NERD PUBLICATION

Editors
Dr. K. Sujatha
Prof. V. M. Venkateswara Rao

Proceedings of International Conference on Intelligent Computing and Applications ICICA-2025

ISBN: 978-81-987483-5-5

NERD PUBLICATION

New Era Research Development Publication
Pune, Maharashtra
www.nerdpublication.com

Foreword

It gives us immense pleasure to welcome you to the *International Conference on Intelligent Computing and Applications (ICICA-2025)*. This conference represents a significant step toward fostering global collaboration and advancing research in intelligent systems, emerging technologies, and their transformative applications across various domains.

The rapid evolution of artificial intelligence, machine learning, data analytics, and smart technologies has opened new horizons for solving complex real-world problems. ICICA-2025 provides an essential platform for researchers and practitioners to present their innovative findings, share diverse perspectives, and engage in meaningful discussions that contribute to the advancement of intelligent computing.

The conference brings together a distinguished group of participants from academia, industry, and research institutions worldwide. Their contributions reflect the growing importance of interdisciplinary research and the need for integrated solutions in areas such as IoT, cybersecurity, cognitive computing, embedded systems, smart cities, and health informatics. By presenting cutting-edge studies, this conference not only highlights current technological achievements but also illuminates the path for future developments.

We extend our deepest appreciation to all authors for their valuable research contributions, to the reviewers for their dedicated evaluation process, and to the keynote speakers and experts who have enriched the conference with their insights. We also acknowledge the tireless efforts of the organizing team whose vision and coordination have made ICICA-2025 a reality.

We believe that the knowledge shared through ICICA-2025 will inspire further research, strengthen academic and professional networks, and contribute meaningfully to the global discourse on intelligent computing.

We warmly welcome all participants and wish you an engaging, productive, and intellectually rewarding experience at ICICA-2025.

Preface

The *International Conference on Intelligent Computing and Applications (ICICA-2025)* stands as a significant forum dedicated to advancing research that shapes the future of intelligent technologies and their real-world impact. As digital transformation accelerates across industries and societies, ICICA-2025 brings together a vibrant community of scholars, innovators, and practitioners committed to exploring the evolving landscape of artificial intelligence, data-driven systems, and smart applications.

Intelligent computing continues to redefine the boundaries of innovation. From breakthroughs in machine learning and deep learning to advances in natural language processing, computer vision, and human–computer interaction, the field is driving new possibilities across sectors. ICICA-2025 serves as a platform to examine how these technologies can solve complex challenges, optimize decision-making processes, and transform the way humans interact with digital systems.

Rapid advancements in computational power, cloud–edge ecosystems, and big data analytics have enabled systems that learn, adapt, and respond to dynamic environments. These emerging capabilities are revolutionizing industries such as healthcare, education, manufacturing, communication, and urban development. ICICA-2025 offers an opportunity to reflect on how intelligent systems can be harnessed not only for efficiency and automation but also for societal well-being, accessibility, and sustainable development.

Health informatics and smart healthcare solutions represent a vital extension of intelligent computing. The integration of AI-driven diagnostics, personalized treatment models, telemedicine, and sensor-based monitoring demonstrates how technology can improve quality of life and transform medical ecosystems. ICICA-2025 highlights these contributions while emphasizing the importance of ethical, secure, and human-centered digital solutions.

Equally important are the contributions from interdisciplinary domains such as educational technology, industrial intelligence, and smart city infrastructures. These areas illuminate how intelligent systems shape learning, productivity, mobility, and environmental sustainability. By understanding the human, organizational, and societal dimensions of technological adoption, researchers can design solutions that are inclusive, responsible, and future-ready.

Environmental and industrial applications also play a critical role. Intelligent sensing technologies, automation systems, and IoT-enabled environments are redefining resource management, manufacturing processes, and large-scale operational decision-making. ICICA-2025 encourages ideas that address global challenges through innovation, collaboration, and scalable technological solutions.

ICICA-2025 is a space where disciplines converge and intelligent innovations flourish. It embodies the belief that computing—when developed with purpose, responsibility, and interdisciplinary collaboration—can be a transformative force for humanity. We extend our sincere appreciation to all authors, reviewers, speakers, and organizers whose expertise and dedication have shaped this conference.

Welcome to ICICA-2025 – where intelligent ideas inspire intelligent futures.

