee (y_i = y_j \wedge x_i < x_j) \quad (1)$$

which approximates human reading flow for the single-column typography of these documents and guarantees reproducibility.

In this notation, $x_i$ and $y_i$are the block’s horizontal and vertical coordinates on the PDF page, and the symbol $b_i \prec b_j$ means ’block $b_i$ comes before block $b_j$’ in reading or-der. The concatenated page text is then normalized: Unicode NFKC removes compatibility variants; ligatures (fi/fl), smart quotes, and long dashes are harmonized to plain ASCII; non-breaking spaces become spaces; and de-hyphenation merges words split across line breaks (rejoining “Procure-\nment” → “Procurement”). This yields a stable token stream that prevents typographic noise from polluting structural matching or embeddings.

Structure is recovered by numbering conventions, not visual layout, to stay robust across formatting. Chapter headers accept Arabic or Roman numerals; Roman tokens are mapped to integers by standard subtractive parsing (e.g. IV=5−1=4, XL=50−10=40). Within each chapter’s span, numbered head-ings (e.g.,1.1, 1.1.1, 1.1.1.1) are located and sorted by their byte offsets; if the i-th heading ends at $e_i$and the next begins at $s_{i+1}$, the section body is the half-open slice

$$\text{body}_i = \text{text}[\, e_i : s_{i+1}) \quad (2)$$

A hierarchy is then built by interpreting the number of dots in a heading as its depth (dot-count +1) and attaching each node to the nearest shallower predecessor (stack discipline). Inside each body, lettered bullets (a),(b),. . . are segmented into a structured list while the preface retains the rule’s leading sentence(s) for clean semantic indexing.

### *C. Chunking and Embedding*

*Type 1: Initial Chunking and Embedding Strategy:* The Type 1 configuration employed a straightforward chunking and embedding pipeline to prepare procurement documents for retrieval. Each parsed section was converted into a text representation consisting of the section heading, body content, and associated bullet points. The text was segmented using a Recursive Character Text Splitter with a chunk size of 500 characters and an overlap of 50 characters, allowing contextual continuity between adjacent chunks while preserving para-graph and bullet boundaries where possible. Each chunk was associated with metadata, including document source, section identifiers, and chunk index, to support retrieval and citation. The resulting text segments were embedded using the {all-MiniLM-L6-v2} SentenceTransformer model, which generates 384-dimensional semantic vectors. These embeddings were stored together with their metadata and indexed in a FAISS vector database to enable efficient semantic retrieval during question answering [12].

*Type 2: Optimized Chunking and Embedding Strategy After Fine-Tuning:* The chunking process can be formalized as follows [13], [14]. Let C denote the full plain-text representation

of a parsed section with character length |C|. A sequence of overlapping chunks is given as :

$C_1, C_2, \ldots, C_n$ where $|C_i| \leq$ chunk_size

Consecutive chunks share an overlap region defined by:

$C_i \cap C_{i+1} =$ suffix($C_i$, chunk_overlap)

(3) where,

$$n \approx \frac{L - \text{chunk_overlap}}{\text{chunk_size} - \text{chunk_overlap}}$$

For Type 1, the number of chunks produced for a section of length L is approximately:

*D. Answer Generation using RAG*

The answer generation module follows a RAG approach in which relevant procurement document chunks are retrieved and supplied as context to a large language model, reduc-ing unsupported responses and improving traceability. In the baseline Type 1 configuration, retrieved chunks are provided to the {llama-3.1-8b-instant} model with a temperature of 0.0 to generate deterministic, document-grounded responses [15]. The optimized Type 2 configuration enhances this process through hybrid retrieval, cross-encoder re-ranking, session memory, and query rewriting before generating responses using the {llama-3.3-70b-versatile} model at the same tem-perature setting. The generated answers are returned together with supporting document references to improve transparency and verifiability.

## IV. Results and Discussion

$$n \approx \frac{L - 50}{500 - 50} = \frac{L - 50}{450}$$

(4)

The 250 questions were constructed manually based on the Goods and Works Procurement Guidelines and the Procure-ment Manual (2006). They cover typical scenarios encoun-

For Type 2, the number of chunks produced for a section of length L is approximately:

$$n \approx \frac{L - 150}{800 - 150} = \frac{L - 150}{650}$$

tered by SMEs, procurement officers, bidding contractors and auditors, spanning simple definitions to multi-step financial decision rules.

The questions were structured across seven functional categories to evaluate different aspects of RAG capability: (1) Gen-eral procedures/definitions, (2) Procurement method selection,

(3) Threshold-based decisions, (4) Boundary value cases, (5) Direct contracting exceptions, (6) Citation/reference checking,

After generating the semantic embeddings, the vectors are

indexed using FAISS (Facebook AI Similarity Search) to create an efficient vector database for similarity-based re-trieval. The vector database stores semantic embeddings of procurement document chunks to support efficient retrieval. In the baseline Type 1 configuration, chunks were embedded using the {all-MiniLM-L6-v2} model and indexed in separate FAISS databases for the procurement documents. The optimized Type 2 configuration replaced this with the retrieval-focused {BAAI/bge-small-en-v1.5} model and adopted a hybrid retrieval strategy that combined FAISS-based semantic search with BM25 keyword retrieval. The outputs of both retrieval methods were merged using Reciprocal Rank Fusion (RRF), improving retrieval precision and relevance, particularly for procurement queries containing specialized regulatory terminology and procedural references.

*B. Query Handling System*

The query handling module prepares user questions for retrieval and answer generation. In the baseline Type 1 con-figuration, queries are processed independently and encoded using the {all-MiniLM-L6-v2} model before semantic retrieval from FAISS indexes. The optimized Type 2 configuration extends this process by incorporating session memory and query rewriting, enabling the system to interpret follow-up questions and maintain conversational context.

and (7) Rephrased query consistency.

Ground-truth answers were manually authored directly from the regulatory text. Each answer includes both the factual answer string and the explicit clause citations against which system outputs were benchmarked. All 250 question-answer-citation triplets underwent a two-pass manual review to eliminate ambiguity and verify that all ground-truth answers were fully supported by the raw PDF source text.

The results are reported across four sub-sections: model-wise evaluation, RAGAS general evaluation, procurement-specific custom metric analysis, and retrieval performance evaluation. The DeepEval 250-

question results are reported at the end of the custom metric sub-section.

*Model-Wise Evaluation - Type 1:* The model-wise evaluation was conducted using five representative real-world procurement questions under both the with-ground-truth and without-ground-truth settings. This controlled comparison uses the same five questions for both models, ensuring that any performance difference is attributable to model capacity rather than question difficulty variation.

Figure 1 summarizes the average Answer Relevancy Scores. The llama-3.3-70b-versatile model achieved 0.925 against 0.769 for llama-3.1-8b-instant, a gap of 0.156 points. This confirms that under the Type 1 retrieval context, the larger model had a substantial advantage in producing answers aligned with expected procurement outcomes.

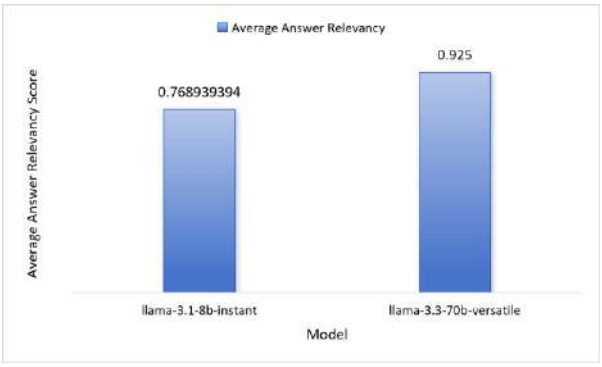



Figure 1. Average Answer Relevancy Score by Model for Type 1 RAG System.

*Procurement-Specific Custom Metric Results - Type 1:* Figure 2 presents category-wise performance. Performance varied substantially across question categories, with the clearest trend being stronger performance in general procedure and definition questions where a correct answer can be assembled from a direct documentary reference and weaker performance in threshold-based questions and boundary value cases, where the correct answer requires the system to apply a financial rule rather than retrieve and cite text.

*Retrieval Evaluation - Type 1:* The retrieval component of the Type 1 system was evaluated independently using the metrics below. The results, summarized in Table I, confirm the high-coverage, low-precision profile identified in the RAGAS evaluation.

TABLE I
. RETRIEVAL EVALUATION RESULTS OF THE TYPE 1 RAG SYSTEM

| Metric | Score |
|---|---|
| Strict Mean Precision@1 | 0.348 |
| Strict Mean Precision@3 | 0.289333 |
| Strict Mean Precision@5 | 0.276 |
| Strict Mean Precision@8 | 0.223 |
| Strict Mean Recall@8 | 0.72019 |
| Strict MRR | 0.53701 |
| Flexible Mean Precision@1 | 0.352 |
| Flexible Mean Precision@3 | 0.294667 |
| Flexible Mean Precision@5 | 0.2808 |
| Flexible Mean Precision@8 | 0.227 |
| Flexible Mean Recall@8 | 0.92 |
| Flexible MRR | 0.54221 |

The Strict Precision@1 score of 0.348 is the most consequential single retrieval finding for Type 1. It indicates that the first chunk provided to the generator which typically exerts the most influence on the generated answer contained the correct evidence for fewer than one-third of all evaluated questions.

*Model-Wise Evaluation - Type 2:* The model-wise evaluation for Type 2 used the same five procurement questions as the Type 1 model comparison, enabling direct assessment of how the optimized retrieval context affected each model's performance.

Figure 3 shows the average Answer Relevancy Scores. The llama-3.3-70b-versatile model achieved 0.709 and llama-

3.1-8b-instant achieved 0.706, a gap of only 0.003 points compared to the 0.156-point gap observed under Type 1. This near-complete convergence confirms that the Type 2 retrieval improvements effectively equalized the two models' answer relevancy performance.

*RAGAS-Based General Evaluation - Type 2:* Context Recall achieved approximately 0.71, marginally higher than the Type

1 score of 0.69. This small addition is attributable to the reduced context window size: using five chunks instead of eight means that some chunks containing relevant evidence that would have appeared in positions six through eight under Type 1 are now excluded. That the recall addition was only

0.02 points despite using three fewer chunks confirms that the reranking stage successfully concentrated the most rele-vant evidence into the top-five positions. Factual Correctness improved to approximately 0.43 from the Type 1 value of 0.34, indicating that the generated answers were more factually aligned with expected content. Faithfulness scored approx-imately 0.38 lower than the Type 1 RAGAS Faithfulness of 0.49 and LLM Context Precision with Reference scored approximately 0.31, also lower than the Type 1 score of 0.6.

*Procurement-Specific Custom Metric Results - Type 2:* This section presents the results of the Type 2 RAG system using 250 procurement-related questions. The questions covered different areas such as procurement method selec-tion, threshold-based decisions, direct contracting exceptions, reference checking, boundary cases, rephrased questions, and general procedure questions.

Figure 4 presents the category-wise performance breakdown. The Type 2 system achieved its strongest results in direct contracting exceptions, citation and reference checking, and rephrased query consistency, categories characterized by explicit procedural rules and clearly identifiable source sec-tions. Threshold-based questions and boundary value cases remained the two weakest categories, confirming that financial threshold application is a persistent challenge that is not fully resolved by retrieval improvements.

TABLE II
. RETRIEVAL EVALUATION RESULTS OF THE TYPE 2 RAG SYSTEM.

| Metric | Score |
|---|---|
| Strict Mean Precision@1 | 0.8000 |
| Strict Mean Precision@3 | 0.6480 |
| Strict Mean Precision@5 | 0.5768 |
| Strict Mean Recall@5 | 0.7100 |
| Strict MRR | 0.8516 |
| Flexible Mean Precision@1 | 0.8280 |
| Flexible Mean Precision@3 | 0.7013 |
| Flexible Mean Precision@5 | 0.6272 |
| Flexible Mean Recall@5 | 0.9560 |
| Flexible MRR | 0.8780 |

*Retrieval Evaluation — Type 2:* The Type 2 retrieval results show strong precision and ranking performance. The strict Precision@1 score of 0.8000 indicates that the first retrieved chunk was correct for most questions. The strict MRR score of 0.8516 further confirms that relevant evidence appeared early in the ranked retrieval list. The strict Recall@5 score of 0.7100

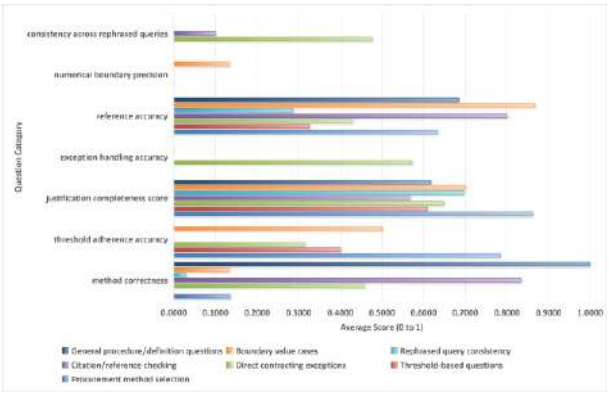



Figure 2. Category-wise Procurement-Specific Performance of Type 1 RAG System

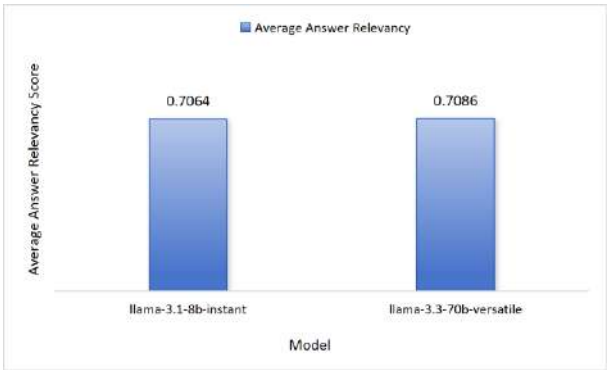



Figure 3. Average answer relevancy score by model for type 2 RAG system

### C. Cross Configuration Comparison

This section directly compares the Type 1 and Type 2 configurations across all evaluation dimensions. Where metrics produce results that appear inconsistent across frameworks, the underlying causes are explained in detail to support accurate interpretation.

*Procurement-Specific Custom Metric Comparison:* Figure

5 presents the overall comparison of Type 1 and Type 2 across the procurement-specific custom evaluation metrics. The results show that the Type 2 RAG system consistently outperformed the Type 1 system across all measured criteria, indicating that the optimized retrieval pipeline had a clear positive impact on domain-specific answer quality.

*RAGAS Metric Comparison:* Figure 6 presents the RAGAS comparison between the Type 1 and Type 2 RAG systems. The results show that Type 2 achieved a slightly higher Context Recall score than Type 1 with values of 0.71 and 0.68, respectively. Although Type 1 achieved higher Context Precision (0.61 vs. 0.31) and Faithfulness (0.49 vs. 0.38), indicating stronger reliance on retrieved context, Type 2 achieved better Factual Correctness (0.42 vs. 0.35).

*Retrieval Performance Comparison:*

### D. Limitations and Future Work

Despite the improvements achieved by the optimized Type 2 architecture, several limitations remain. The system is primarily grounded in the procurement documents (2006), and therefore, may not reflect subsequent regulatory amendments or circulars. Numerical threshold handling also remains challenging, as procurement method selection often requires de-terminist rule-based reasoning that cannot be fully addressed through retrieval improvements alone.

## V. CONCLUSIONS

This paper presented a RAG-based intelligent assistant for navigating Sri Lanka's government procurement

regulations. Using the Procurement Guidelines (2006) and Procurement Manual (2006) as the primary knowledge base, the system was developed to improve access to procurement information for stakeholders such as SMEs, community organizations, new bidders, and procurement officers. Experimental results demonstrate that the optimized retrieval pipeline significantly improved retrieval quality, answer relevance, contextual grounding, and hallucination reduction compared to the baseline configuration. While challenges remain in handling

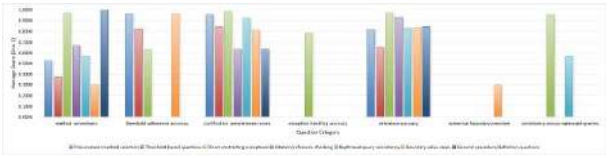


Figure 4. Category-wise procurement-specific performance of Type 2 RAG system.

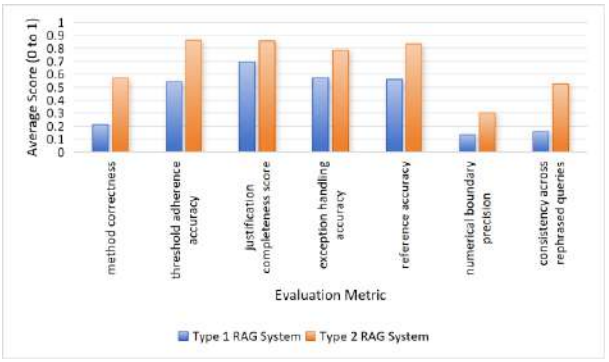



Figure 5. Overall custom metric comparison of Type 1 and Type 2 R[1A6G] systems.

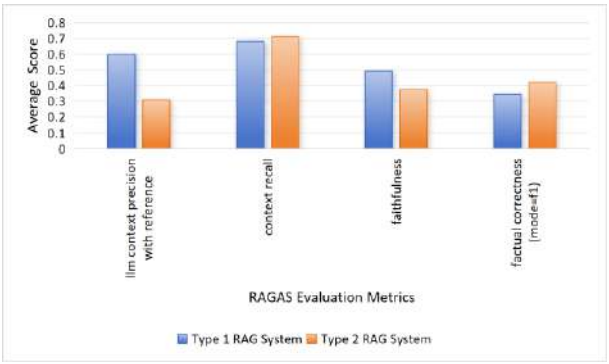



Figure 6. Overall RAGAS Metric Performance of Type 1 and Type 2 RAG Systems.

## REFERENCES


[1] G. Weerasinghe, B. KanaheraArachchi, M. Jayaweera, I. Aryaratne,
[2] U. Samarasinghe, and N. Karunaratna, "Effectiveness of implement-ing e-government procurement in sri lanka," in *Proceedings of the International Conference on Industrial Engineering and Operations Management, Istanbul, Turkey*, 2022.
[3] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.
[4] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-pack: packaged resources to advance general chinese embedding (2023)," *arXiv preprint arXiv:2309.07597*, 2023.
[5] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE transactions on big data*, vol. 7, no. 3, pp. 535–547, 2019.
[6] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, vol. 4. Now Publishers Inc, 2009.
[7] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, 2009.
[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal,
[9] H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
[10] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval aug-mented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11,pp. 1–17, 2023.
[11] S. Sharma, D. S. Yoon, F. Dernoncourt, D. Sultania, K. Bagga,
[12] M. Zhang, T. Bui, and V. Kotte, "Retrieval augmented generation for domain-specific question answering," *arXiv preprint arXiv:2404.14760*, 2024.
[13] M. Stäbler, S. Turnbull, T. Müller, C. Langdon, J. Marx-Goméz, and
[14] F. Köster, "The impact of chunking strategies on domain-specific infor-mation retrieval in rag systems," in *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pp. 1–6, IEEE, 2025.
[15] D. D. A. Bui, G. Del Fiol, and S. Jonnalagadda, "Pdf text classification to leverage information extraction from publication reports," *Journal of biomedical informatics*, vol. 61, pp. 141–148, 2016.
[16] M. Günther, I. Mohr, D. J. Williams, B. Wang, and H. Xiao, "Late chunking: contextual chunk embeddings using long-context embedding models," *arXiv preprint arXiv:2409.04701*, 2024.
[17] T. Nguyen, H. Pham, and K. Tran, "Hierarchical chunking for improved retrieval-augmented generation," in *Proceedings of the 12th Workshop on Vietnamese Language and Speech Processing (VLSP 2025)*, Association for Computational Linguistics, 2025.
[18] B. Smith and A. Troynikov, "Evaluating chunking strategies for re-trieval," tech. rep., Chroma, July 2024.
[19] M. Renze, "The effect of sampling temperature on problem solving in large language models," in *Findings of the association for computational linguistics: EMNLP 2024*, pp. 7346–7356, 2024.

# Ethical Concerns of Artificial Intelligence Usage in Higher Education: A Study Among Undergraduate Students in Sri Lanka

Sanduni Wickramanayake
*Department of Software Engineering & Computer Security*
*NSBM Green University Homagama, Sri Lanka*
wksanduni@students.nsbm.ac.lk

D. Rajasinghe
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
dharani.r@nsbm.ac.lk

G.D.S.Silva
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
gdssilva@students.nsbm.ac.lk

I.G.S.Minodi
*Department of Software Engineering & Computer Security*
*NSBM Green University Homagama, Sri Lanka*
igsmilluppalla@students.nsbm.ac.lk

***Abstract*—The use of Artificial Intelligence (AI) in higher education is growing rapidly, bringing with it both opportunities and ethical concerns. The study explores the ethical issues of AI usage among undergraduate students in Sri Lanka, an area that has not been extensively studied. Using purposive and convenience sampling, 245 undergraduate students from three universities were surveyed on their perceived concern in five areas: academic integrity, data privacy, algorithmic bias, transparency, and critical-thinking dependency. Academic integrity was the most highly rated concern (M = 4.12), followed by data privacy (M = 3.97) and critical-thinking dependency (M = 3.85). There were significant differences in scores across gender, faculty and prior AI-ethics instruction. There was a lack of policy communication, as only 28.6% of respondents knew of a formal institutional AI policy. The results are indicative and not generalizable due to the non-probability sample and self-reported data. The study provides recommendations for students, educators, and policy makers regarding the responsible use of AI in higher education.**

***Keywords—Artificial intelligence ethics, Academic integrity, Higher education, Sri Lanka, Undergraduate students***

## I. INTRODUCTION

In the twenty-first century, Artificial Intelligence (AI) has become one of the most impactful technologies, influencing almost every aspect of human life, including education. In the realm of higher education, AI tools, from the large language models like ChatGPT to adaptive learning platforms and automated assessment systems, are transforming the learning, research, and knowledge consumption experiences of students[1] . These advances have significant pedagogical benefits, but they also raise a complex set of ethical issues that require careful academic study.

As a South Asian emerging economy, Sri Lanka is seeing a gradual but steady adoption of AI tools in the university environment[2] . The government's efforts to incorporate AI education into schools in 2024 represent a broader national push towards digital transformation . However, the ethical aspects of AI use by undergraduate students, who are the main users of AI-supported learning, have not been adequately studied in the Sri Lankan context.

Ethical issues such as AI-assisted academic dishonesty, privacy concerns for personal data, bias in assessment and admission processes, overreliance on AI and the loss of critical thinking skills, and the absence of institutional governance frameworks have been documented in the literature globally[3], [4]. Most of these studies, however, are from the Western or East Asian context, and there is a significant gap in empirical research on students in developing countries like Sri Lanka.

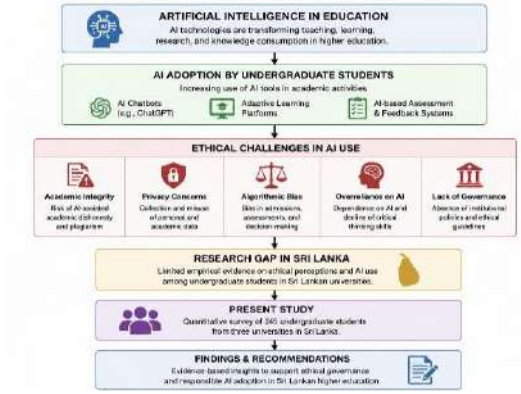



Fig. 1. Overview of AI Adoption and Ethical Challenges in Sri Lankan Higher Education.

This paper fills that gap by reporting the results of a quantitative survey study of 245 undergraduate students from three universities in Sri Lanka. Unlike earlier Sri Lankan studies that focus on AI adoption and academic performance [2] , this study examines perceived ethical concern across five dimensions and compares it across gender, faculty, and prior AI-ethics instruction. The study is designed to achieve the

following objectives: (i) To identify the major ethical issues perceived by the undergraduate students about the use of AI tools in their academic activities; (ii) To understand the extent of the adoption of AI tools and the specific use cases; and (iii) To suggest evidence-based recommendations for ethical governance of AI in Sri Lankan Higher Education Institutions.

The rest of this paper is organized as follows: Section II reviews the relevant literature; Section III states the research gap and problem statement; Section IV outlines the methodology; Section V presents the results and discussion; Section VI provides ethical recommendations; and Section VII concludes the paper.

## II. LITERATURE REVIEW

### *A. AI in Higher Education: Benefits and Prevalence*

The integration of AI into higher education has been characterized by rapid growth. Systematic reviews covering 2016–2022 demonstrate that publications on AI in higher education nearly doubled between 2021 and 2022 alone [5] . AI tools support personalized learning pathways, natural language processing-driven tutoring, automated feedback generation, and administrative efficiency [6] . Studies at Sri Lankan institutions confirm this trend, with students at the Sri Lanka Institute of Information Technology (SLIIT) predominantly using AI for grammar correction, knowledge enrichment, study-note generation, and coding assistance [7].

Fig. 2. Conceptual Framework of Ethical AI Issues in Higher Education.

### *B. Academic Integrity Concerns*

The emergence of generative AI has precipitated what scholars term ‘AI-giarism’—a novel form of academic dishonesty wherein students employ AI-generated content to fulfil academic obligations without genuine intellectual effort [8]. Perkins[9] highlights the difficulty of detecting AI-generated text and the need to revisit academic integrity policies, while evidence that generative AI has increased cheating is mixed: self-reported cheating among high-school students remained relatively stable after ChatGPT’s release[10]. Whether AI use amounts to misconduct also depends on institutional rules and disclosure requirements; for example, most SLIIT students in [7] did not regard AI use as cheating, although about half felt it reduced critical thinking. Studies of ChatGPT in essay assignments likewise point to tensions around learning, grading, and trust [11] .

### *C. Data Privacy and Algorithmic Bias*

Data privacy constitutes a significant ethical dimension. AI educational platforms collect substantial volumes of sensitive student data, and the absence of transparent data-use policies exposes students to potential misuse [12]. Student surveys indicate that privacy consistently ranks among the top ethical concerns, particularly around consent and opt-out provisions [13]. Algorithmic bias represents an equally critical issue: AI systems may perpetuate and amplify existing societal inequities in grading, admissions, and disability support, raising concerns about discriminatory outcomes [3] .

### *D. Dependency and Critical Thinking*

Scholars caution that uncritical reliance on AI-generated content may atrophy students’ analytical and argumentative capabilities[4]. Medical students in Sri Lanka have articulated concerns that AI usage negatively impacts critical thinking, peer interaction, and long-term retention of knowledge[14] . This aligns with broader international evidence suggesting that without structured pedagogical frameworks, AI may displace rather than augment genuine learning.