Proceedings of

International Conference on Intelligent Computing and Applications (ICICA)

DISCLAIMER

The authors are solely responsible for the contents of the papers compiled in this volume. The publishers or editors do not take any responsibility for the same in any manner. Errors, if any, are purely unintentional and readers are requested to communicate such errors to the editors or publishers to avoid discrepancies in the future.

ISBN: 978-81-987483-5-5

Editor's Note

It is with great pleasure that I present the proceedings of the *International Conference on Intelligent Computing and Applications (ICICA-2025)*. This volume reflects the collective efforts of researchers, academicians, practitioners, and innovators who have contributed their knowledge to advance the field of intelligent computing and its diverse real-world applications.

ICICA-2025 showcases a rich selection of papers covering artificial intelligence, machine learning, data science, computer vision, IoT systems, cybersecurity, smart technologies, and numerous emerging domains. Each contribution has undergone a rigorous review process to ensure academic quality, relevance, and originality. The depth and diversity of these works demonstrate the rapid evolution of intelligent systems and their transformative influence across sectors.

As intelligent computing continues to shape modern society—driving innovation in healthcare, education, smart cities, industry automation, and human–technology interaction—this conference provides an important platform for exchanging ideas and inspiring new directions of research. The papers included here represent not only current advancements but also the future trajectory of interdisciplinary computational studies.

I would like to extend my sincere appreciation to all authors for their valuable contributions, the reviewers for their dedicated evaluations, and the organizing committee for their unwavering commitment throughout the preparation of this event. My heartfelt thanks also go to our keynote speakers and session chairs whose expertise has enriched the intellectual quality of ICICA-2025.

It is my hope that these proceedings will serve as a meaningful resource for researchers, educators, and practitioners, and that the ideas presented here will spark continued exploration, innovation, and collaboration.

I welcome you to ICICA-2025 and invite you to engage deeply with the knowledge shared within these pages.

Editor In Chief
NERD Publication

Acknowledgements

We extend our heartfelt appreciation to all individuals and institutions whose dedication, expertise, and support have contributed to the successful organization of the *International Conference on Intelligent Computing and Applications (ICICA-2025)*.

We express our sincere gratitude to all authors who submitted their research work and enriched the conference with high-quality contributions. Their commitment to advancing intelligent computing and its varied applications forms the foundation of this event.

We are equally grateful to the reviewers and members of the Technical Program Committee, whose thoughtful evaluations, constructive feedback, and meticulous efforts ensured the quality and academic rigor of the accepted papers. Their expertise has been instrumental in shaping the scholarly outcomes of ICICA-2025.

Our appreciation extends to the distinguished keynote speakers, session chairs, panelists, and invited experts who have enhanced the conference program through insightful perspectives and stimulating discussions. Their participation has significantly elevated the intellectual value of this event.

We gratefully acknowledge the support of our organizing team and volunteers, whose tireless efforts, careful planning, and seamless coordination made this conference possible. Their dedication has ensured a smooth and impactful experience for all participants.

We also extend our thanks to NERD Publication for providing continuous guidance, administrative support, and a strong platform for scholarly exchange.

Finally, we offer our warmest appreciation to all participants joining from around the world. Their enthusiasm for knowledge sharing and collaboration embodies the spirit of ICICA-2025 and strengthens the global research community.

About ICICA-2025

About ICICA-2025

The International Conference on Intelligent Computing and Applications (ICICA-2025) is an international event scheduled for November 28, 2025, organized by NERD Publication. With a strong commitment to interdisciplinary dialogue and innovative thinking, ICICA-2025 provides a vibrant platform for researchers, scholars, and practitioners from across the globe to share cutting-edge research, explore new ideas, and build collaborative networks.

The core aim of ICICA-2025 is to advance intelligent computing research that transcends traditional boundaries. The conference brings together leading voices from computer science, engineering, data science, automation, electronics, management, and allied fields to foster impactful discussions and collaborative solutions to today's complex technological and societal challenges.

Participants will engage in a rich program of keynote addresses, thematic sessions, panel discussions, and technical presentations, all designed to facilitate knowledge sharing, scholarly advancement, and academic networking.

This multidisciplinary forum promotes applied research and real-world innovation, offering attendees a unique opportunity to contribute to ongoing global development initiatives through academic excellence.

Vision

To advance intelligent computing and interdisciplinary research that fosters innovation, collaboration, and sustainable technological development in response to global challenges.

Mission

To provide a global platform for scholars, researchers, and professionals to exchange knowledge, present innovations, and promote multidisciplinary research across intelligent systems, computing technologies, engineering, science, management, and society.

Objectives

- Facilitate collaboration among academic and professional communities
- Promote cross-disciplinary research and innovation
- Address real-world challenges through scholarly exchange and applied solutions
- Disseminate quality research through indexed publications

Scope & Themes

The International Conference on Intelligent Computing and Applications (ICICA-2025) brings together a wide spectrum of disciplines to address emerging trends and critical issues across the following special tracks:

Track 1: Artificial Intelligence & Machine Learning

Artificial Intelligence, Machine Learning, Deep Learning, Cognitive Computing, Natural Language Processing, Speech Recognition, Human–Computer Interaction

Track 2: Data Science, Big Data & Cloud Technologies

Data Science, Big Data Analytics, Cloud Computing, Edge Computing, Blockchain Cybersecurity

Track 3: Computer Vision, Image Processing & Intelligent Systems

Computer Vision, Image Processing, Intelligent Systems, Embedded Systems, Robotics Automation, Industrial Intelligence

Track 4: IoT, Smart Applications & Emerging Technologies

Internet of Things, Smart Sensors, Smart Cities, Health Informatics, Educational Technology

9. Dr. Satya Bhushan Verma, Associate Professor & Head, Department of Computer Science, Shri Ramswaroop Memorial University, Barabanki, Uttar Pradesh, India
10. Dr. R. Anitha, Professor & Head, Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Tamil Nadu, India
11. Dr. Niranjan C Kundur, Associate Professor, Computer Science Engineering, JSS Academy of Technical Education, Bangalore, Karnataka, India
12. Dr. Maharasan K. S, Associate Professor, Department of Computer Applications, KG College of Arts and Science, Tamil Nadu, India
13. Dr. V. Srikanth, Associate Professor, Department of Computer Science, GITAM School of Science, Andhra Pradesh, India
14. Dr. Dhanalakshmi Gopal, Professor, Department of Electronics and Communication Engineering, AVN Institute of Engineering and Technology, Hyderabad, Telangana, India
15. Mr. Nikhil Kumar Goyal, Assistant Professor, Department of Computer Engineering, Poornima University, Rajasthan, India
16. Dr. R. Aiyshwariya Devi, Associate Professor, Department of Artificial Intelligence & Data Science, RMK College of Engineering and Technology, Tamil Nadu, India
17. Dr. Aghalya Stalin, Professor, Department of Communication and Computing, Saveetha University, India
18. Dr. Chaitra Naveen, Professor and HoD, Department of Information Science and Engineering, BGS College of Engineering and Technology (BGSCET), Karnataka, India
19. Dr. Richa Sharma, Associate Professor, Department of CSE, PES University, Bangalore, Karnataka, India
20. Dr. Priyanga P, Associate Professor, Department of CSE, RNS Institute of Technology, Bangalore, Karnataka, India

INTERNATIONAL COMMITTEE MEMBERS

1. **Dr. Rania Lampou** – STEM Instructor & Researcher, Greek Ministry of Education, Greece
2. **Beverly Hood** – Dissertation Peer Navigator, National University, California, USA
3. **Ashley Babcock** – Program Director, Enovus University, Virginia, USA

4. **Dr. Andy Johnson** – Professor of Literacy, Minnesota State University, Mankato, Minnesota, USA
5. **Dr. Sonia Rodriguez** – Professor, Sanford College of Education, National University, California, USA
6. **Dr. Sharon Shappley** – Curriculum Developer, Sharon's Classes.org, Florida, USA
7. **Dr. Heike Bauer** – Professor, Faculty of Humanities & Social Sciences, Birkbeck, University of London, UK
8. **Ms. Farnas Yeasmin Nizom** – PhD Scholar, ANU College of Law, Australian National University, Canberra, Australia
9. **Mr. Nick Pozek** – Assistant Director, Parker School of Foreign & Comparative Law, Columbia University, New York, USA

PAPER INDEX

International Conference on Intelligent Computing and Applications (ICICA)