### *E. Institutional Policy and Governance*

The literature consistently underscores the absence of robust institutional AI governance frameworks as a primary gap. Researchers advocate for explicit university policies, mandatory AI literacy training, transparent algorithmic audits, and guidelines aligned with international ethical standards[3], [15] . In Sri Lanka specifically, barriers to responsible AI adoption include infrastructural limitations, policy voids, and unequal digital access across socio-economic groups [16] .

## III. RESEARCH GAP AND PROBLEM STATEMENT

Despite a growing body of international literature on AI ethics in education, studies focused on undergraduate students in Sri Lanka and the broader South Asian developing country context remain scarce. Existing Sri Lankan research on AI in higher education tends to focus on adoption patterns and academic performance rather than on the ethical perceptions and concerns of students [2], [7] . A systematic review of AI in Sri Lankan higher education identified this gap, noting that ethical dimensions receive insufficient empirical attention in the local context [17] .

Moreover, the majority of globally published studies on AI ethics in higher education are conducted in high-income countries where digital infrastructure, AI literacy, and institutional governance are comparatively advanced. These contexts do not adequately reflect the realities of students studying in resource-constrained environments with limited institutional guidance and inconsistent internet connectivity.

The problem statement of this study is therefore articulated as follows: Undergraduate students in Sri Lanka are increasingly engaging with AI tools in their academic activities without clear, well-communicated institutional guidance, ethical frameworks, or awareness of the associated risks, creating conditions conducive to academic misconduct, privacy violations, and the erosion of critical learning skills.

The contribution of this study is threefold. It (i) measures perceived ethical concern across five dimensions within a single Sri Lankan sample, rather than focusing on adoption or academic performance [2], [7], [17] ; (ii) tests whether concern differs by gender, faculty, and prior AI-ethics instruction; and (iii) links each recommendation to a measured finding. It thereby extends global findings [3], [4], [13] to a resource-constrained South Asian setting in which students report limited awareness of institutional guidance.

## IV. METHODOLOGY

### *A. Research Design*

This study uses a quantitative, cross-sectional survey design with self-reported data. For the purpose of systematically measuring the dimensions of ethical concern across a large enough sample of students, quantitative methodology was chosen, which would allow for statistical generalization to the larger population of undergraduate students in Sri Lanka.

### *B. Population and Sampling*

The target population were full-time undergraduate students enrolled in Sri Lankan universities. The three universities selected were purposively sampled, one a state university, one a private university and one a specialized technology university. Respondents were recruited through online university platforms and academic groups, so the sample is a non-probability (purposive and convenience) sample. A total of 245 valid responses were obtained: state university n = 79, private university n = 98, and technology university n = 68. Cochran's formula was used only as a benchmark: for an infinite population at 95% confidence and a 5% margin of error ($p = 0.5$) it gives $n \approx 384$, and the sample of 245 corresponds to a margin of error of about ±6.3% under simple random sampling, which was not used here. The results are therefore treated as indicative and are not statistically generalized to all undergraduates in Sri Lanka.

### *C. Instrument Development*

The questionnaire was developed by the authors by adapting the ethical themes reported in prior studies [3], [4], [13] ; it is not a single pre-validated scale. The instrument comprised four sections: (i) demographic information; (ii) AI tool usage patterns (frequency, tool type, academic purpose); (iii) ethical concern perceptions measured on a 5-point Likert scale (1=Strongly Disagree to 5=Strongly Agree) across five constructs—academic integrity, data privacy, algorithmic bias, transparency, and critical-thinking dependency; and (iv) institutional policy awareness. The 25 Likert statements are listed in Table III. Internal consistency was assessed with Cronbach's alpha for the whole instrument ($\alpha = 0.84$, indicating good consistency) and separately for each construct (Table I).

### *D. Data Collection*

Data were collected during February–April 2026 via an online survey distributed through university learning management systems and WhatsApp academic groups. Participation was voluntary, anonymous, and informed consent was obtained electronically prior to questionnaire completion.

### *E. Data Analysis*

IBM SPSS Statistics (v27) was used for data analysis. Descriptive statistics (mean, standard deviation, frequency, percentage) were computed for all variables. Independent samples t-tests and one-way ANOVA with Tukey post hoc tests were performed to examine differences in ethical concern perceptions by gender, faculty, and prior AI training. Effect sizes (Cohen's d and $\eta^2$) were calculated from the test statistics and group sizes. Statistical significance was set at $p < 0.05$.

## V. RESULTS AND DISCUSSION

### *A. Demographic Profile*

Of the 245 respondents, 52.2% (n=128) were female and 47.8% (n=117) were male. The majority (68.2%) were in their second or third year of study. Faculty distribution included Computing & IT (34.3%, n = 84), Business & Management (28.6%, n = 70), Engineering (22.0%, n = 54), and other disciplines, including Arts & Humanities, combined into one group (15.1%, n = 37). Only 31.4% (n = 77) of respondents reported having received any formal instruction on AI ethics from their institution.

### *B. AI Tool Usage Patterns*

A substantial majority of students (87.3%) reported using at least one AI tool in academic activities. ChatGPT was the most widely used tool (74.3%), followed by Grammarly (48.2%), Google Bard/Gemini (31.0%), and Copilot (18.4%). Primary use cases were essay drafting (61.2%), grammar correction (55.9%), generating study notes (49.4%), and coding assistance (38.8%). Notably, 43.7% of respondents self-reported using AI to produce substantial portions of assessed work. The survey did not record whether such use was permitted or disclosed under each institution's assessment rules, so this finding is not equated with misconduct but is relevant to academic integrity.

TABLE I
MEAN SCORES OF ETHICAL CONCERN DIMENSIONS

| Ethical Concern Dimension | Items | α | Mean | SD | Rank |
|---|---|---|---|---|---|
| Academic integrity | 5 | 0.79 | 4.12 | 0.73 | 1 |
| Data privacy and security | 5 | 0.76 | 3.97 | 0.81 | 2 |
| Critical-thinking dependency | 5 | 0.73 | 3.85 | 0.88 | 3 |
| Transparency and explainability | 5 | 0.71 | 3.74 | 0.92 | 4 |
| Algorithmic bias | 5 | 0.77 | 3.61 | 0.95 | 5 |
| Overall instrument | 25 | 0.84 | – | – | – |

### *C. Ethical Concern Perceptions*

Table I summarizes the mean scores for the five ethical concern dimensions. Academic integrity violations recorded the highest mean (M=4.12, SD=0.73), indicating that students view AI-assisted plagiarism as the most critical ethical issue, consistent with international evidence on increasing AI-enabled academic misconduct [9], [10] .

Data privacy ranked second (M=3.97), reflecting concerns about AI platforms collecting and using personal and academic data, consistent with the findings of Han et al.[13] . Respondents

also expressed discomfort with AI tools accessing browsing history and academic records.

Critical-thinking dependency ranked third (M=3.85), with many students reporting reduced motivation for independent problem-solving and deep reading, supporting findings from Sri Lankan medical education research [14] . Transparency concerns (M=3.74) reflected limited understanding of AI-generated outputs, while algorithmic bias received the lowest mean (M=3.61), which may reflect lower awareness of this issue among non-computing students (Table II).

### *D. Group Differences*

Table II reports the group comparisons. Female students reported higher privacy concern than male students ($t(243) = 2.41$, $p = .017$, $d = 0.31$). Algorithmic-bias awareness differed across the four faculty groups ($F(3, 241) = 4.87$, $p = .003$, $\eta^2 = 0.06$); Tukey post hoc tests showed that Computing & IT students scored significantly higher than each of the other three groups, which did not differ from one another. Students who had received formal AI-ethics instruction reported lower academic integrity concern than those who had not ($t(243) = 2.88$, $p = .004$, $d = 0.40$). These are small-to-medium effects, and because the design is cross-sectional the training difference is an association, not evidence that instruction caused lower concern.

TABLE II
GROUP COMPARISONS OF CONCERN AND AWARENESS SCORES

| Comparison (group n) | Test | Statistic | p | Effect |
|---|---|---|---|---|
| Privacy concern: female (128) vs. male (117) | t-test | $t(243) = 2.41$ | .017 | $d = 0.31$ |
| Bias awareness by faculty: Computing & IT (84), Business & Mgmt. (70), Engineering (54), Other (37) | ANOVA | $F(3, 241) = 4.87$ | .003 | $\eta^2 = 0.06$ |
| Integrity concern: no AI-ethics instruction (168) vs. instruction (77) | t-test | $t(243) = 2.88$ | .004 | $d = 0.40$ |

### *E. Institutional Policy Awareness*

Only 28.6% of respondents (n = 70) reported being aware of a formal AI usage policy at their institution, and 61.4% of them (n = 43) described the policy as inadequate or unclear. These figures describe students' awareness and perceptions; they do not show that policies are absent, because a policy may exist without being communicated.

The results therefore point to a gap in policy communication and awareness, consistent with reports of underdeveloped AI governance in Sri Lankan higher education [16] . Institutional policy documents were not audited in this study.

### *F. Limitation*

This study has several limitations. First, purposive and convenience sampling of three universities limits generalizability. Second, the data are self-reported and may be affected by social-desirability bias, especially for questions on AI use in assessed work. Third, the cross-sectional design captures perceptions at one point in time and cannot establish causality. Fourth, the questionnaire was adapted from prior themes rather than a validated scale, and construct validity was not tested (e.g., by factor analysis). Fifth, permission and disclosure rules for AI use were not recorded, and institutional policies were not examined directly. Future work should use probability-based samples, validated instruments, and a review of policy documents.

TABLE III
LIKERT STATEMENTS USED IN THE QUESTIONNAIRE (5-POINT SCALE)

| Construct | Statements |
|---|---|
| Academic integrity | Q11 Undisclosed AI use in assignments is academic dishonesty. Q12 AI content makes it harder to assess students' true abilities. Q13 My institution has clear rules on AI use in assessments. Q14, I know the consequences of submitting AI work as my own. Q15 AI should be allowed in all academic tasks without restriction. |
| Data privacy | Q16 I am concerned about personal data I share with AI platforms. Q17 AI tools I use have privacy policies I understand. Q18 I can control how AI platforms use my data. Q19 Universities should regulate which AI tools students use. Q20, I trust AI platforms not to misuse content I input. |
| Algorithmic bias | Q21 AI grading or feedback could be biased against some backgrounds or languages. Q22 AI tools reflect cultural or gender biases in training data. Q23 I am aware of algorithmic bias. Q24 AI admission or assessment tools could disadvantage some groups. Q25 Universities should audit their AI systems for bias. |
| Transparency | Q26, I understand how the AI tools I use reach their outputs. Q27 AI tools should indicate AI-generated content. Q28 AI in education should explain its reasoning. Q29 I can tell reliable from unreliable AI information. Q30 Lack of transparency creates ethical risks in education. |
| Critical-thinking dependency | Q31 Using AI reduces my need to think independently. Q32, I rely on AI more than I should. Q33 Excessive AI use may weaken students' problem-solving. Q34 I deliberately limit my AI use to keep my analytical skills. Q35 Universities should design assignments AI cannot easily complete. |

## VI. ETHICAL CONCERNS AND RECOMMENDATIONS

Based on the empirical findings, five key ethical concerns and corresponding recommendations are presented for stakeholders in Sri Lankan higher education. Each is linked to a finding of this study; given the sampling limitations, they are indicative rather than definitive.

### *A. Academic Integrity*

Concern: AI-assisted plagiarism and the blurring of authorship boundaries undermine the epistemic foundations of academic assessment. Academic integrity was the highest-rated concern (M = 4.12), and 43.7% of respondents self-reported using AI for substantial parts of assessed work, although permission and disclosure were not recorded.

Recommendation*:* Universities should establish explicit, faculty-specific AI usage policies aligned with international honor codes that define permitted, disclosed, and prohibited uses. Assessment redesign towards oral examinations, reflective journals, and in-class activities that are resistant to AI substitution is strongly advised [11] . AI detection tools, if used, should supplement rather than replace human judgment [9] .

### *B. Data Privacy and Security*

Concern*:* Data privacy was the second-highest concern (M = 3.97) and was higher among female students (Table II). Respondents' unease about how AI platforms collect and use personal and academic data indicates a need for clearer data-handling information.

Recommendation: Institutions should publish transparent data-handling protocols for all AI tools endorsed for academic use. Students must be empowered with meaningful consent mechanisms and opt-out options. National guidance on AI data practices in education should be developed, drawing on frameworks such as the EU's GDPR and UNESCO's Recommendation on the Ethics of AI (2021).

### *C. Algorithmic Bias and Discrimination*

Concern*:* AI systems used for grading, admissions, or student risk identification may embed and amplify biases related to gender, ethnicity, language, or disability status. Bias received the lowest concern rating (M = 3.61), and awareness was significantly higher among Computing & IT students than among other faculties (Table II).

Recommendation*:* Regular algorithmic audits of AI systems deployed by educational institutions should be conducted. Cross-disciplinary AI literacy curricula should address algorithmic bias, ensuring that all undergraduates—not only computing students—develop critical awareness of how AI systems can discriminate[3], [15] .

### *D. Transparency and Explainability*

Concern: Transparency was rated fourth (M = 3.74), reflecting limited understanding of how AI-generated outputs and feedback are produced, which can reduce students' agency in their own educational journey.

Recommendation: Educational AI tools procured by institutions should meet minimum explainability standards. Students should be taught interpretive skills to critically evaluate AI outputs rather than accepting them uncritically. Institutions should favor AI solutions that provide reasoning traces and confidence indicators alongside their outputs.

### *E. Critical Thinking and Cognitive Dependency*

Concern: Critical-thinking dependency ranked third (M = 3.85). Excessive reliance on AI for knowledge generation threatens students' development of analytical reasoning, independent research skills, and intellectual agency—core competencies of higher education.

Recommendation*:* Pedagogical frameworks should be redesigned to position AI as a cognitive scaffold rather than a cognitive substitute. Mandatory AI literacy programs should incorporate ethical use guidelines, metacognitive reflection exercises, and deliberate practice in AI-free analytical tasks. Faculty development initiatives are essential to ensure educators themselves are equipped to model and promote responsible AI use[4] .

## VII. CONCLUSION

This study examined the ethical concerns of AI use in higher education among Sri Lankan undergraduates. It identified five key concerns: academic integrity, data privacy, AI-induced dependency on critical thinking, lack of transparency, and algorithmic bias, with academic integrity rated as the greatest concern. Concern and awareness scores differed significantly by gender, faculty, and prior AI-ethics instruction. The findings also revealed low student awareness of institutional AI policies (28.6%) and limited AI-ethics training (31.4% had received instruction), highlighting the need for clearer communication of policy and stronger governance as AI adoption increases. This study contributes empirical evidence from an underrepresented South Asian context and recommends institutional AI policies, AI literacy initiatives, assessment redesign, stronger national governance, and responsible AI practices among students and educators. Because of the non-probability sample and self-reported data, the findings are indicative rather than generalizable. Future research should investigate evolving ethical perceptions through longitudinal and qualitative studies, use probability-based samples and validated instruments, and extend comparisons across South Asian higher education systems.

## REFERENCES


[1] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," Dec. 01, 2019, *Springer Netherlands*. doi: 10.1186/s41239-019-0171-0.

[2] U. M. L. A. Nawarathne, U. Thayasivam, and K. P. N. S. Dayarathne, "Usability and Academic Impact of Ai Tools on Higher Education in Sri Lanka," *KDU Journal of Multidisciplinary Studies*, vol. 7, no. 2, pp. 149–161, Nov. 2025, doi: 10.4038/kjms.v7i2.213.

[3] Z. Slimi and B. V. Carballido, "Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies," *TEM Journal*, vol. 12, no. 2, pp. 590–602, May 2023, doi: 10.18421/TEM122-02.

[4] M. T. K. Farooqi, I. Amanat, and S. M. Awan, "Ethical considerations and challenges in the integration of artificial intelligence in education: A systematic rReview," *Journal of Excellence in Management Sciences*, vol. 3, no. 4, pp. 35–50, 2024.

[5] R. D. Rajasinghe and D. Wijesinghe, "A Federated Multimodal Deepfake Detection Framework for Enhancing the Integrity of Video Evidence in Court Systems," 2026 9th International Conference on Electronics Technology (ICET), Chengdu, China, 2026, pp. 806-811, doi: 10.1109/ICET69987.2026.11659066.

[6] T. Shaik *et al.*, "A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis," *IEEE Access*, vol. 10, pp. 56720–56739, 2022, doi: 10.1109/ACCESS.2022.3177752.

[7] W. A. S. C. Weerasinghe and H. M. P. P. K. Abeysinghe, "Usage of Artificial Intelligence (AI) Tools for Academic Activities by Undergraduate Students: Quantitative Study at the Sri Lanka Institute of Information Technology (SLIIT) Library," 2024. Accessed: Oct. 31, 2026. [Online]. Available: https://rda.sliit.lk/handle/123456789/4025

[8] C. K. Y. Chan, "Students' perceptions of 'AI-giarism': investigating changes in understandings of academic misconduct," *Educ. Inf. Technol. (Dordr).*, vol. 30, no. 6, pp. 8087–8108, Apr. 2025, doi: 10.1007/s10639-024-13151-7.

[9] M. Perkins, "Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond," *Journal of University Teaching and Learning Practice*, vol. 20, no. 2, 2023, doi: 10.53761/1.20.02.07.

[10] V. R. Lee, D. Pope, S. Miles, and R. C. Zárate, "Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024, doi: 10.1016/j.caeai.2024.100253.

[11] C. C. Tossell, N. L. Tenhundfeld, A. Momen, K. Cooley, and E. J. De Visser, "Student Perceptions of ChatGPT Use in a College Essay

Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1069–1081, 2024, doi: 10.1109/TLT.2024.3355015.

[12] A. M. Tirado, P. Mulholland, and M. Fernandez, "Towards an Operational Responsible AI Framework for Learning Analytics in Higher Education." [Online]. Available: https://www.solaresearch.org/about/what-is-learning-analytics/

[13] R. D. Rajasinghe, "Generative AI and Its Implications for the Authenticity and Legal Use of Video Evidence," 2025 1st International Conference on Data Science and Intelligent Network Computing (ICDSINC), Raipur, India, 2025, pp. 17-22, doi: 10.1109/ICDSINC66221.2025.11448275.

[14] T. Seneviratne, K. Kodikara, I. Abeykoon, and W. Palpola, "Perception of Medical Undergraduates on Artificial Intelligence in Medical Education: Qualitative Exploration," *JMIR Med. Educ.*, vol. 11, 2025, doi: 10.2196/73798.

[15] C. K. Y. Chan, "A comprehensive AI policy education framework for university teaching and learning," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Dec. 2023, doi: 10.1186/s41239-023-00408-3.

[16] R. D. Rajasinghe, "Impact of Artificial Intelligence on Decision-Making in Information Systems," 2025 5th International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 2025, pp. 1-6, doi: 10.1109/ICARC64760.2025.10962922.

[17] W. S. Kodippili, "AI IN SRI LANKAN UNIVERSITIES :A SYSTEMATIC REVIEW OF ITS IMPACT ON EDUCATIONAL SUSTAINABILITY AND STUDENT EXPERIENCE," p. 127, 2024, doi: 10.64752/EWCN1373.

# A Household-Level Profiling Framework for Residential Electricity Consumption in Sri Lanka

Sanuri A. Divyanjalee
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
ksadivyanjalee@students.nsbm.ac.lk

Thilini Bakmeedeniya
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
thilini.b@nsbm.ac.lk

***Abstract*—Household electricity use is a significant energy consumption parameter for residential energy planning, however, there is limited evidence available at the household level that incorporates complete electricity billing histories alongside survey and smart-meter data. This research creates a framework for profiling households at the individual level based on a longitudinal Sri Lankan dataset made available by LIRNEasia which connects monthly electricity consumption data with account metadata, smart meter registers, and household surveys. The billing records for April 2023 to March 2024 were matched with the Wave 1 household descriptors and smart-meter metadata for 1,413 households. Households were classified into Low, Medium, and High consumption tertiles, and then followed by nonparametric profiling, controlled regression, nested cross-validated classification and validation using separate six-hour cumulative registers. Household and dwelling size profiles were larger for high consumption households, as were the numbers of appliances, fans and lights, whilst high consumption households also owned significantly more air-conditioners. Unadjusted depot differences were significant, but depot indicators were not jointly significant after household controls and did not contribute significantly to R-squared. The high-consumption profiling had a receiver operating characteristic-area under the curve (ROC-AUC) of 0.8207 and area under the precision-recall curve (PR-AUC) of 0.7194. The Pearson correlation for the registered reconstruction was 0.930, and the median absolute percentage error was 7.75 percent, both of which were high and agreed strongly with monthly billing records. The framework offers evidence of a household-centered demand analysis, for the matched Sri Lankan sample.**

***Keywords—household profiling, monthly billing, residential electricity consumption, smart-meter, Sri Lanka***

## I. Introduction

The electricity consumption of households is significant for Sri Lanka because the demand of electricity is linked to living conditions, appliances owned by the household, tariff exposure, and electricity-system planning. The national household evidence reveals that residential electricity consumption is a significant portion of total electricity consumption in Sri Lanka and electricity expenditure can be significant under the tariff condition in Sri Lanka [1]. Limited access to modern cooking fuels is also a significant source of deprivation, and there are also income-, socio-demographic- and sub-national-level differences in energy poverty [2]. The distinction of the electricity use profiles of the different household groups is thus important for demand management, consumer support, and evidence-based energy planning.

However, previous studies in Sri Lanka have focused on expenditure on electricity, energy poverty, appliance uptake, response to electricity tariffs, regional household electricity consumption and targeting welfare. These studies provide evidence on the importance of factors such as household size, dwelling characteristics, appliance ownership, cooking energy, and affordability, but there is limited evidence that connects the cross-sections of complete monthly billing histories with detailed household descriptors, smart-meter metadata, controlled location analysis and cumulative-register validation. The LIRNEasia Sri Lankan Residential Electricity Consumption dataset consists of electricity data, high frequency smart-meter electricity data, account characteristics, and three rounds of household surveys [3].

This study aims to create a framework for profiling residential grid-import consumption at the household level. Differences in household size, dwelling characteristics, appliance ownership, air-conditioning, lighting, cooking fuel, solar or export status and meter characteristics are looked at between households as being in the Low, Medium and High consumption groups. Differences in the depots are initially measured descriptively, and then afterward, differences are measured after adjusting for the composition of the house and equipment. A supplementary machine-learning analysis is carried out to investigate whether the household profiles can be separated into the High-consumption households and a separate 6-hour cumulative register is used for validation of the monthly billing result. This contribution is an integrated framework that brings together the concepts of statistical profiling, controlled location analysis, leakage-safe classification and register-based validation in a single Sri Lankan household-energy study. The analysis is not nationally representative, causal, or measures the overall electricity demand of the household, which includes direct behind-the-meter use of solar energy.

The study includes the following research questions (RQ):

RQ1: What differences between the Low, Medium and High groups of residential electricity consumers exist with respect to their household, dwelling, appliance and energy-use characteristics?

RQ2: Do depot-level consumption differences remain after controlling for household and equipment characteristics?

RQ3: Can household survey variables and smart-meter metadata support reliable profiling of high-consumption

households and are monthly billing records consistent with 6-hour cumulative-register evidence?

## II. Related Work

The factors that affect residential electricity use are interacting socio-economic, dwelling, appliance, behavioral, climatic and locational. Reviewing 52 studies from South, East and South-East Asia, Najeeb et al. [4] found 113 determinants reported, such as income, household expenditure, floor area, number of rooms, occupancy, appliance ownership, climate and location. It was also discovered that some socio-demographic variables have context dependent relationships, and that these relationships are not necessarily presumed in other countries.

The most relevant local reference is found in Sri Lankan studies. A multidimensional energy-poverty framework was used by Jayasinghe et al. [2], who identified significant income, socio-demographic and sub-national differences, with lack of modern cooking fuel being one of the key factors contributing to poverty. Kumara and Nimal [1] demonstrated that household scale, urban location, dwelling characteristics and appliance ownership are correlated with electricity expenditure and affordability. Gunewardena and Siyambalapitiya [5] showed the policy relevance of electricity consumption by assessing it as an initial proxy for poverty targeting. These studies prove the significance of affordability and level of consumption, but they do not integrate the household billing histories, detailed survey profiles, location controls and register-based validation.

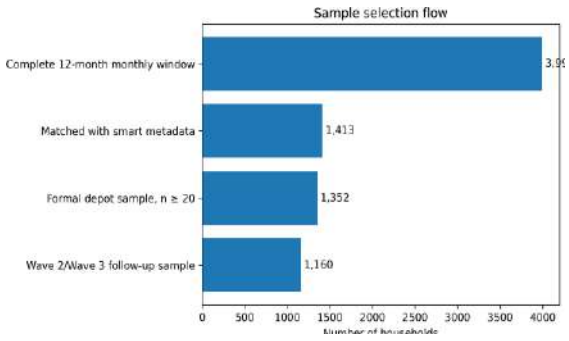
Evidence of appliances and dwellings is also significant. The survey conducted in the Hambantota District revealed that the electricity use rose with the number of family members and house area and was correlated with appliances like air conditioners, refrigerator, kettle, iron, laptops and incandescent lightings [6]. Karunarathna et al. [7] studied the intentions to buy or use energy-efficient appliances in Sri Lanka, demonstrating that the perceived usefulness, ease of use, and affordability of appliances is relevant to adoption, but they did not directly quantify measured electricity consumption. Socio-demographic variables, occupant behavior and appliance characteristics were found to be the most significant relative to general house characteristics in Malaysia, as found by Sena et al. [8]. In Burkina Faso, Tete et al. [9] used a combination of household survey and billing data to conclude that income, tenure, floor area, lighting, televisions, refrigerators, fans, freezers, and air conditioners are key factors. Debebe et al. [10] also demonstrated that cooking-energy choice is influenced by household size, income, education, access to electricity, market access, and household characteristics, which supports the consideration of cooking fuel in more comprehensive household energy profiles.

There is also some evidence that appliance effects vary according to consumption. Park and Yun [11] applied factor analysis and quantile regression to Korean apartment households and revealed that the effects of appliance ownership and use varied across the consumption distribution. Based on data on electricity use in households from 25 countries, de Cian et al. [12] estimated that ownership of an air conditioner contributed 36% to the average increase in electricity use in their household. This evidence can be used to compare consumption groups and not to assume that the household relationship is the same.