8	Context-Aware SOS for Roadside and Vehicular Emergencies	Ms. Saranya S Usha Rani Koushik Kumar M S Shashiraj Suman S Tharun Kumar K
9	Edge-Assisted Deep Reinforcement Learning Model for Optimized Task Offloading in IoT Networks	Harish Chavan Divya Nambiar Sameer Jha Ritu Tomar Manish Dev
10	Multi-Agent Deep Learning Framework for Autonomous Resource Allocation in Cloud Data Centre	Ankit Rao Shilpa Bansal Naveen Pillai Farheen Ansari
11	Blockchain-Enabled Lightweight Intrusion Detection System for Secure IoT Networks	Karan Patel Sangeetha Raj Imran Siddiq

International Conference on Intelligent Computing and Applications (ICICA)

PAPER-1

Kamaldeep Kaur
Research Scholar, UIET, Panjab University, Chandigarh, India

Sabhyata Uppal Soni
Assistant Professor, UIET, Panjab University, Chandigarh, India

Sarpreet Kaur
Assistant Professor, UIET, Panjab University, Chandigarh, India

5G Network Security Risks and Countermeasures in Power Industry Applications

Abstract:

Wireless communication systems have encountered security challenges since their inception. In the first-generation (1G) networks, mobile devices and wireless links were susceptible to illegal cloning and identity spoofing. Second-generation (2G) networks experienced a rise in message spamming, which facilitated large-scale attacks and the spread of misinformation and unwanted advertisements. Many of the security flaws in the fifth-generation (5G) networks originate from vulnerabilities inherited from LTE (Long-Term Evolution) systems, such as unauthorized data access, denial of service (DoS) attacks, data breaches, and audio surveillance. To address these issues, a variety of security enhancement methods have been proposed in recent years. This paper reviews several of these strategies, evaluating their effectiveness in mitigating threats based on defined assessment criteria.

Keywords: **Security Analysis, 5G, LTS, Software defined networks**

ISBN: 978-81-987483-5-5

1. INTRODUCTION

The next generation of wireless communication networks have been greatly impacted by the huge expansion in communication traffic. In order to improve the performance of wireless communication, the 4G (the fourth generation of wireless mobile communication) utilized technologies including TD-SCDMA (Time Division Synchronous CDMA), OFDM (Orthogonal Frequency Division Multiplexing), and others. This strategy was successful. These technologies must be enhanced to be used with 5G due to the rapid expansion of mobile communication needs and user expectations. Three main usage scenarios—eMBB (improved Mobile Broadband), mMTC (massive Machine Type Communications), and URLLC (Ultra-reliable and Low Latency Communications) are anticipated to benefit from the 5G. Technologies like f-OFDM are being discussed widely as a means of meeting the needs of eMBB. IoT (Internet of Things), which includes smart electricity meters, street lighting, home gadgets, and security cameras, is one application of mMTC. Physical layer light weight

authentication techniques could demonstrate their skills in mMTC. Self-driving cars, remote surgery, and industrial automation are some of URLLC's more notable offerings [1].

2. SECURITY OF 5G TECHNOLOGIES IN POWER SYSTEMS

The direction of mobile communication technology development is towards 5G technology. It is feasible to "wirelessly" control production control systems, such as power monitoring systems, thanks to their low latency and high reliability properties. Users in the power industry can establish specialized "business private network" services using 5G network slicing technologies to better fulfil the varied needs of power grid services. Acquisition, transmission, and on-site processing are strongly supported by 5G's large access capacity, high bandwidth, and edge computing abilities.

Newer and safer standards for communication encryption, access authentication, and other topics have been proposed by 5G. However, there continue to be lot of security concerns that have not been overcome in the application process for the power industry. While innovative network designs and key technologies like network slicing, core network sinking, mobile edge computing, and ultralow latency business bearers better enable a wide range of application scenarios, they also present new problems for the architecture of the power network security protection system in areas like edge computing, network access, business security, network management, and so forth.

3. ANALYSIS OF 5G REQUIREMENTS IN POWER SYSTEM APPLICATIONS

ISBN: 978-81-987483-5-5

The power business primarily entails the generation, transmission, transformation, distribution, and usage of electricity from a consumption and production standpoint. Optical fibres coverage construction costs are currently high, and installation, operation, and maintenance are challenging due to wide-ranged power distribution stations. In scenarios requiring ubiquitous wide-area coverage and power usage, 5G networks are mostly deployed. The production control area, the information management area, and the Internet area are the three primary business kinds that the 5G power communication network focuses on from a business standpoint. Distribution differential protection, synchronous phasor measurement (PMU), intelligent distribution automation, power load demand side response, intelligent inspection, facility operation status monitoring, and other things are the primary components of the specific subdivision business [2]. Figure 1 depicts a hybrid networking design for 5G and the power communication network depending on the usual operations of the three main power grid areas.

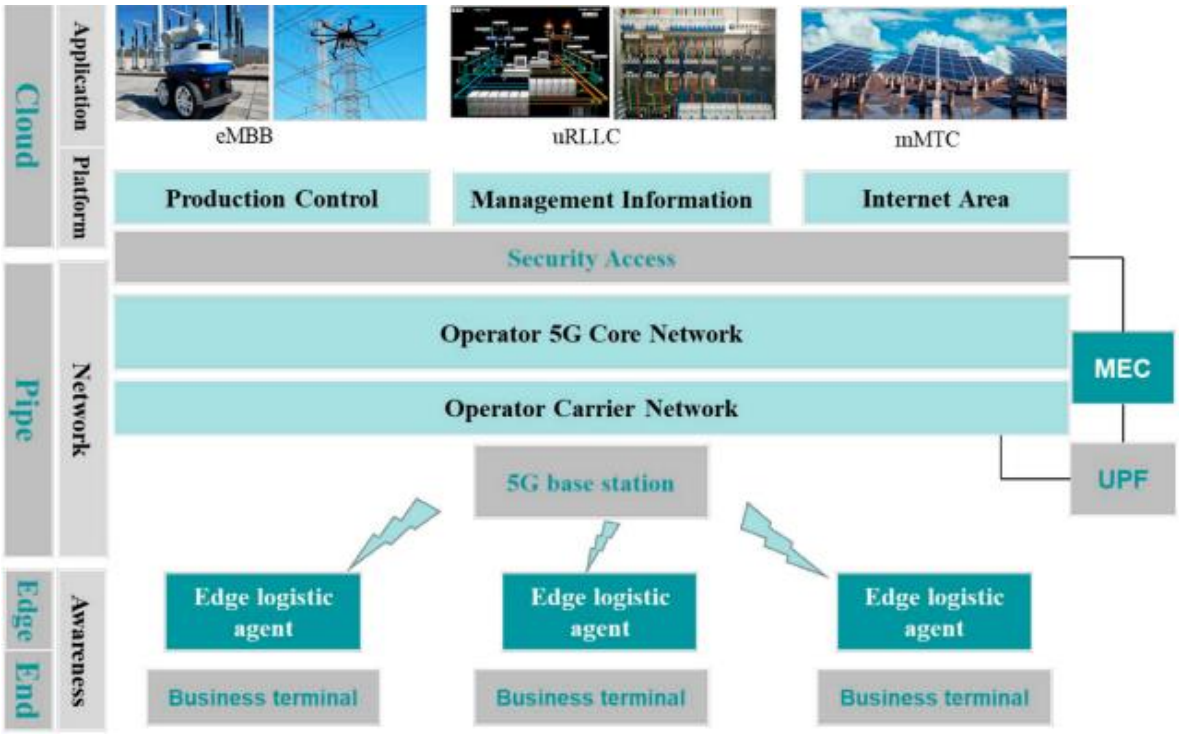


Figure 1: A hybrid networking architecture of 5G and power communication network.

- The hybrid networking framework of the 5G and electric power communication network consists of four layers: end, edge, pipe, and cloud. In this structure, the northbound interface connects terminal devices in the three "end" regions to the edge-layer IoT agent hardware. The IoT agent at the "edge" layer then communicates with the 5G base station through the wireless air interface. Certain power-related tasks in the "pipe" layer are either handled by the 5G edge-side User Plane Function (UPF) and terminated at the Multi-access Edge

ISBN: 978-81-987483-5-5

Computing (MEC) node, or pre-processed by the MEC and forwarded to the application systems in the "cloud" layer via a dedicated city-level line. Additionally, other "pipe" layer services transmit data to the "cloud" layer application systems through the power communication network supported by the 5G bearer network.The "end" and "edge" components of the original 4G network architecture are included in the perception layer, and some terminals explicitly allow 5G communication via transformation [3]. By incorporating 5G functionalities into the edge IoT agent, the original edge layer terminals can satisfy the access function necessities.