Studies using machine-learning approaches can complement traditional explanatory analysis; however, the results must be interpreted carefully. Chidmi et al. [13] used machine-learning and Shapley-based interpretation to identify key predictors of United States residential energy outcomes and identified those predictors to vary across the three energy outcomes: electricity use, total energy use, and expenditure. In low-access environments in Kenya, Fobi et al. [14] were able to use combined electricity bills and satellite imagery to build a typology of household energy use that could be used for planning purposes. The work of Chen et al. [15] proposes backfilling techniques for incomplete smart-meter data and demonstrates the potential use of partially observed high-frequency data for estimating bills and time-of-use profiles, but it is still a preprint. The study on demand response conducted by the Public Utilities Commission of Sri Lanka (PUCSL) is also relevant for the Sri Lankan regulatory context regarding the analysis of household consumption in varying electricity prices [16].

Previous research has focused on electricity expenditure and affordability [1], on energy poverty [2] and on energy consumption associated with appliances [6] in Sri Lanka. Household surveys and machine-learning methods have been applied to analyzing residential energy characteristics across the international literature [9], [13], [14]. The present study is unique in providing matched households with complete annual billing histories, household descriptors, controlled depot analysis, and the High-consumption classification and cumulative-register consistency checks. The contribution is not a new prediction algorithm, it is the integration of complementary analyses.

## III. Methodology




In this study, a quantitative approach on the household level was employed to profile residential grid-import consumption in the matched LIRNEasia sample as in Figure 1. The framework was more descriptive and validation oriented and not causal or nationally representative.

## IV. Sample selection workflow

The consumption outcome was based on monthly billing records and household, dwelling, appliance, account or location characteristics were based on the Wave 1 survey descriptors and smart-meter metadata. The outcome of billing was validated through separately processed six-hour cumulative registers. The grid import for the study is defined as the electricity consumed

by the household from the grid, and not directly from behind-the-meter rooftop solar generation. Depot is the utility service-area that is noted in the smart-meter metadata.

*A. Data Sources and Sample Construction*

This analysis was conducted using monthly billing history, Wave 1 household survey descriptors and smart-meter account metadata. Consumption in the month for billing purposes was the most important grid-import outcome, and the survey and metadata were used to provide household, dwelling, appliance, account, and location characteristics. Six-hour cumulative registers were processed separately and used only for the monthly billing outcome. The monthly file included household_ID, month, and consumption. It initially included 101,575 records. Values that were in comma format were standardized, and 590 truly missing consumption observations were excluded, resulting in 100,985 remaining consumption observations. There were no negative or conflicting household-month values left.

Households were required to have the same 12 consecutive billing months. The search was conducted over all possible windows, and the April 2023–March 2024 window was selected as it would be the one with maximum complete household coverage. This resulted in a panel of 3994 households and 47,928 household-month observations.

Household_ID was used to connect smart-meter metadata. The fields utilized were BRANCH, DEPOT, E_BILLING and EXPORT. Boolean values were normalized prior to aggregation of the households, and an ever-true rule was applied to binary metadata. There were 2,581 households lacking a match for the smart-meter metadata from out of 3,994 households with complete 12-month billing histories. The remaining 1,413 households were matched with Wave 1 profiles and kept for analysis. The data sources and analysis roles are summarized in Table I.

TABLE I. DATA SOURCES AND ANALYTICAL ROLES

| Data Component | Temporal coverage | Sample and analytical role |
|---|---|---|
| Monthly billing records | Raw: October 2022-October 2024; Selected: April 2023-March 2024 | 3994 complete panel households; primary outcome for 1413 matched households |
| Wave 1 household survey | December 2023-October 2024 | Near-contemporaneous household, dwelling, appliance, cooking and meter descriptors |
| Smart-meter metadata | Extracted from six-hour file | Branch, depot, electronic-billing and export indicators for 1413 households |
| Six-hour cumulative registers | January 2023-December 2024 | 4,441,782 valid records across 2950 households; validation over 16,659 household months |

*B. Survey Feature Engineering*

Wave 1 interviews were conducted between December 9, 2023 and October 24, 2024. The variables in the survey were used as near contemporaneous household descriptors and not as perfectly synchronized or causal predictors of this period, since it did not precisely coincide with the April 2023–March 2024 billing window.

Household_ID was used for merging household level files. The appliances, room, fan, light and air-conditioner rosters were combined into one row per household. Household information comprised household size, number of rooms, number of appliances, number of appliance types, number of fans, number of lights, number of air-conditioners, dwelling type, number of electricity meters, survey-reported solar generation and selected meter and dwelling characteristics.

Consolidation followed by the elimination of fully missing or unsuitable variables. The matched analytical sample consisted of 1,413 households with complete outcome and profile data after household-level consolidation and discarding households with missing profile and/or outcome data. Only theoretically relevant variables were chosen for each statistical/machine learning analysis.

*C. Consumption Groups and Statistical Profiling*

The mean monthly grid-import consumption over a 12-month period was used to create tertile groups. This resulted in 472 Low, 470 Medium and 471 High households. Due to the skewness of consumption and appliance variables, Kruskal-Wallis tests were used to compare the numerical characteristics of households. Chi-square tests were used to determine categorical associations, and effect sizes were reported when appropriate. Contextual and sensitivity comparisons were made using PUCSL domestic-consumption bands and solar or export categories. Household-profile tests were performed with Holm adjusted p-values to control for multiple comparisons.

*D. Depot Analysis and Controlled Regression*

Formal depot analysis included only depots containing 20 or more households to reduce the possibility of potentially unstable inferences from small depot sizes. This resulted in 1,352 households in 10 depots. Differences were analyzed by Kruskal-Wallis analysis, which was followed by pairwise Mann-Whitney tests with Holm correction. Nearly zero, export, and survey-reported solar homes were removed from robustness checks.

The outcome variable was the natural log of one plus the mean monthly grid-import consumption. The household-only model was compared with the household-plus-depot model. The reduced heteroskedasticity-consistent type 3 (HC3)-robust model included household, room, meter, appliance, fan, light and AC counts; fan/light/AC use-hours; business, generator, battery, solar, cooking, export and e-billing indicators; and house/meter types. Collinearity control was performed using VIFs. The reduced model excludes appliance-type count due to its high correlation with total appliance count. The exclusion of socio-economic class was due to the lack of external validation of coding matrix.

*E. High-Consumption Machine Learning Profiling*

High/Not-High classification used 16 predictors: household, room, appliance, appliance-type, fan, light, AC and meter counts; AC, business, backup-generator and grid/gas/firewood-

cooking indicators; and house and meter types. Numerical predictors used median imputation (and RobustScaler for Logistic Regression); binary/categorical variables used mode imputation, with categorical one-hot encoding. Five outer and four inner folds, eight tuning configurations per model and inner ROC-AUC selection were compared between Logistic Regression, Random Forest and Gradient Boosting. Training fold upper tertile cutoffs were used for high labels. Calibration (Brier score) and classification thresholds (balanced accuracy) used training out-of-fold predictions. One prediction and fold-specific label were given for each household and used for pooled evaluation. Evaluation methods used were ROC AUC, PR AUC, balanced accuracy, precision, recall, specificity, F1 score and Brier score. Feature-permutation importance, ablation, zero-consumption exclusion and alternative High definitions were used for robustness checks.

The following statistics were used for evaluation: receiver operating characteristic-area under the curve (ROC-AUC), area under the precision-recall curve (PR-AUC), balanced accuracy, precision, recall, specificity, F1-score, calibration, and bootstrap confidence intervals. Robustness checks included feature ablation testing, permutation feature importance, alternative High-consumption thresholds, zero-consumption exclusion, and repeated nested cross-validation that compares 16 features with 15 features without meter type.

### *F. Six-Hour Cumulative-Register Validation*

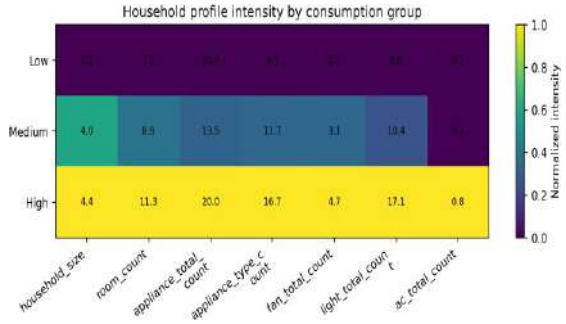
There were 4,441,782 valid, timestamped records in five 6-hour files across 2,950 households. The value of TOTAL_IMPORT was a cumulative register, not as a direct interval consumption. Records were sorted and differenced individually from household by household. Negative deltas, time gaps that were less than or equal to zero, and physically unreasonable positive changes of more than 24 kWh/h were omitted.

No calendar splitting was done on valid same month intervals. Cross-month intervals that were adjacent to each other, and had not exceeded 24 hours, were distributed proportionally across the month, while all other cross-month intervals were excluded. The monthly reconstruction was compared with the billing consumption with Pearson and Spearman correlation as well as by means of median absolute error and median absolute percentage error. There was no extra maximum-gap or minimum monthly-coverage minimum for valid same-month intervals. The 24/48-hour gap limits were tested separately, and percentage errors were not calculated for months where no bills were issued. The six hours of data were not used as the primary modelling predictors but were used for validation and secondary checks.

## V. Results and Discussion

### *A. Consumption Groups and Household Profiles*

The final matched sample consisted of 1413 households: 472 Low, 470 Medium, and 471 High consumption households. Average monthly grid-import consumption was 50.19, 103.27 and 228.16 kWh for the three groups, respectively. The tertiles corresponded closely with the PUCSL consumption bands: 466 of the 472 households in the Low-consumption tertile had a dominant band of 90 kWh or less, and 466 of the 471 households in the High-consumption tertile had a dominant band of at least 121 kWh. Cell values show group means and color intensity is normalized separately for each characteristic




## VI. Selected Household and Appliance Characteristics Across Low, Medium and High Consumption Groups

The Kruskal-Wallis tests showed significant differences in household size, room count, appliance count, appliance-type count, fan count, light count and air-conditioner count following Holm correction. The largest numerical effects were observed for appliance count ($\varepsilon^2 = 0.214$), fan count ($\varepsilon^2 = 0.213$), light count ($\varepsilon^2 = 0.198$), and appliance-type count ($\varepsilon^2 = 0.195$). These results corroborate the Sri Lankan and international studies that have found the relationship between household scale, dwelling characteristics, and appliance ownership with residential electricity consumption [6]–[9].

The strongest categorical differences were in the ownership of an air-conditioner: 6.1% of Low, 7.7% of Medium and 39.1% of High consumption households report air-conditioning (Cramér's V = 0.398). This is in line with the international evidence that indicates that there is a strong correlation between the presence of an air conditioner and the demand for electricity in the home [12].

The distribution of dwelling types was also significantly different between the groups (Cramér's V = 0.200), with more double floor houses in the High group and more single floor houses in the Low group. The amounts of firewood cooking by Low, Medium and High groups were 33.5%, 21.3% and 9.3% respectively, showing a negative trend between these groups (Cramér's V = 0.240). These patterns are not specific effects, but patterns of household-energy configurations [2], [10].

The distribution of survey-reported solar generation and export status was also skewed towards High-consumption households, with about 31.2 % of the High group reporting generation, and 5.1% reporting export, while 3.2 % of the Low group reported export and 0.3 % reported generation. These associations should not be causal as this may reflect differences among participating households in appliances and dwelling characteristics and direct solar self-consumption is not captured in the grid-import records.

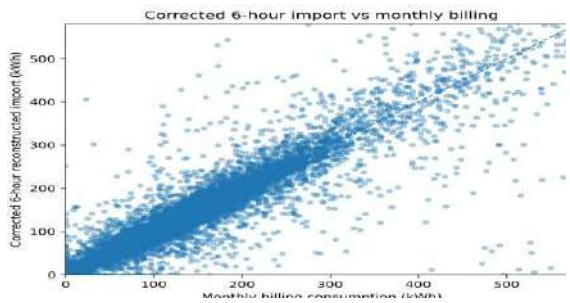
In general, RQ1 reveals that consumption is correlated with a household's combined profile, which includes the size of the household, the features of the dwelling, appliance intensity, cooling equipment, cooking fuel and the solar/export context, instead of one single determinant [4].

*A. Depot Differences After Household Controls*

A total of 1352 households were retained in the formal analysis in 10 depots. The distribution of household mean monthly grid-import consumption was significantly different across depots ($H = 133.35$, $p = 2.43 \times 10^{-24}$, $\varepsilon^2 = 0.093$). The highest descriptive means were recorded in Boralesgamuwa (mean 234.99 kWh) and Nugegoda depot (mean 229.24 kWh) while the lowest was recorded from Moratuwa South C.S.C (mean 101.60 kWh). After Holm correction 16 pairwise depot contrasts were still significant.

This unadjusted depot difference remained when near zero consumers were excluded. Differences were also statistically significant when exporters and survey-reported solar households were removed, but the values of $\varepsilon^2$ were reduced to about 0.030 and 0.033, respectively. The smaller effect sizes suggest a contribution to the raw depot contrast from solar and export composition.

A more conservative outcome was obtained when it was performed in a controlled way. The adjusted $R^2$ value for the reduced household-only model was 0.2571, with an $R^2$ value of 0.2703. The addition of nine depot indicators raised $R^2$ to 0.2756 and adjusted $R^2$ to 0.2574, which represented a change of $\Delta R^2 = 0.0053$. The depot indicators were not jointly significant under the HC3-robust Wald test, $\chi^2(9) = 12.32$, $p = 0.196$.




Household size (β=0.086), AC count (β=0.140), export status (β=0.366) and appliance count (β=0.013) showed positive associations while negative associations were seen for meter count (β=−0.171) and firewood cooking (β=−0.143). These HC3-robust coefficients measure associations with log(1 + mean monthly kWh), rather than causal effects.

As a result, RQ2 would differentiate between descriptive and adjusted geographic variation. Differences between depots were found when comparisons were made, although much less explanatory information was added after household and equipment characteristics were added in. The service-area variation after the adjustment is consistent with the part played by household composition, but does not establish an independent depot effect.

*B. High-Consumption Machine Learning Profiling*

There were 471 households in the 'High consumption' group and 942 households in the 'Not High' group. To avoid target leakage, the High-consumption cut-off has been re-evaluated in each training fold during nested validation. In the final 16-predictor core analysis, Logistic Regression was selected in all five outer folds using inner-fold ROC-AUC, which suggest a consistent preference for the simpler model compared to the models tested that were based on trees.

Using pooled out-of-fold predictions and a threshold selected from training data only, the core model achieved a ROC-AUC of 0.8207 (bootstrap 95% confidence interval: 0.7962–0.8434), PR-AUC of 0.7194, balanced accuracy of 0.7404, precision of 0.5901, recall of 0.7367, specificity of 0.7442, F1-score of 0.6553, and Brier score of 0.1561.

The findings are consistent with the first part of RQ3, which focused on the household survey descriptors and smart-meter metadata that are reproducible to distinguish between High- and Not-High consumption profile. This finding aligns with Chidmi et al. [13] who found that there is potential to use household and dwelling characteristics to assist in machine-learning based analysis of residential energy outcomes, and with Fobi et al. [14] who found the usefulness of classification of household-level energy consumption in a data-constrained environment.

Cross-validated permutation importance derived fan count and household size as the most obvious contributors and then air-conditioner and appliance counts with some uncertainty. Repeated nested validation gave ROC-AUC values 0.8218 with 16 predictors and 0.8216 without meter type (15 predictors).

*C. Six-Hour Register Validation*

A total of 4,441,782 valid timestamped records for 2,950 households were analyzed from six-hourly records. Monthly estimates for the grid-imports were reconstructed by applying cumulative-register differencing at the household level, physical filtering and calendar-boundary allocation. For the final sample of 1,413 households, 16,659 matched household-months were covered by validation.

The degree of agreement was high: Pearson correlation = 0.929997; Spearman correlation = 0.948151; median absolute error = 7.29 kWh; median absolute percentage error = 7.75%. With 48- and 24-hour gap limits, Pearson r was 0.9289 and 0.9246, respectively, with median percentage errors of 8.06% and 9.27%. The results are used to address the validation aspect of RQ3 and suggest that the monthly billing records reflect the ranking of the houses and overall grid-import amplitude when compared to the register estimates processed separately.

Fig. 1. Reconstructed monthly grid import from six-hour records compared with monthly billing consumption in the matched sample

Cumulative readings demanded a procedure of differencing between households and calendar-boundary allocation. Billing and register readings measure the same grid imports, therefore they are only a consistency check and not an external validation. A limitation of irregular gaps still exists [15].

The six-hour records also provided some additional time-of-use and export checks. High consumption households had a relatively higher off-peak proportion and a relatively lower daytime proportion, whereas the proportion of households with both solar and export indicators had the highest export-to-import ratios. These results are secondary, as they need allocation assumptions and grid-import registers are not able to measure all behind-the-meter solar consumption.

### *D. Overall Interpretation and Limitations*

TABLE II
PROVIDES THE SUMMARY OF THE MAIN FINDINGS OF THIS STUDY.

| Component | Final Finding |
|---|---|
| Analytical population | 1,413 matched households with complete 12-month billing records |
| Consumption groups | Low 472, Medium 470, High 471 |
| RQ1 | Largest profile differences involved appliance, fan, light and AC characteristics |
| RQ2 | Unadjusted epsilon-squared = 0.093; adjusted depot increment in R-squared = 0.0053; joint p = 0.196 |
| RQ3 profiling | Nested out-of-fold ROC-AUC 0.8207 and PR-AUC 0.7194 |
| RQ3 validation | Pearson correlation = 0.930, Spearman = 0.948, median absolute percentage error = 7.75% |
| Main interpretation | Household and equipment characteristics were more informative than depot membership after adjustment |

Overall, residential grid-import consumption was more closely related to house and equipment features than to depot membership after adjustment. The uniformity of the descriptive profiling, controlled regression, machine learning and register validation reinforces confidence in the fact that the main results are not specific to an individual analytical method.

There are limitations that set the scope of interpretation. The sample matched is not representative of the Sri Lankan household level. Variables for Wave 1 did not perfectly align with the billing window and extended beyond it, meaning that Wave 1 was not fully simultaneous with the billing window. The study is descriptive and does not allow for conclusions about causality. When billing and six-hour registers are used on a monthly basis, they record the electricity that is imported from the grid, but not the electricity that is directly used on-site, behind the meter, from the rooftop solar power. Irregular gaps in the six-hour recordings exist, and the appliance efficiency, the actual operating time, occupancy patterns, weather exposure, household resources and behavioral preferences are not adequately represented.

The importance of this study is that it offers an integrated and empirically supported framework to identify household characteristics that are linked to varying degrees of residential grid import in the matched population of Sri Lanka. The findings can support demand-management research, consumer segmentation and household-focused energy-planning strategies without emphasizing primarily on broad service-location categories.

## VII. CONCLUSION

This study proposed a framework at the household level to profile residential grid-import consumption in Sri Lanka using comprehensive household-level monthly consumption bills, Wave 1 household descriptors, smart-meter metadata, controlled depot analysis, nested cross-validated classification and cumulative register validation for six hours. The final matched sample consisted of 1,413 households that were observed from April 2023 to March 2024.

Households with high consumption were found to have larger household and dwelling profiles, higher appliance intensity, higher counts of fans and lighting, and much higher air-conditioner ownership. Consumption distributions varied by depot in unadjusted analysis but resulted in an insignificant contribution to $R^2$ (0.0053) and were not significant as a set after household and equipment controls. The classification analysis resulted in reproducible household-profile signal, and the cumulative-register reconstruction was largely in agreement with the monthly billing records.

The importance of the study is that it offers an integrated and empirically backed framework for determining the characteristics of households associated with varying levels of residential grid import in the matched Sri Lankan sample. The results can be used for exploratory consumer segmentation and household-based demand-management research and confirm that a general depot location provides little extra information once household composition is taken into account.

The results remain descriptive. The sample is not nationally representative, the timing of the surveys was not exactly aligned with the billing period, and the grid-import records do not contain direct rooftop solar self-consumption. Future studies should incorporate larger probability-based samples, synchronized longitudinal survey data, direct solar measurements and weather controls as well as external validation in other parts of Sri Lanka.

## REFERENCES


[1] A. S. Kumara and P. D. Nimal, "Drivers of household electricity Expenditure: Insights from the highest electricity tariff-paying nation in South Asia," *Energy Build.*, vol. 334, 2025, doi: 10.1016/j.enbuild.2025.115493.

[2] M. Jayasinghe, E. A. Selvanathan, and S. Selvanathan, "Energy poverty in Sri Lanka," *Energy Econ.*, vol. 101, 2021, doi: 10.1016/j.eneco.2021.105450.

[3] LIRNEasia, "Sri Lankan Residential Electricity Consumption Dataset," 2024, *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/lirneasia/sri-lankan-residential-electricity-consumption

[4] A. Najeeb, S. Sridharan, A. B. Rao, S. B. Agnihotri, and V. Mishra, "Determinants of residential electricity consumption in South, East and South East Asia: A systematic review," *Renewable and Sustainable Energy Reviews*, vol. 198, 2024, doi: 10.1016/j.rser.2024.114400.

[5] D. Gunewardena and S. Siyambalapitiya, "Electricity Usage as A Proxy Indicator For Poverty Targeting," *Sri Lanka Journal of Economic Research*, vol. 10, no. 2, pp. 115–125, 2023, doi: 10.4038/sljer.v10i2.186.

[6] C. E. Kankanamge and P. S. A. D. Gunasekara, "Trends in appliance usage and electricity consumption in households: A case study in Hambantota District, Sri Lanka," *Ruhuna Journal of Science*, vol. 13, no. 1, pp. 29–40, 2022, doi: 10.4038/rjs.v13i1.113.

[7] W. K. S. Karunarathna, W. Jayaratne, S. Dasanayaka, S. Ibrahim, and F. Samara, "Factors affecting household's use of energy-saving appliances in Sri Lanka: an empirical study using a conceptualized technology acceptance model," *Energy Effic.*, vol. 16, no. 3, 2023, doi: 10.1007/s12053-023-10096-7.

[8] B. Sena *et al.*, "Determinant factors of electricity consumption for a Malaysian household based on a field survey," *Sustainability (Switzerland)*, vol. 13, no. 2, pp. 1–31, 2021, doi: 10.3390/su13020818.

[9] K. H. S. Tete, Y. M. Soro, S. F. A. Coulibaly, R. V. Jones, and S. dit S. Sidibé, "Determinants of urban residential electricity consumption in Burkina Faso: A study of dwelling, household, socio-economic and

appliance-related factors," *Buildings*, vol. 14, no. 3, 2024, doi: 10.3390/buildings14030683.

[10] B. Debebe, F. Senbeta, D. Diriba, E. Teferi, and D. Teketay, "Determinants of household energy choice for domestic chores: Evidence from the Semien Mountains National Park and Adjacent Districts, Northwest Ethiopia," *Cleaner Energy Systems*, vol. 4, 2023, doi: 10.1016/j.cles.2023.100063.

[11] J. Park and S. J. Yun, "Social comparison of home appliance ownership and use and heterogeneous impact on electricity consumption: a case of apartment households in Korea," *Energy Effic.*, vol. 17, no. 5, 2024, doi: 10.1007/s12053-024-10219-8.

[12] E. De Cian, G. Falchetta, F. Pavanello, Y. Romitti, and I. Sue Wing, "The impact of air conditioning on residential electricity consumption across world countries," *J. Environ. Econ. Manage.*, vol. 131, 2025, doi: 10.1016/j.jeem.2025.103122.

[13] B. Chidmi, Y. Ettoumi, and E. S. Aljohani, "Analyzing the determinants of U.S. residential energy usage and spending: A machine learning approach," *The Energy Journal*, vol. 46, no. 1, pp. 147–177, 2025, doi: 10.1177/01956574241289375.

[14] S. Fobi, J. Mugyenyi, N. J. Williams, V. Modi, and J. Taneja, "Predicting levels of household electricity consumption in low-access settings," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2213–2222. doi: 10.1109/WACV51458.2022.00227.

[15] X. Chen, S. Cai, and A. F. Smeaton, "Back-filling missing data when predicting domestic electricity consumption from smart meter data," 2024. [Online]. Available: https://arxiv.org/abs/2412.03574

[16] Public Utilities Commission of Sri Lanka, "Study on electricity demand response to price changes," Colombo, 2023. [Online]. Available: https://www.pucsl.gov.lk/wp-content/uploads/2025/04/Final-Report-Study-on-electricity-demand-response-to-price-changes.pdf

# The Impact of Digital Literacy on the Effective Use of E-Learning Platforms Among Undergraduate Students in Sri Lanka.

Vimansa Gamage
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama,Sri Lanka*
kgvdevanmini@students.nsbm.ac.lk

D. Rajasinghe
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama,Sri Lanka*
dharani.r@nsbm.ac.lk

Dewmini Silva
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama,Sri Lanka*
gdssilva@students.nsbm.ac.lk

Rasheena Sewwandi
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama,Sri Lanka*
wlrsewwandi@students.nsbm.ac.lk

Imalsha Sithumi
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama,Sri Lanka Homagama,Sri Lanka*
owhisithumi@students.nsbm.ac.lk

***Abstract*—Digital literacy is one of the most important factors that influence the success of undergraduate students in the rapidly growing e-learning environment in Sri Lanka. Despite the investment in learning management systems (LMS) and virtual classrooms, previous research indicates that the outcomes of students are more related to their capability to use these tools purposefully than to their availability. The present study investigates the association between digital literacy (technical, cognitive and social-emotional) on the effective use of e-learning platforms by undergraduate students in Sri Lanka. A quantitative, cross-sectional survey design was used, and data was collected from 204 undergraduate students from different universities in Sri Lanka through a structured questionnaire that was developed based on the DigComp framework. Descriptive statistics, Pearson correlation and multiple linear regression were used in SPSS to analyse the data. The results indicate that digital literacy is a significant and positive predictor of the effective use of the e-learning platform, explaining about 38.6% of the variance ($R^2 = .386$, $F(3,200) = 41.86$, $p < .001$). Technical digital literacy was the most significant predictor ($\beta = .35$, $p < .001$), followed by cognitive digital literacy ($\beta = .24$, $p < .001$) and social-emotional digital literacy ($\beta = .18$, $p = .030$). The results suggest that although operational ICT skills remain the strongest predictor, critical-thinking and online-collaboration skills also show significant positive associations with effective e-learning engagement. The cross-sectional, self-reported design indicates associations rather than causal effects. The study suggests that higher education institutions in Sri Lanka should incorporate structured digital literacy courses into the undergraduate curriculum, focusing on information evaluation and online collaboration skills, to improve the effectiveness of e-learning and mitigate digital inequality among students.**

***Keywords*—DigComp Framework, Digital Literacy, E-Learning Platforms, Sri Lanka, Undergraduate Students***

## I. INTRODUCTION

Over the last five years, Sri Lankan higher education institutions have experienced a dramatic and largely

irreversible transformation towards digitally mediated learning. The COVID-19 pandemic has transformed this emergency response into a permanent hybrid model, where Learning Management Systems (LMS) like Moodle, alongside platforms like Zoom, Microsoft Teams and WhatsApp, are now integrated into the daily academic routine of both public and private universities [1].

Even with this massive deployment of e-learning infrastructure, institutional investment in platforms has not necessarily led to effective use by students. The national evidence gathered during the pandemic period revealed that less than 40% of Sri Lankan undergraduates had received any formal training in the use of LMS or other e-learning tools, and prior knowledge and experience of ICT were among the most significant statistical factors that influenced a student's willingness and ability to use e-learning [1]. The same is true for medical undergraduates in Colombo, who are still relying on traditional study methods despite the availability of digital resources, suggesting that the challenge is not access but competence and confidence in using digital tools purposefully [2].

Digital literacy, defined as the capacity to find, assess, produce and share information through digital technologies, has thus become a key concept in the study of how some students are able to gain more academic benefit from e-learning platforms than others [3]. Ng's influential model defines digital literacy in three ways: technical, cognitive, and social-emotional [3]. This multi-dimensional perspective is reflected in the European Commission's DigComp 2.2 framework, which has

been adapted and validated in various non-Western higher education settings [4], [5].

Although digital literacy and e-learning adoption have been studied separately, there is comparatively little research in Sri Lanka that has investigated the relationship between the different dimensions of digital literacy and the effective use of e-learning platforms among the general undergraduate population, rather than acceptance during a crisis period or within a single faculty like medicine. This study fills that void.

Based on this, the following research objectives are set for this paper: (i) to explore the level of digital literacy and effective use of e-learning platforms among Sri Lankan undergraduates; (ii) to investigate the relationship between the technical, cognitive, and social-emotional aspects of digital literacy and the effective use of e-learning platforms; and (iii) to determine which dimension of digital literacy is most strongly associated with the effective use of e-learning platforms. The rest of the paper is organized as follows: Section II reviews the relevant literature, Section III presents the research gap and problem statement, Section IV describes the conceptual framework and hypotheses, Section V outlines the methodology, Section VI presents the data analysis, and Section VII presents the discussion and limitations. Section VIII offers recommendations and Section IX offers concluding remarks.

## II. LITERATURE REVIEW

### *A. The Concept of Digital Literacy*

Ng's seminal research on undergraduate digital nativeness suggested that digital literacy should be considered as the intersection of three literacies: technical, cognitive, and social-emotional [3]. The technical dimension is the operational skills required to use hardware, software and online systems effectively; the cognitive dimension is the capacity to search, critically evaluate and synthesize digital information while being mindful of ethical and legal issues; and the social-emotional dimension is the capacity to communicate, collaborate and behave appropriately in online environments [3]. This tri-dimensional structure has since been adopted and empirically validated in various fields, from nursing education to teacher training, and has been found to be valid in all three dimensions, as they all make an independent contribution to overall digital competence.

The European Commission's Joint Research Centre has created a parallel and complementary framework, DigComp 2.2, which breaks down digital competence into five areas: information and data literacy, communication and collaboration, digital content creation, safety, and problem solving, and has been tested in higher education settings as far afield as Brunei and Uganda, proving that the framework travels well beyond its European roots [4], [5]. Both approaches share the perspective that digital literacy is a multi-dimensional phenomenon that goes beyond the mere use of computers.

### B. E-Learning Platforms and Effective Use

E-learning platforms are digital systems that facilitate the delivery, access and evaluation of academic content away from the physical classroom, such as LMS, video-conferencing tools and cloud-based collaboration software. The Technology Acceptance Model (TAM) has long been the focus of research into the reasons why students use and continue to use such systems, claiming that perceived usefulness and perceived ease of use are the most important factors in determining behavioural intention to use technology [6]. A systematic review of 71 studies in educational contexts that examined TAM found that self-efficacy, facilitating conditions, and system accessibility were always among the strongest predictors of successful technology adoption in learning contexts [7].

But the intention to adopt is conceptually different from the actual use. A student might want to use an e-learning platform, and even log in to it regularly, but not get any academic benefit from it, such as passively watching recorded lectures without critically engaging with the content or joining online discussions. This difference is the reason for the present study's focus on effective use, which is defined as the extent to which a student purposefully and competently utilizes the features of the e-learning platform (content access, assessment tools, communication channels) to achieve learning outcomes.

### C. Digital Literacy as a Predictor of E-Learning Outcomes

There is increasing evidence that digital literacy is directly tied to learning outcomes. The results of a large-scale study conducted in Indonesia and Bangladesh showed that digital informal learning engagement, self-efficacy, and digital competence partially mediated the relationship between digital literacy and academic performance, thus indicating that digital literacy has both direct and indirect effects on students' performance in digitally mediated learning environments [8]. Likewise, a large sample of Chinese university students revealed that digital competence was positively correlated with information and data literacy, communication and collaboration, and safety, but there were still significant differences in problem-solving and content-creation competencies, which are more advanced and cognitively demanding aspects of digital competence [9]. The results indicate that operational digital skills are more likely to be acquired than higher-order cognitive skills that enable the critical and productive use of digital tools.

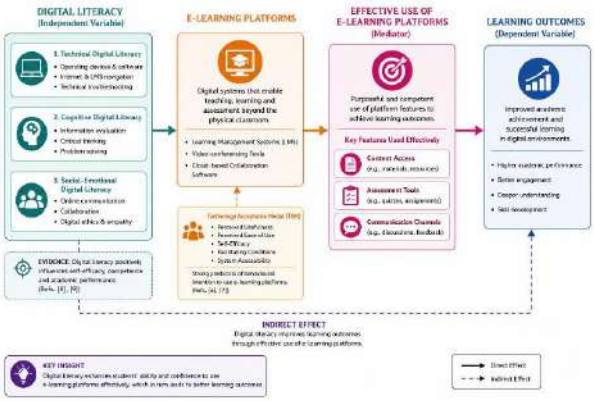



Fig. 1. Digital literacy and e-learning outcomes.

### *B. The Sri Lankan and Developing-Country Context*

In Sri Lanka, a cross-sectional study of 657 undergraduate students from three state universities revealed that only 36.4% had been formally trained in the use of LMS, and that ICT knowledge and prior training were statistically significant factors in e-learning acceptance, along with residential locality

and family income [1]. A similar survey of medical undergraduates and new graduates also revealed that although all students had access to internet-enabled devices, they still preferred traditional learning methods, with preferences differing by year of study and family income [2]. These results all suggest that there is a skills and confidence deficit that cannot be addressed by infrastructure alone.

This is in line with the wider digital divide literature, which has progressed from a first-level divide in access to a second-level divide in digital skills and a third-level divide in the tangible benefits of technology use. In a systematic review of the digital divide in higher education in developing countries, limited digital literacy, along with the affordability of devices and infrastructure, was identified as one of the main factors contributing to educational inequity, and that this gap disproportionately affects students from lower income and rural backgrounds [10]. The literature reviewed indicates that digital literacy, not just the availability of platforms, is a key factor in enhancing the effectiveness of e-learning in resource-poor settings like Sri Lanka in developing countries.

## III. RESEARCH GAP AND PROBLEM STATEMENT

The literature reviewed reveals three gaps. First, most of the digital-literacy-and-e-learning studies in Sri Lanka were conducted during the COVID-19 emergency period and were limited to acceptance or willingness to adopt e-learning, not on the effectiveness of use in a now-stabilized, post-pandemic hybrid learning environment [1], [2]. Second, previous research in Sri Lanka has focused on individual faculties, particularly medicine, and the results are not necessarily applicable to the broader undergraduate student body in other faculties such as business, computing, engineering, and others. Third, although digital literacy has been examined as a unidimensional or loosely defined construct in many of the local studies, international studies have increasingly demonstrated that the technical, cognitive and social-emotional aspects of digital literacy have different and sometimes conflicting relationships with learning outcomes [3], [8], [9], which has not been explored extensively among Sri Lankan undergraduates.

This poses a practical challenge for universities and policy makers: if they do not know which aspect of digital literacy is most predictive of effective use of e-learning, they may be designing and targeting their training interventions in a generic way. Therefore, the problem of this study is: How and in which aspects is digital literacy associated with the effective use of e-learning platforms among undergraduate students in Sri Lanka?

## IV. CONCEPTUAL FRAMEWORK AND HYPOTHESES

Based on Ng's three-dimensional digital literacy model [3] and the DigComp framework [4], this study suggests a conceptual framework in which the technical digital literacy, cognitive digital literacy, and social-emotional digital literacy are independent variables that are hypothesised to predict the effective use of e-learning platforms, which is the dependent variable, as shown in Figure 2. Effective use is defined as the extent to which a student purposefully and competently uses the content-access, assessment and communication features of an e-learning platform to achieve learning outcomes (Fig. 1). Conceptually, technical literacy is expected to underpin operating the platform's content and assessment features, cognitive literacy the critical and purposeful engagement with content, and social-emotional literacy the use of communication channels (e.g., discussions, feedback).

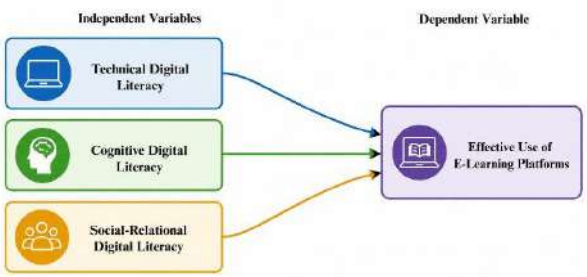



*Fig. 2. Conceptual Framework of the Study.*

Based on this framework, the following hypotheses are proposed:

H1: Technical digital literacy is a significant positive predictor of the effective use of e-learning platforms among undergraduate students in Sri Lanka.

H2: Cognitive digital literacy is a significant positive predictor of the effective use of e-learning platforms among undergraduate students in Sri Lanka.

H3: The effective use of e-learning platforms among undergraduate students in Sri Lanka is significantly positively predicted by social-emotional digital literacy.

H4: Overall digital literacy is a significant positive predictor of the effective use of e-learning platforms among undergraduate students in Sri Lanka.

## V. METHODOLOGY

### *A. Research Design*

The study was of a quantitative, cross sectional survey type with a deductive, positivist research approach. This design was chosen because the study aimed to test hypothesized relationships between pre-defined constructs (digital literacy dimensions and effective e-learning use) across a large undergraduate population at one time, similar to previous studies in this area based on the TAM and DigComp [4], [5].

### *B. Population and Sample*

The target population was undergraduate students who are actively using e-learning platforms as part of their studies in Sri Lankan higher education institutions. The final sample consisted of 204 undergraduate students who were collected using convenience sampling and snowball sampling in several state and private universities through a self-administered online questionnaire. The achieved sample size is greater than the minimum sample size recommended by Krejcie and Morgan's sample size table for populations of this size and is comparable to sample sizes used in similar regional studies [5], [9]. Because non-probability sampling was used, representativeness is limited (Section VII-B).

### C. Research Instrument

A structured questionnaire with three sections was used to collect data. Section A collected demographic data (gender, year of study, field of study, and device ownership). The three dimensions of digital literacy (technical, cognitive, and social-emotional) were assessed in Section B with items adapted from Ng's digital literacy model [3] and the DigComp 2.2 framework [4]. Section C assessed the effective use of e-learning platforms through six items covering three facets of use: purposeful engagement with LMS content, assessment tools, and online communication features; the item scores were combined into a single composite score (mean of items). The items of all constructs were rated on a 5-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree). The full questionnaire is provided as supplementary material.

### D. Validity and Reliability

Content validity was achieved by two academic supervisors who are familiar with the Dig Comp framework and a pilot test with 20 respondents whose data were not used in the final analysis. Internal consistency reliability was determined by Cronbach's alpha for all constructs, and all constructs had an alpha coefficient greater than .70, which is the standard for acceptable internal consistency reliability (see Section VI).

### E. Data Analysis Techniques

Data were analyzed using IBM SPSS. The respondents and constructs were profiled using descriptive statistics (means and standard deviations), bivariate relationships between digital literacy dimensions and effective e-learning use were examined using Pearson correlation, and the hypothesized predictive relationships (H1–H4) were tested using multiple linear regression. Multicollinearity was checked with the variance inflation factor (VIF) and tolerance, and 95% confidence intervals and Cohen's $f^2$ are reported as effect-size information. Gender, year of study and field of study were collected only to profile the sample; because the hypotheses concern the three digital literacy dimensions, they were not entered as control variables (see Section VII-B).

### F. Ethical Considerations

The participation was voluntary and anonymous. An informed consent statement was provided at the beginning of the questionnaire and no personally identifiable information was gathered. Data were used only for academic research. The study followed the research ethics guidelines of the authors' institution for anonymous, minimal-risk surveys.

## VI. DATA ANALYSIS AND RESULTS

### A. Demographic Profile of Respondents

A total of 204 valid responses were received. The sample was reasonably balanced across gender and year of study as indicated in Table I, with the highest number of respondents studying computing/IT related programmes.

TABLE I. DEMOGRAPHIC PROFILE OF RESPONDENTS

| Variable | Category | n | % |
|---|---|---|---|
| Gender | Male | 86 | 42.2% |
| | Female | 118 | 57.8% |
| Year of Study | Year 1 | 44 | 21.6% |
| | Year 2 | 57 | 27.9% |
| | Year 3 | 55 | 27.0% |
| | Year 4 | 48 | 23.5% |
| Field of Study | Computing / IT | 82 | 40.2% |
| | Business / Management | 71 | 34.8% |
| | Engineering | 31 | 15.2% |
| | Other | 20 | 9.8% |

### B. Reliability Analysis

All constructs had Cronbach's alpha values above .70 (Table II), indicating acceptable internal consistency for the digital literacy and effective-use scales employed in this study.

TABLE I. RELIABILITY STATISTICS

| Construct | No. of Items | Cronbach's α |
|---|---|---|
| Technical Digital Literacy | 5 | .812 |
| Cognitive Digital Literacy | 5 | .784 |
| Social-Emotional Digital Literacy | 4 | .763 |
| Effective Use of E-Learning Platforms | 6 | .841 |
| Overall Instrument | 20 | .874 |

### C. Descriptive Statistics

The mean and standard deviation for each construct are shown in Table III. The mean scores of the respondents indicated that they were most confident in social-emotional digital literacy ($M = 3.71$, $SD = 0.55$) and least confident in cognitive digital literacy ($M = 3.45$, $SD = 0.61$), which means that Sri Lankan undergraduates were more confident in their online communication behaviour than in their critical evaluation of digital information.

TABLE II. DESCRIPTIVE STATISTICS

| Construct | Mean | SD |
|---|---|---|
| Technical Digital Literacy | 3.62 | 0.58 |
| Cognitive Digital Literacy | 3.45 | 0.61 |
| Social-Emotional Digital Literacy | 3.71 | 0.55 |
| Effective Use of E-Learning Platforms | 3.58 | 0.62 |

### D. Correlation Analysis

Pearson correlation analysis (Table IV) showed that all three digital literacy dimensions were significantly and positively correlated with effective use of e-learning platforms at the $p < .01$ level. The bivariate correlation between technical digital literacy and effective use was the highest ($r = .517$, $p < .01$), followed by cognitive digital literacy ($r = .468$, $p < .01$) and social-emotional digital literacy ($r = .392$, $p < .01$).

TABLE III. PEARSON CORRELATION MATRIX (**p<.01)

| Variable | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Technical DL | 1 | | | |
| 2. Cognitive DL | .441** | 1 | | |
| 3. Social-Emotional DL | .328** | .389** | 1 | |
| 4. Effective Use | .517** | .468** | .392** | 1 |

*E. Regression Analysis and Hypotheses Testing*

A multiple linear regression was performed using technical, cognitive and social-emotional digital literacy as predictors of the effective use of the e-learning platform. The overall model was statistically significant, $F(3, 200) = 41.86$, $p < .001$, and explained 38.6% of the variance in effective use ($R^2 = .386$, Adjusted $R^2 = .377$). Table V shows that all three predictors contributed significantly and independently to the model, thus supporting H1, H2, and H3. The overall model was significant, and all three dimensions were positive, so H4 is also supported. Multicollinearity was not a concern: VIF values ranged from 1.22 to 1.35 (tolerance = .74–.82), far below the common cut-offs of 5 or 10, consistent with the moderate inter-correlations among the predictors (r = .33–.44, Table IV). The 95% confidence intervals of all three coefficients exclude zero, but that of social-emotional literacy is wide and its lower bound is close to zero, so this coefficient should be interpreted cautiously. The effect size of the model was large (Cohen's $f^2 = .63$).

TABLE IV. MULTIPLE REGRESSION RESULTS

| Predictor | β | 95% CI | t | p | VIF | Hypothesis |
|---|---|---|---|---|---|---|
| Technical Digital Literacy | .35 | [.21, .47] | 5.21 | < .001 | 1.29 | Supported |
| Cognitive Digital Literacy | .24 | [.14, .40] | 4.02 | < .001 | 1.35 | Supported |
| Social-Emotional Digital Literacy | .18 | [.01, .29] | 2.18 | .030 | 1.22 | Supported |

Model summary: $R = .621$, $R^2 = .386$, Adjusted $R^2 = .377$, $F(3, 200) = 41.86$, $p < .001$.Dependent variable: effective use; CI = 95% confidence interval for β; VIF = variance inflation factor.

## VII. DISCUSSION

*A. Interpretation of Findings*

The findings support the hypothesis that digital literacy is a statistically significant and practically meaningful predictor of the effective use of e-learning platforms by Sri Lankan undergraduates, which aligns with the international evidence that digital literacy is associated with digitally mediated learning outcomes [8], [9]. All three hypothesized dimensions (technical, cognitive, and social-emotional) were independently significant predictors, which supported H1 through H4 and further validated Ng's argument that digital literacy should be considered as a multidimensional construct instead of a single operational skill [3].

Prior ICT knowledge and training were found to be the strongest predictors in the Sri Lankan COVID-era literature [1], [2] and were also found to be the strongest predictor in this study (β = .35). This is in line with the literature on the digital divide, which indicates that in the context of developing countries, basic operational skills are the main constraint to the effective use of technology, rather than higher-order skills [10]. That is, many Sri Lankan undergraduates need to be comfortable with the platforms before they can critically evaluate or collaborate around digital content.

The second strongest predictor was cognitive digital literacy (β = .24), which is defined as the capacity to locate, assess, and combine academic information from the internet. This result is consistent with the international evidence that the competencies of cognition and content creation are less developed than more basic technical and communication skills among university students [9] and therefore an area where curricula can contribute beyond just enhancing device access.

Among the three dimensions, social-emotional digital literacy was the least powerful predictor (β = .18). This is a slightly different trend from the western contexts where online collaboration skills are more often highlighted in blended learning pedagogy and may be due to the relatively instructor-centered, content delivery focus of many Sri Lankan e-learning implementations, where student-to-student online collaboration skills are not used as much as content access and assessment skills.

Because effective use was analyzed as a composite of three facets (content access, assessment tools and communication channels), the results show that the three literacy dimensions are associated with effective use overall, but not which facet each dimension relates to. In line with Fig. 1, technical literacy is expected to matter most for content access and assessment tools, which require operating the LMS, cognitive literacy for purposeful engagement with content, and social-emotional literacy for communication channels; the smaller social-emotional coefficient is consistent with the limited use of such channels in instructor-centered courses. These facet-level links were not tested directly and are a priority for future research.

The overall model accounted for 38.6% of the variance in effective e-learning use, suggesting that digital literacy is a significant but not exclusive predictor of effective use, as other factors like internet connectivity, device quality, instructor design of online content, and motivation are likely to be involved as well, in line with the general TAM literature on facilitating conditions and self-efficacy [6], [7].

*B. Limitations*

Several limitations should be noted. (a) Sampling: respondents were recruited through convenience and snowball sampling, so the sample is non-probabilistic, and the largest subgroup consisted of computing/IT students (40.2%, Table I), who may have above-average technical literacy; the findings may therefore not generalize to all Sri Lankan undergraduates. (b) Common-method bias: the predictors and the outcome were

measured with the same self-report questionnaire at the same time, which may inflate the observed associations, and self-reported effective use may differ from actual platform use. Anonymity and voluntary participation were used as procedural remedies, but no statistical check such as Harman's single-factor test was performed. (c) Cross-sectional design: data were collected at one point in time, so the results indicate associations and not causal effects or their direction; the ordering assumed in the framework rests on theory. (d) Model scope: demographic variables were not controlled, residual diagnostics other than multicollinearity were not reported, effective use was analyzed as a single composite rather than by facet, and 61.4% of the variance is not explained by the model, so other factors (e.g., connectivity, devices, course design, motivation) are likely to matter.

## VIII. RECOMMENDATIONS

The following recommendations are made to the higher education institutions and policy makers in Sri Lanka based on the findings:

1) Integrate structured digital literacy modules into first-year undergraduate orientation programmes, not assuming that students are digitally competent.

2) Redesign LMS onboarding to incorporate brief, compulsory practical training on key platform features, especially for students from under-resourced schools or rural areas who were found to have comparatively low levels of prior ICT training in national studies [1].

3) Strengthen information-literacy teaching, for example through library workshops that show students how to evaluate sources and avoid plagiarism, since cognitive digital literacy had the lowest mean score in this study (M = 3.45).

4) Promote more collaborative online learning activities (e.g., discussion forums and peer-review tasks) so that students' social-emotional digital literacy, which they rated highest (M = 3.71), is actually used in the platform's communication features. This dimension had the weakest association with effective use (β = .18), possibly because such activities are limited in instructor-centered courses.

5) The University Grants Commission and individual universities should consider a standardized, DigComp aligned digital skills certification for incoming undergraduates, and ongoing investment in campus and rural internet infrastructure to overcome the underlying access barriers [10].

## IX. CONCLUSION

The study investigated the relationship between digital literacy and the effective use of e-learning platforms among 204 undergraduate students in Sri Lanka. The results indicate that all three dimensions of digital literacy (technical, cognitive, and social-emotional) significantly and positively predict effective e-learning use, jointly accounting for about 39% ($R^2$ = .386) of the variance in the use of e-learning, with technical digital literacy being the strongest predictor. As the data are cross-sectional, self-reported and drawn from a non-probability sample, these results indicate associations rather than causal effects. The findings indicate that Sri Lankan higher education institutions need to go beyond merely providing e-learning infrastructure and invest purposefully in the development of digital literacy in a multidimensional manner, especially in the areas of critical information evaluation and online collaboration, so that students can make meaningful learning gains from the use of the platforms. Future studies should be conducted with larger, probability-based, stratified samples across all provinces of Sri Lanka and with longitudinal or mixed methods designs, objective LMS usage data and facet-level analysis of effective use to further validate and generalize these findings.

## REFERENCES


[1] N. Subashini, L. Udayanga, L. H. N. De Silva, J. C. Edirisinghe, and M. N. Nafla, "Undergraduate perceptions on transitioning into E-learning for continuation of higher education during the COVID pandemic in a developing country: a cross-sectional study from Sri Lanka," *BMC Med. Educ.*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12909-022-03586-2.

[2] K. Samarasekara, "e-Learning in Medical Education in Sri Lanka: Survey of Medical Undergraduates and New Graduates," *JMIR Med. Educ.*, vol. 8, no. 1, Jan. 2022, doi: 10.2196/22096.

[3] W. Ng, "Can we teach digital natives digital literacy?," *Comput. Educ.*, vol. 59, no. 3, pp. 1065–1078, Nov. 2012, doi: 10.1016/j.compedu.2012.04.016.

[4] Riina. Vuorikari, Stefano. Kluzer, and Yves. Punie, *DigComp 2.2, The Digital Competence framework for citizens : with new examples of knowledge, skills and attitudes*. Publications Office of the European Union, 2022.

[5] "Validating the Digital Competence (Dig-Comp 2.1) Framework in Higher Education Using Confirmatory Factor Analysis: Non-Western Perspective," *Canadian Journal of Educational and Social Studies*, vol. 3, no. 6, 2023, doi: 10.53103/cjess.v3i6.184.

[6] M. N. Al-Suqri, "Perceived usefulness, perceived ease-of-use and faculty acceptance of electronic books: An empirical investigation of Sultan Qaboos University, Oman," *Library Review*, vol. 63, no. 4–5, pp. 276–294, Jan. 2014, doi: 10.1108/LR-05-2013-0062.

[7] A. Granić, "Educational Technology Adoption: A systematic review," *Educ. Inf. Technol. (Dordr).*, vol. 27, no. 7, pp. 9725–9744, Aug. 2022, doi: 10.1007/s10639-022-10951-7.

[8] S. Zakir *et al.*, "Digital literacy and academic performance: the mediating roles of digital informal learning, self-efficacy, and students' digital competence," *Front. Educ. (Lausanne).*, vol. 10, 2025, doi: 10.3389/feduc.2025.1590274.

[9] Y. Zhao, M. C. Sánchez Gómez, A. M. Pinto Llorente, and L. Zhao, "Digital competence in higher education: Students' perception and personal factors," *Sustainability (Switzerland)*, vol. 13, no. 21, Nov. 2021, doi: 10.3390/su132112184.

[10] Y. Assefa, M. M. Gebremeskel, B. T. Moges, S. A. Tilwani, and Y. A. Azmera, "Rethinking the digital divide and associated educational in(equity) in higher education in the context of developing countries: the social justice perspective," *International Journal of Information and Learning Technology*, vol. 42, no. 1, pp. 15–32, Jan. 2025, doi: 10.1108/IJILT-03-2024-0058.

# Urban Green Space Proximity and Population Coverage Using GIS and Urban Computing for Sustainable Planning in Colombo District Sri Lanka

W.I.Y.Rodrigo
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
isunimayalindi02@gmail.com

Hiruni Weerasinghe
*Department of Computer and Data Science*
*Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
hiruni.w@nsbm.ac.lk

***Abstract*—This study estimates population proximity to mapped green spaces in Colombo District, Sri Lanka, through a GIS workflow combining OpenStreetMap polygons, administrative boundaries, Grama Niladhari population counts, and occupied housing-unit data. Dissolved Euclidean buffers of 500 m and 1,000 m were intersected with GN polygons, and population was allocated in proportion to the intersected area. Reported coverage is 1,796,880 modeled person-equivalents at 500 m and 2,026,340 at 1,000 m, representing 75.64% and 85.30% of a project reference population of 2,375,420. Increasing the threshold adds 229,460 person-equivalents, or 9.66 percentage points. These values describe potential spatial proximity under a uniform population-distribution assumption; they do not measure walking access, public usability, or observed visitation. Green-space extraction is incomplete in Padukka, Homagama, and Seethawaka, while inventory completeness elsewhere remains unquantified. The study contributes a local application of established spatial methods and a transparent comparison of two distance scenarios. The estimates support preliminary investigation, subject to reconciliation of population provenance, completion of the green-space inventory, and spatial validation.**

***Keywords*—*Urban green space, population coverage, spatial proximity, GIS,Colombo District, urban computing.***

## I. Introduction

Urban green spaces can support recreation, social interaction, and environmental quality, but the amount of mapped green land does not reveal who can use it. Wolch et al. [1] discuss green-space provision in relation to public health and environmental justice. Haaland and Konijnendijk van den Bosch [2] describe the planning pressures that accompany urban densification. Together, these perspectives motivate assessments that relate green-space locations to residential populations while distinguishing physical proximity from usable access.