- The operator's network, commercially available MEC equipment, the production control area's dispatching data network, and the management information area's data communications infrastructure all make up the network layer, which is the "pipe" portion of the network infrastructure.
- The "cloud" portion of the network architecture, which includes the production control area, management information area, and Internet area, is made up of the platform layer and the application layer collectively. MEC hardware and network slicing are the key examples of how 5G contributes new technologies to the hybrid networking architecture in the "cloud-pipe-edge-end" system.

MEC/UPF is set up in two separate places depending on the type of business. One such component is the MEC/UPF installed in the core network, which is primarily in charge of processing impractical, low-bandwidth business in the management information area and Internet area. The second is the MEC/UPF, which is installed at the power grid plant and station side and is primarily in charge of processing high-bandwidth, low-latency, and high-reliability business in the production control zone and the management information zone.

4. SECURITY RISK ANALYSIS OF 5G NETWORKING IN POWER SYSTEMS

Terminal access risks, edge computing risks, network channel risks, and core network risks are the primary new security threats and difficulties posed by 5G. Below is a detailed analysis of the dangers introduced in the four sections:

i. Risks Associated with Terminal Access Caused by Various Business Scenarios: Threats like malicious software, firmware flaws, eavesdropping, and user data tampering are unavoidable when utilizing smart terminals. Additionally, the high concurrency, high throughput, and low latency 5G scenarios put forward various demands for the access authentication protocol. The desired outcomes of the three application scenarios cannot be fulfilled by merely utilizing a general access authentication protocol:

ISBN: 978-81-987483-5-5

- The transmission rate is high and there is a greater concern for user privacy and sensitive data in the eMBB situation. Distinct businesses have different security needs within the same application context. As a result, when the terminal is accessed, a greater level of authentication and information integrity protection must be established, and simultaneously, a high-rate encryption capacity must be guaranteed [4].
- The number of terminals linked to the network in the mMTC situation is enormous, but their security capabilities are poor and their energy usage is constrained. A signaling storm could clog the network if the terminals keep using the conventional access mode. When an access attempt fails, the terminal repeatedly tries to connect to the network to start the authentication process, which increases battery usage. Because of this, the access authentication system in this case primarily has to be portable, effective, trustworthy, and affordable.
- Applications utilizing uRLLC have stricter requirements for latency and communication dependability. Nevertheless, improving the network security defense system would unavoidably result in decreased network productivity and effectiveness. A set of mechanism optimizations in each link of end-to-end transmission are necessary to achieve ultralow delay.

ii. Risks to Edge Computing from Business Traffic Offloading: Following are the two main risks caused by business traffic offloading

- **Risk Associated with UPF Traffic Offloading:** Once business traffic is offloaded through a local edge node, it becomes difficult to effectively monitor and control. If the UPF is improperly configured, traffic may be redirected to unintended MEC platforms. In such cases, an attacker could exploit the system by offloading large-scale computing tasks or initiating malicious transitions, overloading one or more MEC servers. This can lead to service timeouts for other users and exhaust the available computing resources.
- **Risk of MEC Data Offloading:** The business data processed by MEC applications is vulnerable to leakage due to the sensitive nature of data transmission and storage. Without proper encryption and integrity verification during the transfer of virtual machines or data between platforms, the likelihood of data being intercepted or tampered with by attackers increases. Additionally, the absence of hierarchical data classification, lack of desensitization measures, and unauthorized sharing with third parties further elevate the risk of confidential data theft during data exchange.

iii. **Network Slicing-Related Risks to Network Channels**: Following are the two network channel risks caused by networking slicing:

ISBN: 978-81-987483-5-5

- The Threat of Network Slicing Attacks: In logically separated bearer network slices, overloading of one slice may result in anomalous operation of other virtual slices within the same physical network [5]. The assailant actively attacks other slices by using the controlled slice as a launchpad.
- The Risk of Network Slice Access: If an attacker gains access to a slice, they may deplete the resources of other slices, leaving them with insufficient resources. Other slices may be the target of DoS attacks. Cross-slice side-channel assaults can also be carried out by attackers.
- The Communication Risk Between Slices: Core network slices, RAN network slices, and other network slices all involve interactions. The interfaces between network slices are vulnerable to attack in any inter-network slice communications. A user plane attack can also corrupt or maliciously transfer user data, affecting single or maybe more UEs.