Colombo has an established body of green-space research. Senanayake et al. [3] integrated satellite-derived green-space information with population and environmental indicators at GN level in Colombo Municipal Council. Li and Pussella [4] examined urban green-space provision and Kumuduni and Weerakoon [5] considered how green spaces should be defined across multiple uses in Colombo District. Subasinghe et al. [6] documented urban growth in the Colombo Metropolitan Area. These studies provide local context and show that GIS integration or GN-level analysis alone cannot be presented as a new methodology.

The present study asks how much of a project reference population is estimated to lie within 500 m and 1,000 m of an available mapped green-space inventory, and how much that estimate changes when the threshold increases. It integrates spatial and demographic information through an applied urban-computing workflow using the data-integration sense of the term described by Zheng et al. [7]. The contribution is a case-specific population-coverage assessment using established methods. It does not introduce a new algorithm or infer statistically significant clusters of disadvantage.

The objectives are to combine the available green-space and administrative layers with GN population and housing counts, define population and housing density as contextual indicators, estimate population within each distance scenario and examine the difference between the two estimates. The interpretation is restricted to potential proximity to the mapped inventory. Incomplete extraction and unresolved data provenance constrain district-wide planning conclusions.

The World Health Organization reviews the links between urban greenery and health and distinguishes availability, accessibility, and use [8]. Kabisch et al. [9] assess green-space availability across European cities, illustrating the value of combining land-use information with population data. These findings motivate population-based assessment, but they do not establish local health benefits or a universal access standard for Colombo.

## II. Related Work and Analytical Scope

Oh and Jeong [10] used GIS network analysis to examine pedestrian access and park serviceability in Seoul. Their approach relates the location of parks to the population potentially served, rather than relying only on total park area. Chen et al. [11] used fine-resolution population and greenspace information to examine global exposure inequality. Their exposure measures and geographic scope differ from the buffer-based coverage indicator used here.

*Distance measures.* Network analysis offers a more direct representation of movement than an unrestricted buffer. Comber

et al. [12] used a GIS network approach to examine differences in green-space access among population groups. Gupta et al. [13] assessed access across green-space hierarchies using network distances. Their work demonstrates the value of matching the access measure to the destinations and users being studied. A buffer around a polygon measures proximity to its nearest mapped part; a walking analysis should represent usable entrances and permitted routes.

La Rosa [14] examined GIS indicators for green-space accessibility, while Ekkel and de Vries [15] evaluated different accessibility metrics in the health literature. These approaches answer related but distinct questions. A local proximity percentage cannot be compared directly with a published per-capita area, network-distance statistic, or health association. Meaningful quantitative benchmarking requires consistent boundaries, population dates, destination definitions, and measurement units.

*Quality and equity.* Stessens et al. [16] combined proximity and quality in a GIS model for Brussels. Rigolon et al. [17] reviewed disparities in green-space access across Global South cities. These studies support extending a proximity assessment with quality and socioeconomic information, but neither quality nor socioeconomic inequality can be inferred from the aggregate Colombo counts alone. No demographic-group comparison or observed-use survey forms part of the present analysis.

*Inventory uncertainty.* Ludwig et al. [18] combined OpenStreetMap and Sentinel-2 information to map public urban green spaces while addressing source uncertainty. Their work is relevant because a mapped green polygon may be vegetated without being publicly accessible, and an unmapped location may still contain eligible green space. The Colombo inventory therefore requires both completeness checks and verification of public usability. The present scope is limited to a two-threshold Euclidean assessment; hotspot statistics and pedestrian-network results are not claimed.

Table 1 compares selected methods by the question they answer and the inputs they require. It positions the Colombo analysis as a transparent proximity baseline. The comparison is methodological; no numerical superiority over these studies is asserted.

TABLE 1
SELECTED METHODS AND THEIR RELATIONSHIP TO THE COLOMBO ASSESSMENT

| Study | Method and purpose | Relationship to this study |
|---|---|---|
| Oh and Jeong [10] | GIS network analysis of park serviceability | Shows how pedestrian routes can extend a distance-only assessment |
| Comber et al. [12] | Network access for population groups | Requires route and demographic information beyond GN totals |
| Gupta et al. [13] | Network access by green-space hierarchy | Supports distinguishing destinations by size and function |
| Stessens et al. [16] | Proximity and green-space quality model | Quality information is outside the present indicator |
| Senanayake et al. [3] | Remote sensing and GN-level analysis in Colombo | Provides local precedent; the present output compares two population-proximity scenarios |

## III. STUDY AREA AND DATA

*Geographic coverage.* Colombo District is the administrative context. Grama Niladhari divisions provide the population-allocation units and Divisional Secretariat divisions provide broader geographic context. The source analysis identifies incomplete green-space extraction in Padukka, Homagama, and Seethawaka. These areas remain in the administrative and population framework. Their missing mapped features are treated as inventory uncertainty rather than evidence of absent green space.

*Input datasets.* The analytical inputs comprise OSM green-space polygons extracted using QuickOSM, GN and DS boundaries, GN population counts and occupied housing-unit counts. Reported green-space categories include parks, gardens, forests, wetlands, grass areas, greenfields, and village greens. This is a broad mapped-land inventory category membership does not by itself establish permission to enter or suitability for recreation. Housing counts provide contextual density information and are not used as a second weight in population coverage.

*Population denominator.* Both scenarios use 2,375,420 as the reference population. This value is retained as a project reference total rather than presented as a verified contemporary district census count. Its census or reference year, the complete list of contributing GN units, DS subtotals and join completeness have not been independently reconciled. Retaining the incomplete-inventory areas maintains a common denominator for the two scenarios, but it also means that missing green-space features can affect coverage estimates for residents included in that denominator.

Removing Padukka, Homagama, or Seethawaka after calculation would not automatically correct the result. An assessment restricted to areas with an audited inventory must recompute both its numerator and denominator for the same GN subset. The published aggregate totals do not contain enough information to calculate that restricted-area percentage. Table 2 distinguishes the inputs used by the model from metadata that remain necessary for independent spatial replication.

TABLE 2
ANALYTICAL INPUTS AND UNRESOLVED PROVENANCE

| Input | Analytical role | Provenance still required |
|---|---|---|
| OSM green-space polygons | Destinations for Euclidean buffers | Extraction date, exact query and tags, eligibility checks, counts and union area |
| GN and DS boundaries | Population allocation and geographic context | Boundary source, reference date, included unit list and geometry checks |
| GN population counts | Weights and reference total | Census year, source table, matched keys and DS subtotals |
| Occupied housing counts | Contextual housing density | Reference year, source and join completeness |

## IV. Methodology

The analysis follows the sequence in Fig. 1: data collection, preprocessing, green-space mapping, dissolved-buffer modelling, GN intersection and population allocation, and interpretation. Visual checking supports data preparation; it is not a measured validation of inventory completeness or public access.

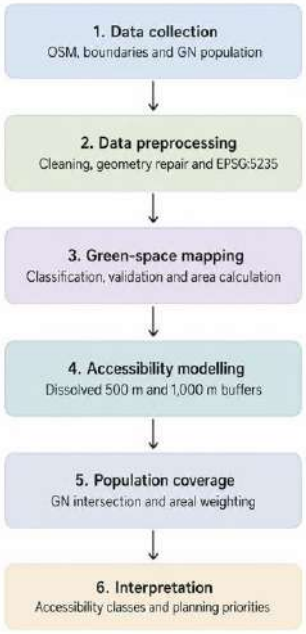



Fig. 1. GIS workflow for the green-space proximity and population-coverage assessment. The validation stage denotes reported visual checking and preprocessing; public usability and inventory completeness have not been independently measured.

*Preprocessing and joins.* The reported workflow transforms spatial inputs to the metric analytical coordinate reference system EPSG:5235 before distance and area calculations. Preprocessing comprises geometry repair, removal of duplicate and irrelevant features, standardization of categories, and joining population attributes through cleaned GN identifiers. These operations must be checked for changes in feature count and area. Visual inspection can identify obvious problems but does not provide a measured inventory-accuracy estimate.

GN identifiers must match uniquely between spatial and demographic tables. Unmatched boundaries, duplicate population keys, null counts, and excluded units can alter the reference population and its spatial allocation. A reproducible implementation should record the number of boundaries before and after joining, the counts of matched and unmatched identifiers, and the population sum for the retained set. The reference total must be reconciled against that retained set before any coverage percentage is interpreted.

*Contextual density.* Population density is the GN population divided by GN area in square kilometers; housing density is occupied housing units divided by the same area. These indicators describe the residential context of the study units. They do not resolve the distribution of residents inside a GN polygon and are not evidence of socioeconomic disadvantage. Numerical GN density rankings are not reported because a verified GN results table is unavailable.

*Euclidean buffer construction.* For each distance d, buffers around retained green-space polygons are dissolved to form a single union B(d). Dissolving is essential because overlapping buffers must not count the same residential area repeatedly. The service-zone union is then intersected with each GN polygon $G_i$. The proportion of GN area inside the union is calculated using Eq. 1. All areas in this ratio use the same planar coordinate system.

$$r_i(d) = \frac{area(G_i \cap B(d))}{area(G_i)}. \quad (1)$$

For a valid GN polygon with positive area, the ratio must lie between zero and one. Values outside this range require checks for geometry errors, overlapping inputs, duplicate intersections, or inconsistent measurement units. Buffering measures proximity to the entire mapped polygon, including its edge. It does not identify an entrance, a permitted path, or a crossing over a canal, railway, or major road.

*Population allocation.* Population inside the buffer is estimated by multiplying the GN population $P_i$ by its covered area fraction, as shown in Eq. 2. This assumes that residents are uniformly distributed across the GN. A covered nonresidential area receives the same modeled population density as a covered residential area within that unit. The output is therefore a modeled person-equivalent quantity, even when displayed as a whole number.

$$P_i(d) = P_i \times r_i(d). \quad (2)$$

Let I be the GN units retained in the analysis. The aggregate estimate and reference population use exactly the same unit set. Coverage C(d) is expressed as a percentage, and U(d) is the modeled population outside the mapped buffer. Equations 3 to 5 define these quantities.

$$P(d) = \sum_{i \in I} P_i(d), \quad P = \sum_{i \in I} P_i. \quad (3)$$

$$C(d) = 100 \times \frac{P(d)}{P}. \quad (4)$$

$$U(d) = P - P(d). \quad (5)$$

For any GN with positive population, its estimated within-buffer percentage equals 100 times its covered area fraction. Population weighting changes the district aggregate and the estimated number of people outside buffers; it does not produce a finer within-GN population model. A GN with zero population has no defined population-coverage percentage. Intermediate estimates should retain precision, with rounding applied only after aggregation.

*Boundary effects.* Clipping green spaces to the administrative boundary before buffering can remove destinations immediately outside the district that may be near residents inside it. A future spatial rerun should retain eligible destinations at least 1,000 m beyond the population-study boundary, construct buffers, and clip the final service zones to the study area. The effect of this edge treatment has not been quantified in the reported totals.

*Verification and reproducibility.* Available verification consists of recalculating percentages and complements from the aggregate counts and checking consistency between the two thresholds. QGIS and QuickOSM were used for extraction, cleaning, joins, buffers, overlays, field calculations and map design; Microsoft Excel supported summary checks. The exact QGIS and QuickOSM versions, OSM extraction date, original query and tag filters, buffer-segment settings, classification break values, and complete processing log are not documented.

The green-space feature count, dissolved green-space area, distribution by DS or GN, and audited inventory-coverage proportion are not established by the available aggregate results. Neither the spatial intersections nor final map outputs have been independently reproduced. No network-distance comparison, alternative population-allocation experiment or statistical hotspot test is reported. Arithmetic consistency is evaluated separately from these unresolved spatial-validation requirements.

## V. Results

*Spatial outputs.* The analysis includes GN maps for both thresholds, presented in Figs. 2 and 3. The legends show qualitative classes but do not specify numerical class boundaries and use different colour schemes, so colours cannot be compared directly across maps or used to quantify changes in GN class. Low mapped coverage in Padukka, Homagama, and Seethawaka remains confounded by incomplete inventory coverage. The maps therefore provide spatial context without supporting definitive GN rankings.

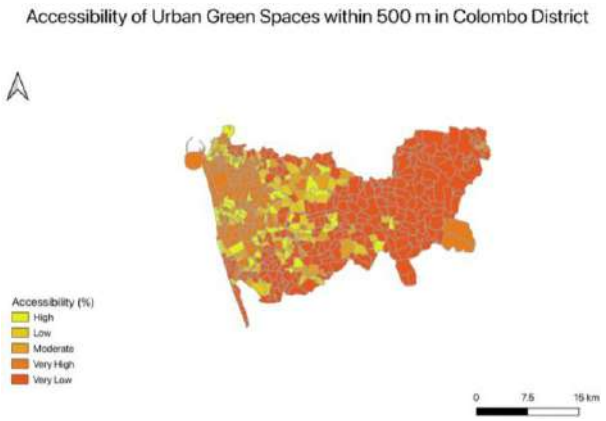



Fig. 2. GN proximity classes at 500 m. The analytical CRS reported by the study is SLD99 / Sri Lanka Grid 1999, EPSG:5235. Green-space source: OpenStreetMap contributors. Numerical class breaks and extraction date are undocumented; the classes indicate mapped proximity rather than verified walking access.

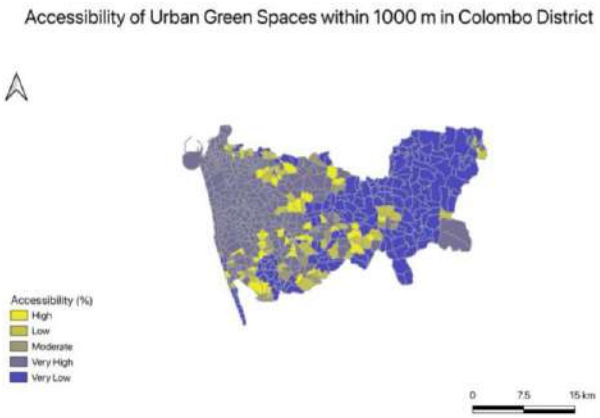



Fig. 3. GN proximity classes at 1,000 m. The analytical CRS reported by the study is EPSG:5235. Green-space source: OpenStreetMap contributors. Class breaks and extraction date are undocumented, and the colour scheme differs from Fig. 2. Incomplete mapping must not be interpreted as an absence of green space.

*Population within each threshold.* Table 3 summarizes the reported counts and the percentages recalculated from the common reference population. At 500 m, estimated coverage is 1,796,880 person-equivalents, leaving 578,540 outside the mapped union. At 1,000 m, estimated coverage rises to 2,026,340, leaving 349,080 outside. These values indicate modeled proximity to the available inventory. They are not observed counts of park users or verified counts of residents with a usable walking route.

TABLE 3
POPULATION PROXIMITY UNDER THE TWO DISTANCE SCENARIOS

| Measure | 500 m | 1,000 m |
|---|---|---|
| Project reference population | 2,375,420 | 2,375,420 |
| Estimated population within buffer | 1,796,880 | 2,026,340 |
| Share within buffer [%] | 75.64 | 85.30 |
| Estimated population outside buffer | 578,540 | 349,080 |
| Share outside buffer [%] | 24.36 | 14.70 |

Counts are modeled person-equivalents reported as whole numbers. Percentages and outside-buffer totals are recalculated from the reported reference total; source provenance and inventory completeness remain unresolved.

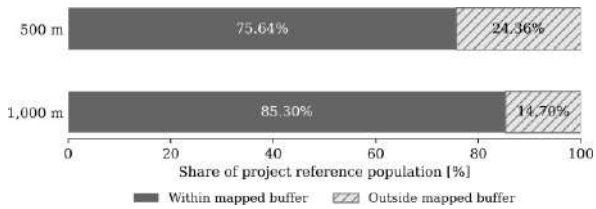



Fig. 4. Shares of the project reference population within and outside the mapped buffers. Values are reconstructed from reported aggregate counts under uniform GN population allocation. The figure is a numerical summary, not a map or a walking-access validation.

*Threshold comparison.* Increasing the threshold adds 229,460 modeled person-equivalents. Dividing this difference by 2,375,420 gives an increase of 9.66 percentage points.

Relative to the 500 m covered estimate, the increase is 12.77%. Relative to the 500 m outside-buffer estimate, the reduction is 39.66%. These percentages answer different questions and use different denominators the relative increases must not be substituted for the percentage-point difference.

The two nested scenarios also permit a three-band summary: 1,796,880 person-equivalents within 500 m, 229,460 in the additional area reached between 500 m and 1,000 m, and 349,080 beyond 1,000 m. Their shares are 75.64%, 9.66%, and 14.70%, respectively. This decomposition is derived from the reported totals and does not constitute a new spatial overlay. Fig. 4 shows the same threshold effect as two population-share bars.

*Consistency checks.* Within-buffer and outside-buffer estimates sum to 2,375,420 at each threshold. The 1,000 m estimate exceeds the 500 m estimate as expected for nested buffers around the same destinations. These checks detect inconsistencies in the summary arithmetic. They do not establish that polygons are complete, that joins are correct or that the population model accurately represents residential locations.

The comparison demonstrates dependence on the selected distance, but it cannot establish robustness across other thresholds, population distributions, or spatial units. Aggregate results do not reveal which GN divisions move between coverage classes or whether local priority rankings remain stable. Accordingly, no GN is ranked as definitively underserved and no statistically significant hotspot or coldspot is inferred from these totals.

## VI. Discussion

*Meaning of the threshold effect.* The principal finding is the size of the change between the two proximity scenarios. An additional 229,460 person-equivalents fall inside the expanded mapped buffer under the same population model. The result shows why a single distance threshold can conceal a substantial intermediate-distance population. It does not demonstrate that extending an acceptable walking distance would improve actual access, because residents may face barriers or lack usable entrances.

*Comparison with alternative methods.* For a fixed set of publicly usable destinations and consistent distance definitions, a route through a pedestrian network cannot be shorter than the straight-line distance between the same locations. In this study however, Euclidean distance is measured to mapped polygons rather than verified entrances. Differences in destination eligibility, entrance placement, network connectivity and population representation prevent application of a universal correction factor. The network approaches in [12], [13] are methodological comparators, not measured baselines for these Colombo totals.

Population allocation is a second source of uncertainty. Uniform areal weighting is transparent and can use available GN totals but it distributes residents over every included part of the polygon. A residential-building or population-grid allocation could place more or fewer residents inside the same buffer. The direction of change depends on the relationship between settlement locations and green-space proximity. The aggregate counts alone cannot establish that relationship.

*Inventory and measurement uncertainty.* Incomplete mapping and unsuitable mapped destinations affect interpretation in different ways. Adding omitted eligible polygons to a fixed inventory can only maintain or increase its Euclidean buffer coverage when all other rules remain unchanged. Yet this property does not make the current estimate a lower bound on actual usable access. Private or inaccessible mapped land and ignored barriers may overstate usable access while omitted features or boundary clipping may understate mapped proximity. No overall error direction or confidence interval is justified without further evidence.

Population, boundary and OSM reference dates also require reconciliation. A population total from one period combined with green-space mapping from another describes neither period exactly. Changing the administrative zoning can further affect local area-weighted estimates and rankings. The study does not measure opening hours, safety, maintenance, capacity, travel behaviour, socioeconomic equity or observed health outcomes. These limits apply even where a mapped polygon and its buffer are geometrically correct.

*Planning implications.* The immediate planning use is to guide further investigation. Padukka, Homagama and Seethawaka require inventory completion before a low mapped-coverage result can be interpreted as a green-space deficit. Other areas also need an audit because their completeness has not been quantified. Where a verified analysis eventually identifies a large population far from eligible destinations, field checks should distinguish a shortage of green land from a lack of entrances, crossings, paths or public permission. Each problem calls for a different intervention.

*Further validation.* Further validation should begin by reconciling GN membership and the reference population, then reporting green-space counts and dissolved areas by DS and GN. A matched walking-network assessment should use the same population base and verified public destinations, document route restrictions and disconnected components, and evaluate both 500 m and 1,000 m. Absolute count differences, percentage-point differences, and changes in GN rankings would allow a quantitative comparison with the current model.

Additional scenarios could compare 300 m and 750 m thresholds, uniform allocation with residentially constrained allocation, and GN outputs aggregated to DS level. The same GN totals should constrain alternative population surfaces so that differences reflect allocation rather than a changed population base. An audited-area analysis should explicitly define its retained units and recompute both totals. These are proposed extensions not completed experiments or evidence that the present estimates are robust.

*Ethics and interpretation.* The study uses aggregated secondary population and housing information and does not analyze individually identifiable records. Aggregation limits personal data exposure, but it also limits what can be inferred about individual access, behaviour or disadvantages. Area-based screening should guide local investigation rather than label residents or communities from an unverified map class.

## VII. Summary

The reported GIS assessment places 75.64% of a project reference population within 500 m of mapped green-space polygons and 85.30% within 1,000 m. Expanding the threshold adds 229,460 modeled person-equivalents, equivalent to 9.66 percentage points. The calculations are arithmetically consistent and provide a clear comparison of two straight-line proximity scenarios under uniform GN population allocation.

The contribution is an applied population-coverage assessment for Colombo District using established GIS methods. The findings remain conditional on the mapped inventory and population denominator. Incomplete extraction in Padukka, Homagama and Seethawaka unresolved source metadata, and the absence of spatial and network validation limit their use for definitive district-wide prioritization. Completing the inventory, reconciling demographic provenance and evaluating routes to verified entrances are the next requirements for stronger planning evidence.

## References


[1] J.R. Wolch, J. Byrne, J.P. Newell, Urban green space, public health, and environmental justice: The challenge of making cities "just green enough", Landscape and Urban Planning 125 (2014) 234–244. doi: 10.1016/j.landurbplan.2014.01.017.

[2] C. Haaland, C. Konijnendijk van den Bosch, Challenges and strategies for urban green-space planning in cities undergoing densification: A review, Urban Forestry & Urban Greening 14 (2015) 760–771. doi: 10.1016/j.ufug.2015.07.009.

[3] I.P. Senanayake, W.D.D.P. Welivitiya, P.M. Nadeeka, Urban green spaces analysis for development planning in Colombo, Sri Lanka, utilizing THEOS satellite imagery -A remote sensing and GIS approach, Urban Forestry & Urban Greening 12 (2013) 307-314. doi: 10.1016/j.ufug.2013.03.011.

[4] L. Li, P.G.R.N.I. Pussella, Is Colombo city, Sri Lanka secured for urban green space standards? Applied Ecology and Environmental Research 15 (2017) 1789-1799. doi: 10.15666/aeer/1503_17891799.

[5] Y. Kumuduni, K.G.P.K. Weerakoon, Defining urban green spaces in the Colombo District across multiple uses, Journal of Real Estate Studies 21 (2024) 1-12. doi: 10.31357/jres.v21i2.7543.

[6] S. Subasinghe, R.C. Estoque, Y. Murayama, Spatiotemporal analysis of urban growth using GIS and remote sensing: A case study of the Colombo Metropolitan Area, Sri Lanka, ISPRS International Journal of Geo-Information 5 (2016) 197. doi: 10.3390/ijgi5110197.

[7] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: Concepts, methodologies, and applications, ACM Transactions on Intelligent Systems and Technology 5 (2014) 38. doi: 10.1145/2629592.

[8] World Health Organization Regional Office for Europe, Urban Green Spaces and Health: A Review of Evidence, WHO Regional Office for Europe, Copenhagen, 2016. https://www.who.int/europe/publications/i/item/WHO-EURO-2016-3352-43111-60341.

[9] N. Kabisch, M. Strohbach, D. Haase, J. Kronenberg, Urban green space availability in European cities, Ecological Indicators 70 (2016) 586-596. doi: 10.1016/j.ecolind.2016.02.029.

[10] K. Oh, S. Jeong, Assessing the spatial distribution of urban parks using GIS, Landscape and Urban Planning 82 (2007) 25-32. doi: 10.1016/j.landurbplan.2007.01.014.

[11] B. Chen, S. Wu, Y. Song, C. Webster, B. Xu, P. Gong, Contrasting inequality in human exposure to greenspace between cities of Global North and Global South, Nature Communications 13 (2022) 4636. doi: 10.1038/s41467-022-32258-4.

[12] A. Comber, C. Brunsdon, E. Green, Using a GIS-based network analysis to determine urban greenspace accessibility for different ethnic and religious groups, Landscape and Urban Planning 86 (2008) 103-114. doi: 10.1016/j.landurbplan.2008.01.002.

[13] K. Gupta, A. Roy, K. Luthra, S. Maithani, Mahavir, GIS based analysis for assessing the accessibility at hierarchical levels of urban green spaces, Urban Forestry & Urban Greening 18 (2016) 198–211. doi: 10.1016/j.ufug.2016.06.005.

[14] D. La Rosa, Accessibility to greenspaces: GIS based indicators for sustainable planning in a dense urban context, Ecological Indicators 42 (2014) 122-134. doi: 10.1016/j.ecolind.2013.11.011.

[15] E.D. Ekkel, S. de Vries, Nearby green space and human health: Evaluating accessibility metrics, Landscape and Urban Planning 157 (2017) 214-220. doi: 10.1016/j.landurbplan.2016.06.008.

[16] P. Stessens, A.Z. Khan, M. Huysmans, F. Canters, Analysing urban green space accessibility and quality: A GIS-based model as spatial decision support for urban ecosystem services in Brussels, Ecosystem Services 28 (2017) 328-340. doi: 10.1016/j.ecoser.2017.10.016.

[17] A. Rigolon, M.H.E.M. Browning, K. Lee, S. Shin, Access to urban green space in cities of the Global South: A systematic literature review, Urban Science 2 (2018) 67. doi: 10.3390/urbansci2030067.

[18] C. Ludwig, R. Hecht, S. Lautenbach, M. Schorcht, A. Zipf, Mapping public urban green spaces based on OpenStreetMap and Sentinel-2 imagery using belief functions, ISPRS International Journal of Geo-Information 10 (2021) 251. doi: 10.3390/ijgi10040251.

# A Role-Sensitive Framework for Assessing Organizational Readiness Under Sri Lanka's Personal Data Protection Act

Hishara Samaranayake
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwatta,Pitipana, Sri Lanka*
hnsamara@students.nsbm.ac.lk

T A H Dilpriya
*Department of Computer and Data Science*
*NSBM Green University*
*Mahenwatta,Pitipana, Sri Lanka*
hirushi.d@nsbm.ac.lk

Madusanka Mithrananda
*Department of Software Engineering and Computer Security*
*NSBM Green University*
*Mahenwatta,Pitipana, Sri Lanka*
madusanka.m@nsbm.ac.lk

***Abstract*—Translating data-protection legislation into measurable organizational practices remains a challenge, particularly when responsibilities and levels of awareness vary across organizational roles. This paper proposes a role-sensitive framework for assessing organizational readiness under Sri Lanka's Personal Data Protection Act (PDPA), No. 9 of 2022, as amended in 2025. The framework establishes a traceable mapping from selected statutory requirements to organizational and technical controls, readiness criteria, and role-specific assessment statements for executive, middle-management, and operational personnel. To provide a structured assessment of readiness, the framework incorporates a weighted Organizational Readiness Index (ORI) and a Cross-Level Discrepancy Index (DSI), enabling organizations to examine overall readiness alongside differences in perspectives across organizational levels. The framework was evaluated in three organizations involving 17 respondents. Two organizations achieved Level 4 maturity, with ORI scores of 62.29% and 62.08%, while the third achieved Level 2 maturity, with a score of 40.83%. The observed DSI values further highlighted differences in readiness perceptions across organizational levels. In addition, feedback from 15 experienced professionals in information systems audit, information security, privacy, and technology governance informed the assessment of statement clarity and relevance. The findings demonstrate how combining role-specific assessment with quantitative indicators can help organizations identify readiness gaps, understand differences in internal perspectives, and establish priorities for improvement. The proposed framework serves as a privacy-preserving diagnostic tool to support organizational readiness assessment, rather than a certification of legal compliance.**

***Keywords—Legal-to-Control Mapping, Organizational Readiness, Personal Data Protection Act (PDPA), Privacy Governance, Role-Sensitive Assessment***

## I. Introduction

The growing use of personal data has increased the need for organizations to translate data-protection law into practical internal controls. In Sri Lanka, the PDPA, No. 9 of 2022 [1] , together with the Personal Data Protection (Amendment) Act, No. 22 of 2025 [2] , establishes requirements relating to processing, accountability, data-subject rights, security, and controller and processor responsibilities. Translating such legal requirements into organizational practice can be difficult for non-legal users because privacy regulations often involve specialized terminology and interpretation [3] . Formal assessments can also require considerable expertise, evidence review, and time [4] .

Many existing assessment approaches are designed around the GDPR and cannot be directly applied to Sri Lanka's current PDPA without remapping the legal requirements. Legislative assessment research also emphasizes the importance of maintaining traceability between legal obligations and the questions or controls used to assess them [5] . Organizational readiness may further differ across responsibility levels: executive management establishes governance, middle management implements controls, and operational employees apply them in daily work. A single organizational response may therefore hide important differences.

This study investigates how selected PDPA requirements can be translated into assessable organizational and technical controls and adapted to executive, middle-management, and operational responsibilities. The proposed framework maintains traceability from statutory provision to control, readiness criterion, and assessment statement, while supporting multiple respondents within the same organizational level. It is developed as the basis for a software-supported organizational readiness assessment.

## II. Related Work and Research Gap

### *A. Privacy Compliance Frameworks*

Research on privacy compliance has explored different ways of converting regulatory requirements into structured assessment mechanisms. Chatsuwan et al. developed a quantitative approach for evaluating compliance with Thailand's PDPA using defined privacy criteria [6] . Zaeem and Barber showed how privacy requirements can be systematically transformed into assessable elements [7] . Ngobeni et al. extended this idea through a software-supported GDPR assessment toolkit incorporating assessment management, analytics, reporting, and improvement identification [8] . Brodin approached the problem from an organizational perspective, highlighting the difficulty smaller organizations may experience when regulatory knowledge and resources are limited [9] . These studies demonstrate that privacy assessment can combine regulatory interpretation, measurable criteria, and software support, although the underlying legal context differs from Sri Lanka.

### *B. Self-Assessment in Cybersecurity Data Protection*

Data-protection readiness has also been studied through organizational and cybersecurity-related factors. Nnko et al. examined readiness through policies, staff training, technical resources, awareness, and management support [10] . Ndichu and Wausi considered regulatory, organizational, and technological readiness while involving respondents with different organizational responsibilities [11] . Agusti and Rakhmawati emphasized governance and technical safeguards [12] , whereas Baloyi and Kotzé examined knowledge, policies, standards, training, and awareness [13] . Wijaya et al. focused more specifically on employee information-security awareness, including knowledge, attitudes, and behaviour relevant to personal-data protection [14] . Together, these studies indicate that readiness depends on both organizational capability and the people responsible for applying controls.

Sri Lankan research provides further context. Mithrananda examined data-protection readiness in private-sector higher-education institutions using the 2019 draft data-protection framework [15] . Ranathunga and Wickramarachchi later proposed a data-governance model to support GDPR adoption in Sri Lankan businesses [16] . These studies are useful local foundations for understanding organizational preparedness and governance.

### *C. Research Gap*

The existing literature addresses several relevant components separately: regulatory assessment, organizational readiness, employee awareness, governance, and software-supported compliance tools. What remains limited is an approach developed around the current Sri Lankan PDPA that preserves traceability from statutory requirements to assessable controls while representing different organizational responsibilities within the same framework. The present study addresses this gap through a role-sensitive readiness assessment designed for executive, middle-management, and operational respondents and structured for later software implementation.

## III. Methodology: Framework and Assessment Engine

This study follows a Design Science Research approach because the research develops an artefact to address an identified organizational problem [17] . The development process was structured around problem identification, artefact design, preliminary evaluation, and subsequent refinement, consistent with the Design Science Research Methodology described by Peffers et al. [18] . The artefact at this stage consists of the PDPA readiness framework, its legal mapping structure, role-specific assessment statements, and prototype assessment process.

### *A. Legal-to-Technical Control Mapping*

The framework required a systematic method for translating statutory obligations into controls that could be assessed within an organization. The mapping procedure was adapted from the legal-to-control alignment approach presented by Breaux et al. [19] , where regulatory requirements are related to corresponding information-technology controls while retaining the relationship with the original legal source. Rather than directly adopting an existing control catalogue, the present study applies this principle to the organizational context of Sri Lanka's PDPA.

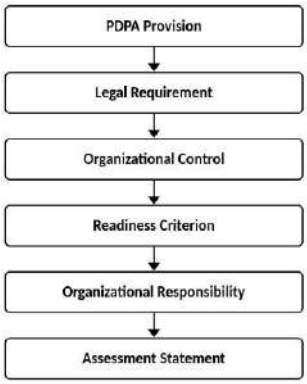



Fig. 1. Legal-to-Control Mapping

Relevant PDPA provisions were first examined to identify requirements applicable to organizational data-processing responsibilities. Each requirement was then expressed as assessable control. Depending on the requirement, the control may represent a technical safeguard, governance responsibility, procedure, accountability mechanism, or expected operational practice. The originating statutory provision is retained throughout the mapping so that each assessment statement remains traceable to its legal basis.

The resulting readiness criterion is then adapted according to responsibility. Executive statements address areas such as governance, oversight, authorization, and resource provision. Middle-management statements focus on implementation, supervision, monitoring, communication, and escalation. Operational statements examine awareness and expected actions during routine handling of personal data. This allows the same underlying requirement to be assessed from different organizational perspectives without changing its regulatory basis.

TABLE I
EXAMPLE OF ROLE-SENSITIVE TRANSLATION OF PDPA REQUIREMENTS

| **Regulatory area** | **Executive perspective** | **Middle-management perspective** | **Operational perspective** |
|---|---|---|---|
| Data-subject request handling | A clear process and support are established. | Requests are handled, monitored, and escalated through the defined process | Employees know how to respond or escalate requests. |
| Personal-data security | Responsibilities and safeguards are established. | Security procedures are followed and monitored. | Employee follows required procedures when handling personal data |

| | | | |
|---|---|---|---|
| Personal-data incident handling | Reporting responsibilities and procedures are established. | Incidents are managed and escalated appropriately. | Employees know how to report suspected incidents. |

TABLE I illustrates the principle of translating a common regulatory area according to organizational responsibility while retaining a shared legal basis. The final assessment inventory maintains the corresponding statutory reference for each statement. For example, a requirement concerning data-subject requests may correspond to a control requiring an established process for receiving and handling such requests. Executive respondents are assessed on whether the process has been established and supported, middle management on whether it is implemented and monitored, and operational respondents on whether they know the appropriate action when a request is received. Fifteen experienced professionals reviewed the assessment statements and provided feedback to improve their clarity and relevance.

*B. Quantification and Scoring Algorithm*

TABLE II
QUANTIFICATION AND SCORING FORMULAS

| | |
|---|---|
| **Individual Response Normalization** | $S_{i,c,m} = \left(\frac{r_{i,c,m} - 1}{5 - 1}\right) \times 100$ |
| **Category Readiness Score per Role ($C_{c,m}$)** | $C_{c,m} = \frac{1}{N_m \cdot K} \sum_{i=1}^{N_m} \sum_{k=1}^{K_c} S_{i,c,k,m}$ |
| **Organizational Role Score ($R_m$)** | $R_m = \frac{1}{C} \sum_{c=1}^{C} C_{c,m}$ |
| **Aggregate Organizational Readiness Index (ORI)** | $\text{ORI} = (\omega_{\text{Exec}} \times R_{\text{Exec}}) + (\omega_{\text{Mid}} \times R_{\text{Mid}}) + (\omega_{\text{Oper}} \times R_{\text{Oper}})$ Where $\omega_{\text{Exec}} = 0.40$, $\omega_{\text{Mid}} = 0.35$, $\omega_{\text{Oper}} = 0.25$ ($\sum\omega = 1.0$) |
| **Cross-Level Discrepancy Index (DSI)** | $\text{DSI} = \frac{1}{C} \sum_{c=1}^{C} \lvert C_{c,\text{Exec}} - C_{c,\text{Oper}} \rvert$ |

TABLE II
SUMMARIZES THE FORMULAS FOR CALCULATING ORGANIZATIONAL READINESS AND CROSS-LEVEL DISCREPANCIES.

Individual Response Normalization:

Each Likert-scale item $r_{i,c,m} \in \{1,2,3,4,5\}$ (where and ) answered by respondent $i$ in regulatory category $c$ for role tier $m$ is normalized to a percentage $S_{i,c,m} \in [0,100]$:

Category Readiness Score per Role ($C_{c,m}$)

For a given regulatory category $c$ (e.g., Data-Subject Requests, Completeness, Erasure) and a specific organizational role $m$ (Exec, Middle, Oper), the score is the arithmetic mean across all $N_m$ respondents and $K_c$ items in that category.

Organizational Role Score ($R_m$):
The total readiness score for a specific organizational tier $m$ across all $C$ categories.
Aggregate Organizational Readiness Index (ORI)

The overall readiness score balances executive governance, middle-management implementation, and operational execution through weighted contribution.

Cross-Level Discrepancy Index (DSI)

To detect communication gaps or misaligned perception across tiers, the Discrepancy Score Index (DSI) calculates the mean absolute difference between Executive perception and Operational execution across all categories.

*C. Maturity Level Mapping*

The computed (ORI) maps directly into five governance maturity bands (TABLE III) and the weights assigned to each organizational role reflect their different responsibilities in maintaining data protection readiness. Executive management is assigned the highest weight (0.40) because it is responsible for strategic direction, governance, and resource allocation. Middle management is assigned a weight of 0.35 because it plays a key role in implementing policies, supervising processes, and monitoring data protection practices. Operational employees are assigned a weight of 0.25 because they are primarily involved in the day-to-day handling of personal data and following established procedures. These weights represent the proposed contribution of each organizational level to the overall readiness score.

| **Score Range (ORI) %** | **Maturity Level** | **Diagnostic Meaning** |
|---|---|---|
| 81 - 100 | Level 5 | Fully institutionalized controls, continuous monitoring, proactive PDPA governance. |
| 61 - 80 | Level 4 | Formal policies & technical controls active across tiers; minor operational gaps. |
| 41 - 60 | Level 3 | Controls documented at executive/middle tier but inconsistently applied at operational level. |
| 21 - 40 | Level 2 | Ad-hoc compliance practices; lack of formal role-specific responsibilities. |
| 0 - 20 | Level 1 | Unaware or non-compliant; no mapped controls or operational procedures. |

TABLE III
PDPA READINESS MATURITY MAPPING AND SCORE CLASSIFICATION

## IV. IMPLEMENTATION AND SYSTEM FEATURES

Each assessment was available in English, Sinhala, and Tamil. The forms collected limited organizational context, including sector, organization type, employee-size category, organizational level, and an optional organization code. The

code supported grouping responses from the same organization while maintaining confidentiality.

IBM SPSS Statistics was used to support analysis of the evaluation dataset. The refined assessment logic will be transferred to a Python-based application using Stream lit. The implementation remains lightweight, with emphasis on assessment logic and regulatory traceability rather than unnecessary technical complexity.

### *A. System Architecture*

The proposed architecture organizes the readiness assessment into six connected layers, as shown in Fig. 1. The PDPA Legal Layer provides the statutory basis for the framework using the PDPA of 2022 and its 2025 amendment. Relevant requirements are passed to the Legal-to-Control Mapping Layer, where they are translated into organizational controls, readiness criteria, responsible organizational levels, and assessment statements.

The Role-Based Assessment Layer presents the relevant statements to executive, middle-management, and operational respondents according to their responsibilities. Submitted responses are processed through the Response and Scoring Engine and subsequently examined within the Analytics and Comparison Layer. This supports within-level, cross-level, and contextual analysis. The final Readiness Reporting Layer presents role-level summaries, priority areas, and reportable readiness information. The architecture preserves the connection between the original legal requirement and the resulting assessment output while separating legal interpretation, assessment delivery, analysis, and reporting.

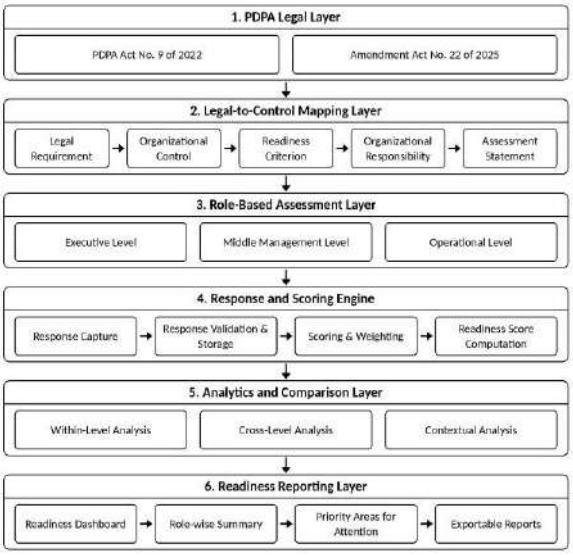



Fig. 2. Proposed architecture of the PDPA readiness assessment system.

### *B. Core System Features*

The system supports role-specific assessment, multilingual presentation, response grouping, comparative analysis, and readiness reporting. A key feature is the ability to accept several responses from the same organizational level rather than relying on one representative. This allows variation among respondents with similar responsibilities to remain visible.

The analytical component examines within-level and cross-level patterns. These comparisons support interpretation of organizational readiness rather than evidence of legal non-compliance.

The application follows a low-disclosure approach. Participants are not required to upload internal policies, audit reports, breach records, screenshots, or other confidential organizational evidence. The resulting output represents reported readiness and should be used as an initial diagnostic rather than a verified compliance determination.

Technology acceptance can be examined after the complete application has been implemented. The Unified Theory of Acceptance and Use of Technology (UTAUT) provides an established basis for evaluating factors such as performance expectancy, effort expectancy, social influence, and facilitating conditions [20] .

## V. EVALUATION AND RESULTS

The framework was evaluated using role-specific Google Form prototypes across three organizations. The evaluation included 17 organizational respondents from executive, middle-management, and operational levels. Additional feedback was obtained from 15 professionals with experience in information systems audit, information security, privacy, and technology governance.

### *A. Evaluation Procedure*

Participants completed the assessment corresponding to their organizational level. Limited contextual information was collected, including sector, organization type, employee-size category, organizational level, and an optional organization code. The code supported grouping responses from the same organization without collecting its name.

The evaluation considered two forms of information. Organizational responses were used to calculate role-level readiness, the Organizational Readiness Index (ORI), maturity levels, and the Cross-Level Discrepancy Index (DSI). Professional feedback was analyzed separately to identify areas requiring clarification or contextual interpretation.

### *B. Analysis*

The evaluation data were analyzed using descriptive statistics and the proposed scoring model. Frequencies and percentages were used to summarize the responses. Role-level readiness scores, the Organizational Readiness Index (ORI), maturity levels, and the Cross-Level Discrepancy Index (DSI) were calculated for each organization. Differences across executive, middle-management, and operational levels were examined to identify variations in reported readiness.

### *C. Organizational Evaluation Results*

The evaluation covered three organizations representing professional services, utilities, and tourism and hospitality. To maintain confidentiality, the organizations are referred to as Organization 1, Organization 2, and Organization 3. Table I summarizes the role-level readiness scores, ORI, maturity level, and DSI.

TABLE IV
ORGANIZATIONAL READINESS EVALUATION RESULTS

| Assessment Criteria | Org. 1: Professional Services | Org. 2: Utilities | Org. 3: Tourism & Hospitality |
|---|---|---|---|
| Respondents (n) | 7 | 7 | 3 |
| Executive (%) | 75.00 | 83.33 | 50.00 |
| Middle (%) | 62.50 | 58.33 | 41.67 |
| Operational (%) | 41.67 | 33.33 | 25.00 |
| ORI (%) | 62.29 | 62.08 | 40.83 |
| Maturity Level | Level 4 | Level 4 | Level 2 |
| DSI (%) | 33.33 | 50.00 | 25.00 |

As shown in table IV, Organization 1 and Organization 2 achieved similar overall readiness scores and were classified at Level 4. However, Organization 2 showed the highest DSI at 50.00%, with a notable difference between executive and operational readiness. Organization 3 recorded the lowest ORI at 40.83% and was classified at Level 2, with lower readiness reported across all three organizational levels.

### *D. Professional fedback*

Fifteen professionals with experience in information systems audit, information security, privacy, and technology governance completed the assessment and provided feedback on the framework. Their feedback supported the role-specific structure and legal traceability of the framework while identifying areas for further refinement. A key concern was that questionnaire responses alone may not provide sufficient context to determine actual PDPA readiness. Professionals noted that readiness and control applicability may vary according to organizational factors such as business processes, personal-data processing activities, regulatory obligations, retention requirements, and third-party or cross-border processing.

The feedback also highlighted the risk of false assurance when responses are converted directly into numerical readiness scores. A positive response to a control may not indicate that the control is sufficient or applicable to a particular organizational context. Professionals therefore recommended incorporating contextual information, Records of Processing Activities (ROPA), and qualitative readiness or risk bands to strengthen interpretation. These findings support the use of the framework as an indicative diagnostic rather than a definitive assessment of legal compliance. The professionals' assessment responses were considered separately from the organizational readiness scores.

## VI. Discussion and Conclusion

The proposed framework approaches PDPA readiness as an organizational issue rather than a single compliance response. Its main value lies in combining legal traceability with responsibility-based assessment. Separating executive, middle-management, and operational perspectives allows the assessment to capture differences that may be concealed when one representative responds for an entire organization.

The framework should not be interpreted as a substitute for a formal compliance audit. Its results are based primarily on reported organizational practices and do not independently verify policies, technical controls, or documentary evidence. Differences between respondents may indicate differences in implementation, awareness, or communication but do not by themselves establish non-compliance. The framework is therefore intended as a diagnostic tool to identify areas requiring further review.

This study establishes the framework, legal-to-control mapping, and prototype assessment process for evaluating PDPA readiness. Evaluation across three organizations from professional services, utilities, and tourism and hospitality showed differences in role-level and overall readiness. Similar overall scores also reflected different patterns across organizational levels, supporting the value of role-sensitive assessment. Professional feedback highlighted the need for contextual interpretation of numerical results. Future research can validate the framework with larger samples across sectors and explore sector-specific readiness factors and contextual risk indicators.

## References


[1] Parliament of the Democratic Socialist Republic of Sri Lanka, *PERSONAL DATA PROTECTION ACT, No. 9 OF 2022*. Sri Lanka, 2022, pp. 1–64. Accessed: Oct. 01, 2026. [Online]. Available: https://www.parliament.lk/uploads/acts/gbills/english/6242.pdf

[2] Parliament of the Democratic Socialist Republic of Sri Lanka, *PERSONAL DATA PROTECTION (AMENDMENT) ACT, No. 22 OF 2025*. Sri Lanka, 2025, pp. 1–12. Accessed: Oct. 01, 2026. [Online]. Available: https://www.parliament.lk/uploads/acts/gbills/english/6384.pdf

[3] S. Sirur, J. R. C. Nurse, and H. Webb, "Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR)," in *Proceedings of the ACM Conference on Computer and Communications Security*, Association for Computing Machinery, Oct. 2018, pp. 88–95. doi: 10.1145/3267357.3267368.

[4] S. Wairimu, L. H. Iwaya, L. Fritsch, and S. Lindskog, "On the Evaluation of Privacy Impact Assessment and Privacy Risk Assessment Methodologies: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 19625–19650, 2024, doi: 10.1109/ACCESS.2024.3360864.

[5] S. Agarwal, S. Steyskal, F. Antunovic, and S. Kirrane, "Legislative Compliance Assessment: Framework, Model and GDPR Instantiation," in *6th Annual Privacy Forum, APF 2018, Barcelona, Spain, June 13–14, 2018, Revised Selected Papers*, Barcelona, Spain: Springer, 2018, pp. 131–149. doi: 10.1007/978-3-030-02547-2_8.

[6] P. Chatsuwan, T. Phromma, N. Surasvadi, and S. Thajchayapong, "Personal data protection compliance assessment: A privacy policy scoring approach and empirical evidence from Thailand's SMEs," *Heliyon*, vol. 9, no. 10, p. e20648, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20648.

[7] R. N. Zaeem and K. S. Barber, "The Effect of the GDPR on Privacy Policies," *ACM Trans. Manag. Inf. Syst.*, vol. 12, no. 1, Mar. 2021, doi: 10.1145/3389685.

[8] S. Ngobeni *et al.*, "Towards a GDPR Compliance Assessment Toolkit," in *Proceedings of the 23rd European Conference on Cyber Warfare and Security (ECCWS)*, Academic Conferences and Publishing International Limited, 2024, pp. 313–321. doi: 10.34190/eccws.23.1.2278.

[9] M. Brodin, “A Framework for GDPR Compliance for Small- and Medium-Sized Enterprises,” *European Journal for Security Research*, vol. 4, no. 2, pp. 243–264, Oct. 2019, doi: 10.1007/s41125-019-00042-z.

[10] N. Nnko *et al.*, “Assessing the organisational readiness and compliance with the personal data protection legislation in Tanzania,” *Telematics and Informatics Reports*, vol. 21, Mar. 2026, doi: 10.1016/j.teler.2026.100299.

[11] F. Ndichu and A. Wausi, “The Assessment of Organizations Readiness to Comply with the Kenya Data Protection Act: A Case of a Humanitarian Arganization,” *International Journal of Computer and Information Technology*, vol. 14, no. 2, pp. 52–56, 2025, doi: 10.24203/xvcq8z18.

[12] N. I. N. Agusti and N. A. Rakhmawati, “A Preliminary Governance-Centered Readiness Framework for Personal Data Protection Law Implementation: Evidence from Indonesia,” in *2026 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2026*, Institute of Electrical and Electronics Engineers Inc., 2026, pp. 432–436. doi: 10.1109/ICETSIS68266.2026.11549181.

[13] Paul. Cunningham and Miriam. Cunningham, *2017 IST-Africa Week Conference: 31 May-02 June 2017, Windhoek, Namibia*. IIMC International Information Management Corporation Ltd., 2017.

[14] R. R. Wijaya, F. A. Putri, and S. Karya, “Measuring Employee Information Security Awareness in Personal Data Protection Compliance,” in *2025 1st International Conference on Emerging Trends in Information Systems and Informatics, ICETISI 2025*, Jakarta, Indonesia: Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICETISI67983.2025.11405918.

[15] Madusanka Chamara Mithrananda, “ANALYZE THE INSTITUTIONAL READINESS ON IMPLEMENTING PERSONAL DATA PROTECTION REGULATIONS, WITHIN PRIVATE SECTOR HIGHER ACADEMIC INSTITUTES IN SRI LANKA, A STUDY OF DATA PROTECTION DRAFT 2019,” Sri Lanka.

[16] P. A. I. Ranathunga and A. P. R. Wickramarachchi, “General Data Protection Regulation (GDPR) Adoption in Sri Lankan Businesses: A Data Governance Model,” in *ICARC 2022 - 2nd International Conference on Advanced Research in Computing: Towards a Digitally Empowered Society*, Sri Lanka: Institute of Electrical and Electronics Engineers Inc., Feb. 2022, pp. 266–271. doi: 10.1109/ICARC54489.2022.9753790.

[17] A. R. Hevner, S. T. March, J. Park, and S. Ram, “DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH 1,” 2004.

[18] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.

[19] Daniel. Amyot, A. I. Antón, Travis. Breaux, Aaron. Massey, and Peter. Swire, *2013 6th International Workshop on Requirements Engineering and Law : RELAW: proceedings: July 16, 2013: Rio de Janeiro, Brasil ; [held in conjunction with the 21st IEEE International Requirements Engineering Conference]*. IEEE, 2013.

[20] A. M. Momani, “The unified theory of acceptance and use of technology: A new approach in technology acceptance,” *International Journal of Sociotechnology and Knowledge Development*, vol. 12, no. 3, pp. 79–98, Jul. 2020, doi: 10.4018/IJSKD.2020070105.

# Data Privacy Awareness Among Sri Lankan Higher Education Students: Comparing Student Perception with the Documented Data Governance Practices of Generative AI Platforms

Jayani Adikari
*Department of Computer and Data Science*
*NSBM Green University*
*Homagama, Sri Lanka*
*0009-0001-5890-1288*

M. C. Mithrananda
*Department of Software Engineering & Computer Security*
*NSBM Green University*
*Homagama, Sri Lanka*
*0009-0003-5206-6984*

***Abstract*—Generative AI tools are now embedded in student routines, yet what platforms do with submitted content is governed by policies that are seldom read. This study measures the gap between what Sri Lankan higher education students believe about AI data handling and what platform policies document. A bilingual cross-sectional survey of 56 students recorded awareness, perceived risk and reported disclosure across eight matched information categories, alongside a structured audit of five widely used platforms across eight governance dimensions. The audit found that all five train on consumer data by default that human review is documented, that deletion does not correspond to removal and that paid consumer tiers confer no additional protection. Only 12.5% of respondents identified both training use and human review, 33.9% believed AI tools as confidential as a doctor and 80.4% remained unprotected by available opt-out settings. A further 64.3% reported disclosing information they themselves rated high risk, most commonly citing utility, habit or the absence of a human alternative. Awareness of the Personal Data Protection Act No. 9 of 2022 was minimal. The gap is specific rather than general, students rank sensitivity soundly but misunderstand backend processing and no platform acknowledges Sri Lankan law.**

***Keywords—privacy awareness, generative AI, data governance, information sensitivity, opt-out settings***

## I. Introduction

Generative artificial intelligence tools have become embedded in the daily routines of university students worldwide[1],[2]. Systems such as ChatGPT, Gemini, Claude, Copilot and Meta AI are now used to draft assignments, generate and debug code, seek career guidance and increasingly to discuss personal difficulties and interpret medical documents[3]. Adoption of generative AI tools among Sri Lankan undergraduates is already well established, with surveys across state universities reporting widespread use of these tools for academic work[4]. Research in the Sri Lankan context, however has addressed patterns of use rather than privacy. Studies of students in adjacent settings have found that disclosure on social media decreases as privacy concern rises and increases with perceived benefit[5] and that high levels of personal disclosure coexist with limited use of available privacy settings[6]. Whether these patterns extend to generative AI has not been examined.

The tasks for which these tools are used differ considerably in sensitivity, yet the interface remains identical and users receive no indication that the handling of their data differs. What becomes of submitted content is determined by the provider's privacy policy, and such policies are lengthy, technically worded and seldom read, McDonald and Cranor estimated that reading those encountered by a typical user would require approximately 201 hours annually[7]. Users therefore accept terms they have not examined, and their understanding derives from assumption rather than documentation. The distance between what students believe and what providers document constitutes an awareness reality gap and it determines the extent to which students expose themselves to risks they have not recognized. Two studies have examined awareness in relation to generative AI [1], [2] though existing work does not compare the awareness recorded against the documented practices of the platforms concerned. This compounds a wider limitation, published work on AI awareness clusters within a small number of national contexts, with emerging economies substantially underrepresented[8]. Sri Lanka has enacted the Personal Data Protection Act No. 9 of 2022[9], the first comprehensive data protection statute in South Asia, yet no published study has established whether students are aware of its existence. This paper addresses these gaps. A survey of Sri Lankan higher education students measured awareness of AI data handling and reported disclosure behaviour, alongside an independent audit of the published privacy policies of five widely used AI platforms across eight governance dimensions. By measuring perception and reported practice across an identical set of information categories, the resulting gap can be quantified rather than assumed. This context is specific in three respects, no prior study compares awareness against documented practice across matched categories, Sri Lankan users receive the weakest default terms available and PDPA awareness has not previously been measured.

## II. Literature Review

### A. Awareness of AI Data Practices

Research on awareness of artificial intelligence has expanded rapidly, though its distribution is uneven. A systematic review found that published work clusters within a small number of national contexts, with consumers and emerging economies substantially underrepresented[8]. This limits the extent to which existing findings apply to student populations in South Asia. Two studies have examined awareness in relation to generative AI. Dahabiyeh et al. surveyed university students on their awareness of four dimensions of the ChatGPT privacy policy, namely data collection, data usage, data disclosure and privacy-enhancing features[2]. Wang et al. interviewed university students and identified optimism bias and peer influence as significant factors shaping privacy concerns[1]. Both establish that student awareness is an appropriate object of measurement, but neither compares the awareness recorded against the documented practices of the platforms concerned. In the Sri Lankan context, existing research has addressed the adoption of generative AI rather than its privacy implications [4]. Work on students in adjacent settings has examined privacy more directly disclosure on social media decreases as privacy concern rises and increases with perceived benefit[5], and high levels of personal disclosure have been documented alongside limited use of available privacy settings[6]. Awareness of the Personal Data Protection Act No. 9 of 2022 has not to the authors' knowledge, been measured empirically.

### B. Disclosure and Information Sensitivity

Milne et al. rated fifty-two information types on perceived risk alongside willingness to provide, producing a typology of information sensitivity[10]. Schomakers et al. applied a comparable approach in a German sample, retaining a substantial proportion of the original information types so that results could be compared across national contexts[11]. This paired structure, in which the same categories are rated for sensitivity and for disclosure, provides the methodological basis for the present instrument. Norberg et al. demonstrated that stated disclosure intentions substantially understate actual disclosure[12]. In the context of conversational AI, Kelly et al. found that willingness to disclose declined as the sensitivity of health information increased, though disclosure was not eliminated[3], supporting the treatment of sensitivity as a graded rather than binary property.

### C. Explaining Disclosure and Auditing Practice

Several frameworks explain continued disclosure under perceived risk. The privacy calculus model holds that disclosure follows a trade-off in which perceived benefit and trust offset perceived risk[13], while privacy cynicism describes a condition in which users perceive protection as futile and cease to attempt it[14]. Reviews of the wider phenomenon report that divergence between attitude and behaviour is most pronounced in hypothetical settings and less evident where concrete behaviour is measured[15], [16]. This literature has attracted substantive criticism. Solove argues that inferring general privacy values from narrow, context-specific decisions constitutes a logical error [17]. Colnago et al. demonstrated that misalignment can operate in the opposite direction [18] and Dienlin and Trepte showed that apparent inconsistency diminishes when attitudes and behaviour are measured at equivalent specificity[19]. The present study responds by measuring perception and reported disclosure across an identical set of categories. Establishing what platforms actually do requires examining their published policies, a task users are unlikely to undertake. McDonald and Cranor estimated that reading the policies encountered by a typical user would require approximately 201 hours annually[7], which provides the rationale for the audit component of this study.

## III. Methodology

### A. Methods

This study examined Sri Lankan higher education students' privacy awareness in generative AI use through a quantitative method based on a questionnaire survey with a cross-sectional time frame complemented by a structured audit of platform privacy policies. Data were collected through an online questionnaire distributed via convenience and snowball sampling, yielding 56 valid responses. The two-phase design addresses three objectives: to measure student awareness of AI data handling to establish the documented practices of widely used platforms, and to evaluate the gap between the two. A survey records what students believe but cannot establish whether those beliefs are correct. Since privacy policies are lengthy and effectively unread in practice [7] an independent audit is required to establish what platforms do.

### B. Survey Design and Instrument

The instrument was developed in stages. Published studies were reviewed to establish how comparable research had been structured, instruments measuring information privacy concern [20], information sensitivity and disclosure[10],[11]and privacy awareness in generative AI[2] informed the choice of constructs and the wording of items, while prior work on Sri Lankan student populations [4],[5],[6] informed the framing of demographic and usage items. Each item was written so that it could be mapped to a defined analysis. The draft was tested with a small group, items producing confusion were revised, and early responses were reviewed before wider circulation. The instrument comprises seventeen questions in four sections, two of which use matrix formats, producing twenty-eight measured items. Section A records respondent profile and frequency of AI use across five task types of increasing sensitivity, establishing the exposure baseline for the later analysis; prior work shows disclosure declines as sensitivity rises[3]. Section B measures awareness of how platforms handle user inputs, perceived risk across eight information categories, and awareness of the Personal Data Protection Act No. 9 of 2022 [9]. Following Dahabiyeh et al., who measured awareness across four policy dimensions rather than as a single construct[2], these items are written against the audit dimensions so that each belief has a documented counterpart wording draws on the IUIPC scale [20]. Section C records reported disclosure across the same eight categories together with protective behaviours. Section D records disclosure under perceived risk and the reasons given, with response options reflecting two competing explanations: that disclosure follows a benefit–risk trade-off [13] and that users disengage from protection because they consider it futile [14]. Items were derived either from published instruments or

written directly against the audit dimensions, which supports content validity. Items that produced confusion in the pilot were reworded. Reliability coefficients are not reported, the eight categories are intended to be distinct rather than interchangeable measures of one construct, so internal consistency is not an appropriate criterion.

### C. Survey Measurements

Perceived risk is measured on a five-point scale from very high risk to no risk at all across eight information categories. Reported disclosure is a multiple-selection item across the same categories. Frequency items use five- and four-point ordinal scales and trust is measured through agreement with a single statement. Rating the same categories twice follows established practice[10],[11]. The pairing is methodologically necessary: Dienlin and Trepte found that apparent inconsistency between attitudes and behaviour largely disappears once both are measured at the same specificity[19], and Solove argued that inferring general privacy values from narrow behavioural decisions constitutes a logical error[17]. Measuring both across identical categories ensures any difference reflects the gap under study rather than a mismatch in measurement.

### D. Data Collection and Participants

Participants were higher education students in Sri Lanka. Students across all disciplines were eligible, with field of study recorded as a segmentation variable rather than an inclusion criterion. Convenience sampling has been used in comparable studies of Sri Lankan undergraduates[5]. Sri Lanka is a multilingual country in which higher education is delivered in more than one language and English proficiency varies between institutions and disciplines. The instrument was therefore administered in both English and Sinhala with respondents directed to their preferred version through conditional branching. Technical terms without established Sinhala equivalents were retained in English reflecting how these terms are used in practice.

### E. Platform Audit

The second phase applied documentary content analysis to the published privacy documentation of five widely used generative AI platforms - ChatGPT, Gemini, Claude, Microsoft Copilot and Meta AI [21],[22],[23],[24],[25]. Documentary content analysis is a qualitative method in which existing documents are examined against categories defined in advance, allowing consistent comparison across sources. The platforms were selected based on prevalence of use within the target population. The wider ecosystem of AI tools, including voice assistants and image generators was not audited, this constraint is noted as a limitation. Each platform was assessed across eight dimensions: data collection, data retention, model training, human review, third-party sharing, user controls, regional coverage and differences between free and paid tiers. These dimensions functioned as pre-determined coding categories, applied consistently to each platform. Sources were the official privacy policies, terms of service and support documentation published by each provider. Where provider statements were ambiguous or diverged from independent reporting, the discrepancy was recorded rather than resolved, since ambiguity is itself relevant to what users can reasonably know. A dimension was recorded as documented only where the provider's own documentation stated the practice explicitly. Where wording was conditional or unclear, the dimension was treated as qualified and reported in the text rather than in Table I, which is why four of the eight dimensions appear only in narrative form. Policies were consulted on the dates given in the references. Coding was carried out by a single coder, so no inter-rater agreement is reported.

### F. Analysis and Ethics

Responses from the English and Sinhala versions were merged before analysis and risk ratings reverse-coded so that higher values indicate higher perceived risk. Frequencies and percentages are reported for all items. Awareness items are compared against the corresponding audit findings to establish the proportion of respondents holding accurate and inaccurate beliefs. For each of the eight categories the mean risk rating and the percentage reporting disclosure are presented together allowing the two to be compared category by category. Disclosure of categories rated as high risk is additionally reported at respondent level. Group comparisons by subscription status, field of study and self-rated familiarity are reported descriptively as the subgroup sizes do not support inferential testing. Ethical approval was obtained prior to data collection. Participation was voluntary and anonymous.

## IV. Platform Audit Findings

### A. Overview

The audit examined five generative AI platforms across the eight dimensions defined in the methodology. The platforms differ in interface, capability and market position but converge closely in how they treat consumer data. Table I summarizes the position of each platform across four dimensions on which practice can be stated categorically. Data collection, retention, third-party sharing and regional treatment require qualification and are reported in the text.

TABLE I. DATA GOVERNANCE PRACTICES OF FIVE GENERATIVE AI PLATFORMS

| Platform | Trains on consumer data by default | Opt-out available | Human review of content | Stronger defaults for EU users |
|---|---|---|---|---|
| ChatGPT | Yes | Yes | Yes | Yes |
| Gemini | Yes | Yes | Yes | Yes |
| Claude | Yes [a] | Yes | Yes [b] | Yes |
| Copilot | Yes [c] | Yes | Yes | Yes |
| Meta AI | Yes | No [d] | Yes | Yes |

[a] Since October 2025. [b] Safety-flagged content. [c] Signed-in consumer accounts. [d] Not available outside the EU and a small number of other jurisdictions.

### B. Data Collection and Retention

All five platforms collect submitted content including text prompts, uploaded files and images, together with account and device information. The scope differs materially, Gemini operates within the wider Google account and where the relevant features are enabled, can draw on Gmail, Drive, Calendar and Photos[22], while Meta AI is informed by public posts, photographs and comments across the Meta platform family[25]. Retention practices differ in duration but share a common feature deletion by the user does not correspond to

immediate removal. Content deleted from ChatGPT remains on provider systems for approximately thirty days[21]. Gemini retains conversations for eighteen months by default and those subject to human review for up to three years independently of the user's deletion of activity [22]. Claude retains content for thirty days where training has been disabled and substantially longer where it has not [23]. Copilot retains chat history for eighteen months[24]. Meta publishes no retention period specific to its assistant.

### C. Model Training, User Controls and Human Review

All five platforms use consumer conversations to improve their models by default. The user is enrolled automatically and must act to withdraw. Opt-out settings exist on all five but sit within sub-menus and are not labelled in terms that clearly indicate their privacy function. Anthropic's position is of note, Claude previously did not use consumer conversations for training, but consumer accounts have been enrolled by default since late 2025 unless the setting is disabled[23]. Meta AI is the least accommodating, offering a formal objection mechanism only in the European Union and a small number of other jurisdictions[25]. Human review occurs in some form across all five platforms. Google states explicitly that trained reviewers may read conversations that such conversations are retained separately from the user account and advises users not to enter information they would not wish a reviewer to see[22]. The other providers describe review for safety, quality or product improvement purposes [21],[23],[24],[25].

### D. Sharing, Regional Coverage and Tier Differences

None of the providers states that personal data is sold though data may be shared with sub processors, affiliates and where legally required authorities. Meta represents the clearest departure from December 2025, content submitted to Meta AI informs advertising personalization across the Meta platform family[25]. Users in the European Economic Area, the United Kingdom and Switzerland receive stronger default protection across all five platforms because of regulatory obligation rather than provider preference. Users elsewhere, including Sri Lanka are governed by general terms under which training is enabled by default. No platform references Sri Lanka or the Personal Data Protection Act No. 9 of 2022[9], whose substantive provisions had not been brought into force at the time of the audit. Paid consumer subscriptions do not alter this position. ChatGPT Plus, Gemini Advanced, Claude Pro and Copilot Pro operate under the same defaults as their free counterparts the distinction governing training use is consumer against enterprise, not free against paid [21],[22],[23],[24].

## V. Results

### A. Respondent Profile

The sample comprised 56 valid responses drawn from private, government and semi-government higher education institutions. Computing and related disciplines accounted for 53.6% of respondents and 53.6% were in their final year of study while 57.1% used free versions of AI tools only and 41.1% had paid or shared access. Over half the sample consisted of computing students in their final year of study, a group that would be expected to be more familiar with digital systems and data handling than students in other fields. The awareness levels reported below therefore describe a comparatively well-placed group rather than the student population as a whole. Whether awareness is lower in other disciplines or earlier years was not tested.

### B. Usage Across Task Sensitivity

Regular use defined as daily or weekly, was reported by 95% of respondents for academic work, 73% for social media and creative tasks, 66% for guidance and career advice, 39% for well-being and mental health discussions and 27% for interpreting medical reports. Usage declines steadily as sensitivity rises, indicating that respondents exercise some caution. The audit establishes that this caution is not reciprocated, a mental health conversation is processed on identical terms to a request to edit a social media caption and no platform applies reduced handling to sensitive content. The 39% and 27% who report regularly submitting well-being and medical content are therefore exposed on the same basis as casual users.

### C. Awareness Against Documented Practice

TABLE I. STUDENT AWARENESS AGAINST AUDIT FINDINGS

| Audit finding (Section IV) | Student response | % |
|---|---|---|
| All five platforms train on consumer data by default | Identified training use | 26.8 |
| Human review documented on all five | Identified human review | 23.2 |
| Both practices documented | Identified both | 12.5 |
| | Identified neither | 62.5 |
| Nothing is deleted immediately, retention 30 days to 3 years | Believed deleted immediately | 17.9 |
| | Believed stored for personal history only | 50.0 |
| Human review documented, one provider warns against sensitive input | Believed AI as confidential as a doctor | 33.9 |
| Opt-out exists on all five but is buried in sub-menus | Did not know it existed | 35.7 |
| | Aware but had not acted | 33.9 |
| | Aware and had disabled it | 19.6 |
| No platform references the PDPA, provisions not in force | Never heard of the Act | 46.4 |
| | Knew only the name | 33.9 |
| | Knew what it covers | 16.1 |

Percentages of all respondents. Blank cells indicate continuation of the audit finding in the row above.

Three comparisons stand out. First, the most common belief in the sample that conversations are stored only for personal chat history and are inaccessible to others is contradicted by the audit on both of its claims. Second, the belief that AI tools offer doctor-equivalent confidentiality is contradicted absolutely rather than partially, platform terms expressly permit third-party review, whereas doctor patient confidentiality excludes it. Third, 80.4% had not disabled the opt-out with unawareness and inaction accounting for most of this.

### D. Perceived Risk Against Reported Disclosure

TABLE II. PERCEIVED RISK AND REPORTED DISCLOSURE BY CATEGORY

| Category | Mean risk | Shared |
|---|---|---|
| Financial information | 4.04 | 11% |
| Someone else's information | 3.95 | 21% |
| Personal identity information | 3.77 | 46% |
| Emotional / well-being | 3.73 | 23% |
| Photos or videos | 3.68 | 68% |
| Medical / health information | 3.64 | 39% |
| Public or creative content | 2.48 | 75% |
| Academic information | 2.34 | 95% |

Risk coded 1–5, where 5 = very high risk. Disclosure expressed as the percentage of all respondents reporting they had entered that category of information. Higher risk ratings are expected to correspond to lower disclosure.

Table III shows that the two measures generally move in opposite directions, the categories rated most risky are those shared least and the categories rated least risky are those shared most. Students ordering of sensitivity is broadly sound. At individual level the position differs 64.3% reported disclosing at least one category they personally rated as high or very high risk. Respondents reported disclosing a mean of 3.79 of the eight categories and 51.8% reported four or more. Two categories diverge from the pattern and the audit explains why both matter. Photographs and videos were rated high risk by 57% yet shared by 68% the only sensitive category where disclosure exceeds perceived risk. Images may contain biometric data which the PDPA treats as a special category. Medical information received the lowest mean risk rating of the six sensitive categories, below financial information, despite health data also constituting a special category and receiving no distinct handling from any platform audited.

### E. Protective Behavior and Stated Reasons

TABLE III. PROTECTIVE BEHAVIOUR AND REASONS FOR DISCLOSURE

| Item | % |
|---|---|
| Rarely or never remove personal details | 39.3 |
| Always remove personal details | 14.3 |
| Never read privacy policies | 37.5 |
| Share despite perceived risk, sometimes or more often | 62.5 |

Percentages of all respondents. Reasons for sharing were a multiple-selection item, so percentages do not sum to 100.

Table IV reports protective behavior. Reasons given for sharing were instant answers under time pressure (75.0%) guidance too useful to forgo (57.1%), habitual use displacing risk assessment (37.5%), less embarrassment than asking a person (35.7%), no accessible human expert (32.1%), trust in the platform (23.2%), and the view that one's own data is unimportant (12.5%). Respondents could select more than one reason. The audit establishes that removing personal identifiers is the only protective measure effective irrespective of platform settings, region or subscription tier. It is the measure most neglected 39.3% rarely or never apply it. Policy reading shows a comparable divergence. Although 44.6% claimed to read policies somewhat or very carefully only 26.8% identified training use a practice stated in every policy examined. This confirms the rationale for the audit McDonald and Cranor estimated that reading the policies a typical user encounters would require approximately 201 hours annually[7] and students cannot realistically establish these facts unaided.

### F. Group Differences

The audit established that paid consumer tiers carry identical privacy defaults to free tiers, with only enterprise agreements excluding training use. Respondents with paid or shared access (n = 23) nonetheless reported disclosing more categories on average than free-tier users (n = 33), 4.26 against 3.45, while showing comparable awareness of training use, 30% against 24%. Computing students identified training use more often than others, 40% against 12%, though most in both groups did not. Self-rated familiarity was a weak indicator of actual knowledge, 32% against 24%. Given the subgroup sizes these comparisons are descriptive.

## VI. Discussion

### A. The Nature of the Gap

The results locate the awareness reality gap precisely. Students do not misunderstand privacy in general, their ranking of information sensitivity is consistent with established typologies [10], [11]. What they misunderstand is what occurs after content is submitted. Each of the three practices the audit documents across all five platforms training by default, human review and retention beyond deletion was identified by roughly a quarter of respondents or fewer. In this sample the gap is therefore produced by invisibility rather than by carelessness. Nothing in the interface signals that a well-being conversation becomes training data and the audit found the controls that would prevent this located in sub menus and labelled in terms that do not indicate their function. Information alone is not sufficient. That 33.9% knew of the opt-out setting and had not used it is as significant as the 35.7% who did not know it existed. For the former group the barrier is not knowledge, whether the explanation is the effort required resignation[14] or a considered judgement that benefit outweighs risk[13], awareness measures alone would not change their position. The same pattern appears in the disclosure data. That 64.3% reported sharing a category they themselves rated high risk is consistent with the intention behavior divergence demonstrated by Norberg et al.[12], and the stated reasons argue against reading this as indifference: utility under time pressure dominates at 75.0% while only one respondent in eight considered their data unimportant. Dahabiyeh et al. measured awareness without a documented counterpart [2] and comparing the two shows that self-reported awareness overstates accuracy.

### B. Sensitive Use Without Alternatives

Health and emotional disclosure attract the lowest usage rates, yet 39% and 27% respectively engage regularly and 32.1% of those disclosing despite perceived risk cited having no accessible human expert. For this group, AI use in sensitive domains constitutes substitution rather than convenience. These are precisely the interactions the audit shows to receive no distinct treatment. Under the PDPA, health data would attract special-category protection, that protection is not yet in force and no platform audited acknowledges the Act. The risk is

concentrated where the consequences are most personal and where, for a substantial minority, no alternative exists.

### C. Implications

Respondents with paid access reported disclosing more while receiving no additional protection suggesting an assumption that payment confers privacy benefit. The audit shows that assumption to be unfounded since the distinction governing training use is consumer against enterprise, not free against paid. The pattern is exploratory at this sample size but identifies a testable hypothesis for future work. More broadly, interventions framed as general privacy awareness are unlikely to close a gap that lies specifically in backend processing, particularly since nearly two-thirds of those already possessing that knowledge do not act on it. The weak relationship between self-rated familiarity and actual knowledge cautions against evaluating any intervention through self-report. Finally, the regional position is structural, Sri Lankan users receive the weakest default treatment available, under a law that is unknown to most, not yet operative and unacknowledged by any provider. Individual behavior cannot correct that position.

## VII. Conclusion

This study measured the gap between what Sri Lankan higher education students believe generative AI platforms do with their data and what those platforms published policies establish, comparing a bilingual survey of 56 students against a structured audit of five platforms across eight governance dimensions. The gap is specific rather than general. Students rank information sensitivity soundly and moderate their use as tasks become more sensitive. What they do not know is what happens after submission only 12.5% identified both training use and human review 33.9% believed AI tools as confidential as a doctor and 80.4% had not disabled opt-out settings, inaction being almost as common as unawareness. Meanwhile 64.3% reported disclosing information they themselves rated high risk, most commonly citing utility, habit or the absence of human alternatives and 80.3% did not know what the Personal Data Protection Act covers. Several limitations apply. The sample of 56 obtained through convenience and snowball sampling is skewed towards private institutions, computing disciplines and final-year students and subgroup comparisons are descriptive only. Disclosure was self-reported and is likely understated. The audit covered five platforms rather than the wider AI ecosystem and two audited dimensions data collection and third-party sharing had no corresponding survey measure. Minor wording differences between the language versions of two items are acknowledged and platform policies change frequently. Future work should test the payment assurance relationship with a larger sample extend the audit to specialized well-being and medical applications and examine whether interventions addressing backend processing specifically produce protective behavior where general awareness does not. As generative AI becomes a default source of guidance for students without access to human alternatives the distance between what they believe and what platforms do constitutes a daily and unchosen exposure.

## References


[1] N. Wang, Y. Li, and F. Cong, "University Students' Privacy Concerns Towards Generative Artificial Intelligence," *J. Acad. Ethics*, vol. 23, no. 4, pp. 2401–2422, 2025.

[2] L. Dahabiyeh, N. Taha, M. Thneibat, and M. A. Bhat, "Privacy awareness in generative AI: the case of ChatGPT," *Interactive Technology and Smart Education*, vol.23, no.1, pp.25-48, 2026.

[3] S. Kelly, K. M. White, S.-A. Kaye, and O. Oviedo-Trespalacios, "Tapping into Key Drivers: Self-Disclosure in Sensitive Health Conversations with ChatGPT," *Int. J. Hum. Comput. Interact.*, vol. 41, no. 24, pp. 15668–15678, 2025.

[4] S. Sabraz Nawaz, M. B. Fathima Sanjeetha, G. Al Murshidi, M. I. Mohamed Riyath, F. B. Mat Yamin, and R. Mohamed, "Acceptance of ChatGPT by undergraduates in Sri Lanka: a hybrid approach of SEM-ANN," *Interactive Technology and Smart Education*, vol. 21, no. 4, pp. 546–570, Oct. 2024.

[5] C. Maduragoda, U. S. Senarath, and L. I. Manorathne, "Factors associated with the privacy disclosure on social media by state university students in Sri Lanka," J. Manage. Matters, vol. 10, no. 1, pp. 83–92, 2023.

[6] J. Lavanya and S. Santharooban, "Information disclosure and attitude towards privacy on Facebook among undergraduates: A survey at the Eastern University, Sri Lanka," *Journal of the University Librarians Association of Sri Lanka*, vol. 24, no. 1, pp. 39–56, Apr. 2021.

[7] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *I/S: A Journal of Law and Policy for the Information Society*, vol. 4, no. 3, pp. 543–568, 2008.

[8] Y. Zhu and W. M. H. Wan Hussain, "Artificial intelligence (AI) awareness (2019–2025): A systematic literature review using the SPAR-4-SLR protocol," *Social Sciences & Humanities Open*, vol. 12, p. 101870, 2025.

[9] Parliament of Sri Lanka, "Personal Data Protection Act, No. 9 of 2022," 2022. Accessed: Aug. 12, 2026. [Online]. Available: https://www.parliament.lk

[10] G. R. Milne, G. Pettinico, F. M. Hajjat, and E. Markos, "Information Sensitivity Typology: Mapping the Degree and Type of Risk Consumers Perceive in Personal Data Sharing," *Journal of Consumer Affairs*, vol. 51, no. 1, pp. 133–161, Mar. 2017.

[11] E.-M. Schomakers, C. Lidynia, D. Müllmann, and M. Ziefle, "Internet users' perceptions of information sensitivity – insights from Germany," *Int. J. Inf. Manage.*, vol. 46, pp. 142–150, Jun. 2019.

[12] P. A. Norberg, D. R. Horne, and D. A. Horne, "The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors," *Journal of Consumer Affairs*, vol. 41, no. 1, pp. 100–126, Jun. 2007.

[13] T. Dinev and P. Hart, "An Extended Privacy Calculus Model for E-Commerce Transactions," *Information Systems Research*, vol. 17, no. 1, pp. 61–80, Mar. 2006.

[14] C. Lutz, C. P. Hoffmann, and G. Ranzini, "Data capitalism and the user: An exploration of privacy cynicism in Germany," *New Media Soc.*, vol. 22, no. 7, pp. 1168–1187, Jul. 2020.

[15] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon," *Comput. Secur.*, vol. 64, pp. 122–134, Jan. 2017.

[16] N. Gerber, P. Gerber and M. Volkamer, "Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior," *Comput. Secur*, vol. 77, pp. 226–261, Aug. 2018.

[17] D. J. Solove, "The myth of the privacy paradox," *George Washington Law Review*, vol. 89, no. 1, pp. 1–51, 2021.

[18] J. Colnago, L. Cranor, and A. Acquisti, "Is There a Reverse Privacy Paradox? An Exploratory Analysis of Gaps Between Privacy Perspectives and Privacy-Seeking Behaviors," *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 1, pp. 455–476, Jan. 2023.

[19] T. Dienlin and S. Trepte, "Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviors," *Eur. J. Soc. Psychol.*, vol. 45, no. 3, pp. 285–297, Apr. 2015.

[20] N. K. Malhotra, S. S. Kim, and J. Agarwal, "Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal

Model," *Information Systems Research*, vol. 15, no. 4, pp. 336–355, Dec. 2004.

[21] OpenAI, "Privacy policy." Accessed: Jul. 01, 2026. [Online]. Available: https://openai.com/policies.

[22] Google, "Gemini Apps Privacy Hub." Accessed: Jul. 01, 2026. [Online]. Available: https://support.google.com/gemini/answer/13594961.

[23] Anthropic, "Privacy policy." Accessed: Jul. 01, 2026. [Online]. Available: https://www.anthropic.com/legal/privacy.

[24] Microsoft, "Privacy FAQ for Microsoft Copilot", Accessed: Jul. 01, 2026. [Online]. Available: https://support.microsoft.com/en-us/microsoft-copilot/privacy-faq-for-microsoft-copilot.

[25] Meta,"Privacy policy", Accessed: Jul. 01, 2026. [Online]. Available: https://www.facebook.com/privacy/policy.

# AI-Enabled Secure Digital Support Platform for Women's Safety, Wellbeing, and Empowerment in Sri Lanka

R. G. Risini Senara Wijewickrama
*Faculty of Computing,*
*NSBM Green University*
*Homagama,Sri Lanka*
grswijewickrama@students.nsbm.ac.lk

Hirushi Dilpriya
*Lecturer, Faculty of Computing*
*NSBM Green University*
*Homagama, Sri Lanka*
Hirushi.d@nsbm.ac.lk

***Abstract*—Digital safety tools for women in Sri Lanka address isolated needs such as emergency alerting, cyber harassment reporting or reproductive health, which forces women to juggle several disconnected services during a crisis. This paper presents Suraksha, a prototype AI-enabled platform that brings emergency response, trilingual (Sinhala, Tamil and English) harassment detection, encrypted evidence management, anonymous reporting, AI-assisted legal guidance, mental health screening and an economic empowerment entry point into one role-based system. The work follows the Design Science Research Methodology. Requirements were derived from a literature synthesis, a hybrid classical-to-transformer language processing pipeline was specified for harassment detection, and an architecture with five roles (User, Police, Counselor, Legal Advisor and Admin) linked by a single case reference scheme was designed. A functional prototype of thirty-seven screens was demonstrated through scenario-based walkthroughs, and a structured comparison with existing Sri Lankan and international platforms shows that no reviewed system combines these capabilities. Of ten secondary objectives, six are demonstrated at prototype level, two are partially realized, one is addressed through analysis and one awaits formal evaluation; the economic empowerment module is limited to a single entry point. This paper reports no experimental performance results: classifier accuracy, security strength and usability have not yet been measured, and the per-language figures shown on the monitoring screen are illustrative placeholders. Formal evaluation using per-language precision and recall, an independent security assessment and the System Usability Scale is defined as the next stage of the research cycle.***

***Index Terms*—Women's safety, AI harassment detection, Sinhala-Tamil NLP, encrypted evidence management, Design Science Research**

## I. Introduction

This paper presents *Suraksha*, a prototype AI-enabled platform that brings these facets together. Its contributions are:

(1) a requirements-driven, five-role architecture linked by a single case-reference scheme; (2) a functional prototype of 37 screens, demonstrated through scenario-based walkthroughs;

(3) a structured comparison with existing Sri Lankan and international platforms; and (4) an explicit separation of implemented features from proposed and unvalidated ones, together with a defined evaluation protocol. The primary aim is to design, build and assess an integrated platform; the secondary objectives are: O1, analyse the physical, cyber, social and economic challenges faced by Sri Lankan women; O2, design a multilingual harassment-detection module; O3, implement a secure, concealable evidence vault; O4, provide emergency SOS, safe-route guidance and geo-fenced alerts; O5, provide anonymous reporting; O6, provide AI-assisted mental-health screening and referral; O7, establish a moderated peer-support community; O8, create an educational knowledge hub; O9, provide an employment and skills module; and O10, evaluate usability, security and user acceptance.

**Scope.** This paper reports the design and demonstration stages of the research cycle. It does not report experimental results: classifier performance, security strength and usability have not yet been measured. Section II reviews related work, Section III describes the methodology and evaluation protocol, Section IV the system, Section V the prototype verification and discussion, and Sections VI and VII the limitations and conclusions.

Gender-based violence, inequity at work and cyber harassment continue to limit women's safety and their access to legal and psychological support. As smartphone and social-media use has grown, harm has moved online in the form of cyber-stalking, non-consensual image sharing and cyber-blackmail. In Sri Lanka, the 2019 Women's Wellbeing Survey found that one in five ever-partnered women (20.4%) had experienced physical and/or sexual intimate-partner violence, and most women who experienced violence did not report it to the police [1].

Technology responses include one-touch alert and location-sharing apps such as bSafe and Noonlight, crowdsourced safety scoring such as Safetipin, machine-learning classifiers for cyberbullying text [2] and cognitive-behavioural chatbots such as Woebot. A 2026 survey of women's-safety applications reports that most solutions are limited to reactive features such as SOS alerts, GPS tracking and geofencing, with little support for legal, psychological and economic needs [3]. Sri Lankan tools such as 2six4 [4], Cyber Care [5] and Fio each address one facet of the problem, and, on the comparison in Section II, none combines Sinhala and Tamil harassment detection, encrypted evidence storage, anonymous legal guidance, mental-health support and economic empowerment in one platform.

## II. Related Work

### A. Emergency Response and Physical Safety

Panchard et al. [6] validated mobile-enabled emergency alerting in urban settings, an early foundation for one-touch SOS. Recent prototypes include ANARKI (spy-camera detection, SOS, siren) [7], SakhiAlert [8] and SafeHer (SOS, offline SMS and community alerts) [9]. None links the emergency event to downstream legal, psychological or evidentiary support.

### B. Harassment Detection and Low-Resource NLP

Dinakar et al. [2] showed that Support Vector Machines can classify cyberbullying text, and ReThink [10] blocks harassing messages using NLP; both target English and neither is connected to reporting or evidence preservation. Most harassment against Sri Lankan women occurs in Sinhala and Tamil. Sandaruwan et al. [11] built an early Sinhala hate-speech classifier on about 3,000 labelled comments (Multinomial Naïve Bayes, accuracy 0.92), and Amali and Jayalal [12] reported 94.5% accuracy for a hybrid rule-based/ML Sinhala cyberbullying detector. The SOLD dataset

[13] and Subasa [14] extend this work with transformer models. For Tamil, a 47,692-tweet cyberbullying dataset [15] and a hate-speech API [16] exist. Both languages remain under-resourced relative to English [17], and none of these studies is part of a deployed multi-feature safety system.

### C. Digital Mental-Health and Legal-Assistance Tools

Woebot (CBT) and 7 Cups (anonymous peer support) lower barriers to seeking help but operate separately from reporting and evidence tools. In the legal domain, LEGALBOT [18] handles a narrow task set, retrieval-augmented generation has been used to reduce fabricated answers in court self-help [19], user research documents the demand and risks of legal question answering [20], and AI for victim support is documented internationally [21]. None is adapted to Sri Lankan evidence practice or to Sinhala and Tamil users.

### D. Sri Lankan Applications and Comparison

2six4, developed by Women In Need, sends SOS alerts with location to selected contacts, supports reporting of street harassment to the police and referral of cyber harassment to the organisation, in English, Sinhala and Tamil [4]. Cyber Care, supported by UNDP, offers awareness content, reporting guidance and a resource library on cyber violence [5]. Fio focuses on reproductive health. Table I compares these and other systems with Suraksha using only features reported in the cited sources.

TABLE I
FEATURE COMPARISON WITH EXISTING SYSTEMS

| System | Alert | Det. | Si/Ta | Evid. | Legal | MH | Econ. | Anon. |
|---|---|---|---|---|---|---|---|---|
| 2six4 [4] | ● | – | – | – | P | – | – | – |
| Cyber Care [5] | – | – | – | – | P | – | – | – |
| Fio (reproductive health) | – | – | – | – | – | – | – | – |
| bSafe, Noonlight | ● | – | – | – | – | – | – | – |
| Safetipin | P | – | – | – | – | – | – | – |
| SafeHer [9], SakhiAlert [8] | ● | – | – | – | – | – | – | – |
| Woebot, 7 Cups | – | – | – | – | – | ● | – | P |
| LEGALBOT [18] | – | – | – | – | ● | – | – | – |
| ReThink [10], Dinakar et al. [2] | – | ● | – | – | – | – | – | – |
| Sinhala/Tamil NLP studies [11], [13], [14], [15] | – | ● | ● | – | – | – | – | – |
| **Suraksha (this work)** | ○ | ○ | ○ | ○ | ○ | ○ | P | ○ |

Alert: SOS, location or safe-route alerting; Det.: AI harassment detection; Si/Ta: Sinhala/Tamil support in detection; Evid.: encrypted evidence storage; MH: mental-health support; Econ.: economic empowerment; Anon.: anonymous reporting or participation. ● feature reported in cited source; P partial or information/referral only; ○ prototype only, not validated; – not reported in the cited sources.

The comparison yields four gaps. **G1 (fragmentation):** no reviewed system combines emergency response, harassment detection, encrypted evidence, legal advice, mental-health support and economic empowerment. **G2 (language):** Sinhala/Tamil detection research is not integrated with English detection in one deployed application. **G3 (anonymity versus identifiability):** reviewed systems do not reconcile the anonymity needed for safe reporting with the identifiability needed for emergency dispatch. **G4 (evidence to empowerment):** tools stop at alerting, detecting, venting or informing, and do not continue to legal action, recovery and economic independence.

## III. Methodology

### A. Research Design

Design Science Research [22] suits this study because its main output is a working artefact rather than a tested theory. The Design Science Research Methodology (DSRM) [23] was operationalised through six activities: problem identification, objective definition, design and development, demonstration, evaluation and communication. The problem is both technical and social, so a pragmatic mixed-methods stance [24] is adopted. This paper covers the first four activities and defines the protocol for the fifth.

### B. Participants and Data Collection Plan

Target groups are women across urban, semi-urban and rural districts, police officers, counsellors and NGOs supporting survivors. Stratified purposive sampling (age, district, language, prior harassment experience) is complemented by snowball sampling through partner NGOs to reach rural, Tamil-speaking and lower digital-literacy participants. Planned (not achieved) targets are 150 to 200 survey respondents, 15 to 20 interview or focus-group participants and 30 to 40 usability participants; achieved samples will be reported with the evaluation.

### C. AI Model Development Plan

Text is normalised and tokenised for Sinhala and Tamil script, code-mixing and romanised spelling, and annotated in four classes (non-offensive, offensive, threatening, blackmail/extortion) with inter-annotator agreement measured by Cohen's kappa. Training data are drawn from public corpora [13], [11], translated or code-mixed English examples [2] and consented user-contributed samples. A TF-IDF/embedding classifier [2] serves as an interpretable baseline, to be compared with a fine-tuned lightweight multilingual transformer [25]. The engine is specified and integrated at interface level in the

prototype; trained-model performance has not yet been measured.

*D. Evaluation Protocol (to be executed)*

**Classifier:** precision, recall, F1 and accuracy on a held-out set, reported per language and per class, with the baseline and transformer compared, so that gaps caused by Sinhala and Tamil data scarcity [17] are visible. **Usability and acceptance:** the ten-item System Usability Scale [26] against the benchmark of about 68 [27], task success and time, and acceptance items based on the technology acceptance model [28]. **Effect:** a pre/post design for legal awareness, reporting confidence and help-seeking intention, with six-phase thematic analysis of qualitative feedback [29]. **Security:** an independent assessment of the evidence vault covering access control, key management, integrity verification and tamper tests. **Ethics:** WHO recommendations for violence research [30] and Sri Lanka's Personal Data Protection Act [31], including informed consent, data minimisation and referral to local support services.

## IV. System Design and Implementation

*A. Architecture Overview*

Suraksha provides a common AI-enabled case-management backbone through which Police, Counselors, Legal Advisors and Administrators connect with survivors. Five role-specific interfaces share one design language and one case-reference scheme (for example Case #SL-2291) used consistently across roles. The data model is an Enhanced Entity-Relationship (EER) model in which a generalised User entity specialises into Woman, Police, Counsellor and Admin, with the relationships SUBMITS, TRIGGERS, GENERATES, BOOKS and

RESPONDS_TO. The User app is a mobile application, while the Police, Counselor, Legal Advisor and Admin interfaces are web portals; implementation is ongoing. Table II lists the functional requirements and their status.

*B. User Application*

The app is disguised as an ordinary application (Calculator, Notes or Weather) and opens into "Guardian mode" only with a personal PIN or biometric, so an abuser with physical access to the device is less likely to discover it. Language (Sinhala, Tamil, English) and trusted contacts are set at onboarding. The home dashboard (Fig. 1a) gives fast access to SOS, the Legal Aid chat, the Evidence Vault, a mental-health check-in and a jobs tile. A deliberate hold-to-alert gesture reduces false alarms,

then notifies the nearest police unit and trusted contacts with live location; a Safe Route feature computes walking routes that avoid user-reported danger zones.

Submitted screenshots pass through the harassment-detection module, whose result screen (Fig. 1b) shows a risk class and a confidence value; the "92%" and "threat + blackmail" values in the figure are illustrative mock-up content, not measured output. Results are saved to an evidence vault designed around AES-256 symmetric encryption [32], with a SHA-256 integrity hash computed when an item is attached to a report to support chain of custody (Fig. 1c). Anonymous reporting uses category selection and a structured form with a "submit anonymously" toggle. The AI Legal Aid chat is designed to draw on a curated Sri Lankan legal knowledge base, in line with retrieval-grounded approaches [19], and to route queries to human Legal Advisors; a searchable "Know Your Rights" hub complements it. Wellness support comprises a moderated Community screen and a GAD-7-style check-in explicitly labelled "not a diagnosis".

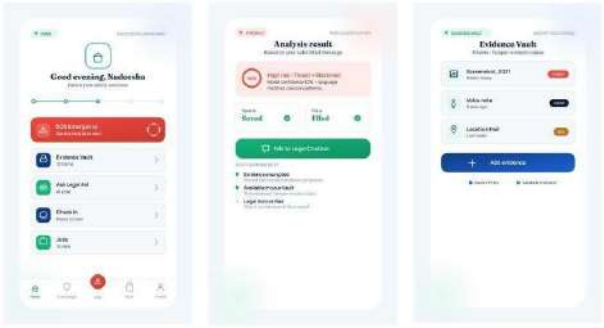



Fig. 1. User application: (a) home dashboard, (b) harassment-detection result (illustrative values), (c) encrypted Evidence Vault with timestamped, tamper-evident items.

*C. Police, Counselor, Legal Advisor and Admin Interfaces*

The Police dashboard has a live SOS map, an alert queue with one-click "Respond" and a case-investigation file (Fig. 2) holding the sealed evidence bundle, reconstructed location trail and an auditable status stepper (Filed, Under investigation, Suspect contacted, Resolved). The Counselor dashboard manages sessions, an anonymised client wellbeing snapshot, session notes and scheduling. The Legal Advisor dashboard queues incoming Legal Aid chat inquiries and administers the public legal-resource library. The Admin dashboard provides user verification, user management, a reports queue, an AI risk-assessment banner, community content moderation and an AI Model Monitoring screen (Fig. 3) showing per-model and per-language coverage and drift status.

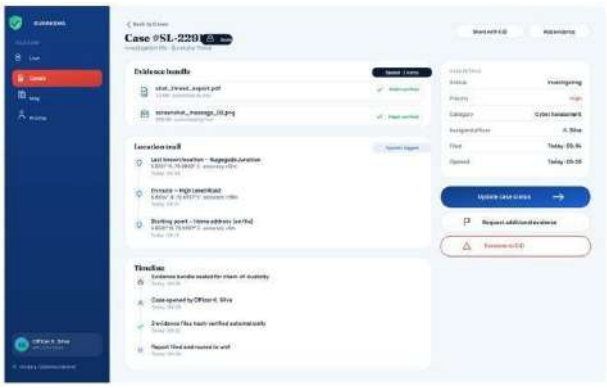



Fig. 2. Police case-investigation file for Case #SL-2291: sealed evidence bundle, location trail and case timeline.

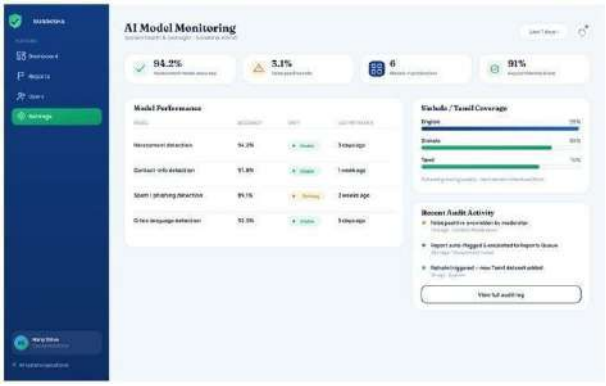



Fig. 3. Admin AI Model Monitoring screen. All accuracy and per-language coverage values (including English 99, Sinhala 82 and Tamil 74) are illustrative placeholders in the mock-up and are not experimental results.

## V. Prototype Verification and Discussion

### A. Status of Requirements and Objectives

TABLE II
Requirements, Objectives and Status

| ID | Functional requirement | Interface | Obj. | Status |
|---|---|---|---|---|
| FR1 | Disguised app icon; PIN/biometric "Guardian mode" | User app | O3 | Demo. |
| FR2 | Live location, Safe Route, hold-to-alert SOS | User app | O4 | Demo. |
| FR3 | Classify Sinhala, Tamil or English text as harassment, threat or blackmail | User app | O2 | Demo. |
| FR4 | Store analyzed evidence in an AES-256 encrypted, SHA-256 hashed vault | User app | O3 | Demo. |
| FR5 | Report incidents anonymously or identified, in structured form | User app | O5 | Demo. |
| FR6 | AI Legal Aid chat, Know Your Rights hub, check-in, counsellor booking, community | User app | O6–O8 | Demo. |
| FR7 | Live SOS map, alert queue, case file and status stepper | Police | O4 | Demo. |
| FR8 | Counselling sessions, wellbeing snapshots, session notes | Counselor | O6 | Demo. |
| FR9 | Legal-resource library; routing of Legal Aid chat queries | Legal Advisor | Main | Demo. |
| FR10 | Verify users, moderate content, monitor per-language AI coverage | Admin | O2, O7 | Demo. |
| EE | Economic empowerment: single "Jobs" tile only; no listing or skills screens | User app | O9 | Partial |
| EV | Formal usability, security and acceptance evaluation | — | O10 | Pending |

Demo.: screens and interaction flow demonstrated in scenario walkthroughs; not empirically validated. Partial: only an entry point exists. Pending: not yet performed.

Of the ten secondary objectives, six (O2, O3, O5, O6, O7, O8) are demonstrated at prototype level, two are partial (O4, because geo-fenced alerts are not yet demonstrated, and O9), O1 is addressed through the literature analysis, and O10 awaits the evaluation stage. Demonstrating a screen shows that a feature exists, not that it works or is effective.

### B. Implemented versus Proposed

**Implemented:** a functional prototype of 37 screens across five roles with a shared case-reference scheme, demonstrated in scenario walkthroughs. **Specified but not validated:** the multilingual detection engine, the cryptographic design of the vault, retrieval-grounded legal assistance and per-language monitoring; no accuracy, security or usability figure is claimed. **Proposed:** job listing and skills development, the trilingual interface on every screen, geo-fenced alerts, offline and error handling, and the full evaluation. The figures shown on the Admin monitoring screen (English 99, Sinhala 82, Tamil 74) and the 92 confidence value on the result screen are mock-up placeholders; no conclusion in this paper depends on them, and they should not be read as evidence of effectiveness. The expectation that Sinhala and Tamil detection will lag English is a hypothesis grounded in data scarcity [17], to be tested.

### C. Comparison with Existing Solutions

As Table I shows, each existing system covers one or two feature clusters. SOS apps stop at alerting; Sinhala and Tamil detection studies stop at classification, whereas in Suraksha classification feeds the Evidence Vault and the reporting process. The Police map and queue extend the mobile emergency alerting validated by Panchard et al. [6] with staff assignment and priority markers for a full dispatch workflow, and booking with a Counselor dashboard is intended to close the loop that Woebot and 7 Cups leave open. A single case reference (Case #SL-2291 flowing from the user's report tracker to the Police case file and the Admin escalation screen) and separate anonymous and identified paths are the design response to gaps G1 and G3. This is design intent only. Existing

Sri Lankan apps such as 2six4 [4] have been used by real users since 2017, whereas Suraksha has not been tested with any.

the economic-empowerment module, extend the trilingual interface to all screens, add error and offline states, and execute the evaluation protocol of Section III-D: per-language precision and recall, an independent security assessment, and System Usability Scale and pre/post evaluation with a diverse pilot group.

## References

[1] Department of Census and Statistics and United Nations Population Fund (UNFPA), *Women's Wellbeing Survey 2019: Findings from Sri Lanka's First Dedicated National Survey on Violence Against Women and Girls*. Colombo, Sri Lanka: Dept. of Census and Statistics, 2019.

[2] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. AAAI Conf. Web and Social Media*, vol. 5, no. *Design-Level Security Analysis*

[3] J. Panchard, P. Papadimitratos, J.-P. Hubaux, and F. Legendre, "Mobile-enabled solutions for emergency alerting in urban environments," in *Proc. ACM Workshop Wireless Networks and Systems for Developing Regions*, 2010.

[4] T. Prabhu, "ReThink: A tool to combat cyberbullying using natural language processing," ReThink Inc., 2016.

[5] P. K. S. Sandaruwan, S. A. S. Lorensuhewa, and M. A. L. Kalyani, "Sinhala hate speech detection in social media using text mining and machine learning," in Proc. 19th Int. Conf. Advances in ICT for Emerging Regions (ICTer), 2019.

[6] W. D. A. S. D. Amali and S. Jayalal, "A hybrid approach for cyberbullying detection in Sinhala social media," in Proc. Int. Conf. Advances in ICT for Emerging Regions (ICTer), 2020.

[7] T. Ranasinghe, I. Anuradha, D. Premasiri, K. Silva, H. Hettiarachchi, L. Uyangodage, and M. Zampieri, "SOLD: Sinhala offensive language dataset," arXiv preprint, 2022.

[8] S. Haturusinghe, T. C. Weerasooriya, M. Zampieri, C. M. Homan, and

S. R. Liyanage, “Subasa: Adapting language models for low-resourced offensive language detection in Sinhala,” in Proc. 2025 Conf. Nations of the Americas Chapter of the Assoc. Comput. Linguistics: Human Language Technologies (Vol. 4: Student Research Workshop), 2025.

[9] “Multi-tier linguistic and emotional modeling for cyberbullying detection in Tamil social media,” Expert Syst. Appl., 2025, in press.

[10] Omdena and DreamSpace, “AI for Tamil hate speech detection.”

[11] N. De Silva, “Sinhala text classification: Observations from the perspective of a resource poor language,” arXiv preprint, 2019.

[12] “LEGALBOT: AI law advisor chatbot,” Int. J. Novel Res. Dev.

[13] A. Ajmi, “Revolutionizing access to justice: The role of AI-powered chatbots and retrieval-augmented generation in legal self-help,” The Brief, vol. 53, no. 2, p. 10, 2024.

[14] “User research report into AI for access to justice: Legal questions and answers,” Stanford Legal Design and Innovation / Justice Innovation, 2024.Organisation for Economic Co-operation and Development (OECD), AI in Justice Administration and Access to Justice: Governing with Artificial Intelligence. Paris, France: OECD, 2025. A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” MIS Q., vol. 28, no. 1, pp. 75–105, 2004.

[15] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” J. Manage. Inf. Syst., vol. 24, no. 3, pp. 45–77, 2007.

[16] J. W. Creswell and V. L. Plano Clark, Designing and Conducting Mixed Methods Research, 3rd ed. Thousand Oaks, CA, USA: SAGE, 2018.

[17] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[18] J. Brooke, “SUS: A ‘quick and dirty’ usability scale,” in Usability Evaluation in Industry, P. W. Jordan et al., Eds. London, U.K.: Taylor & Francis, 1996,

[19] A. Bangor, P. T. Kortum, and J. T. Miller, “An empirical evaluation of the System Usability Scale,” Int. J. Hum.-Comput. Interact., vol. 24, no. 6, pp. 574–594, 2008.

[20] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” MIS Q., vol. 13, no. 3, pp. 319–340, 1989.