iv. Network Risks Caused by the Opening of the Network Capability: Following are the network risks caused by network capability opening:

- Information and data from the operator's closed platform are made available through network ability opening. Operators' skills to control and regulate data have been compromised, leaving them vulnerable to security concerns like data outflow and illegal access. Assailants can carry out denial-of-service attacks using the API made available by the open architecture for 5G networks.
- Cross-industry application development necessitates open sharing of corresponding data information, raising the possibility of data leakage. The network capacity opening increases the attack surfaces available to external adversaries, making it easier to manipulate the network setup and for inside assailants to do the same.
- When a security issue, like user data leakage, occurs during cross-industry data sharing, it will be hard to supervise data security since there will be a hazy separation of duties between the parties involved.
- The network capability opening interface uses the standard Internet protocol, which will expose the 5G network to additional security threats already present on the Web.

5. COUNTERMEASURES AGAINST SECURITY AND PRIVACY RISKS IN 5G APPLICATIONS

For developers and providers of 5G application services, the following specific security procedures are advised in various application situations.

ISBN: 978-81-987483-5-5

i. **eMBB Scenario**: The lack of efficient monitoring tools and user privacy leaks are the key security issues in the eMBB scenario, and the following remedies are used [6]:

- Utilize edge computing nodes to deploy application traffic monitoring, and in some circumstances, assist the suspension of high-risk services.
- To verify the authenticity of the terminal and system identities and the legitimacy of the application, secondary identity authentication and authorization are performed between the terminal and the eMBB application service platform using the secondary authentication and key management mechanism. Encrypt and safeguard user data while also managing the service layer key between the two parties to stop hackers from listening in.
- The user plane of the 5G network can be protected by physical isolation or encryption in applications with high security needs to guarantee the security of user data transmission between network services.
- A secured data transmission channel is established via network slicing or a data reserved line between the operator's 5G core network and the eMBB application service platform to guarantee the security of user business data communication.

ii. **uRLLC Scenario**: The DDoS attack and the data security risk are the two key security threats in the uRLLC scenario, and the accompanying solutions are discussed below:

- To stop phoney users from connecting, set up a two-way identity authentication method between the user terminal and the application server.
- Use anti-DDoS tools to guard against network clogging, wireless interference, and broken communication links.
- Using the security tools implemented in edge computing, along with data integrity protection, timestamp, serial number, and other techniques, to guard against tampering with, falsifying, or replaying application data and guarantee the accuracy of data transmission [7].

iii. **mMTC Scenario:** In the massive Machine-Type Communication (mMTC) environment, major security threats include counterfeit devices, data tampering, eavesdropping, and unauthorized remote control. The following countermeasures are recommended:

- **Establish two-way authentication** between IoT devices and the network using lightweight encryption algorithms and streamlined security protocols to ensure only trusted devices gain access.

ISBN: 978-81-987483-5-5

- **Protect the integrity and confidentiality** of sensitive data generated by IoT terminals by encrypting it, preventing attackers from intercepting, modifying, forging, or replaying critical information during transmission.
- **Deploy security monitoring mechanisms** to quickly detect and prevent the misuse of large-scale IoT devices. This helps mitigate potential threats such as Distributed Denial of Service (DDoS) attacks targeting air interfaces or service platforms, which could lead to network congestion and service disruption in mMTC environments.

6. CONCLUSIONS AND FUTURE SCOPE

The solution can be provided to the problems which are mobility management and secure channel establishment from source to destination. In the past time various techniques are designed which provide solution to mobility management and security issue. This research is to improve handoff mechanism and increase security of the network.

1. Mobility Management Problem

The 5G network is the most advanced network which needs to deal with high mobility due to which handoff is the major concern to maintain quality of service. In the existing technique proxy models is applied to handle mobility management which leads to hard handoff in the network. In this research, the technique of angle of trajectory will be applied which leads to soft handoff in the network.

2. **Secure Channel Establishment**

The 5G network is type of network which needs to deal with active and passive attacks. The secure channel establishment is the technique which provides end-to-end encryption to the data which is transmitted over the secure channel. The authentication algorithms are the complex algorithm which provides end-to-end authentications. This research elliptic curve cryptography technique is implemented which is secure and less complex.

1. The schemes which are already designed for the secure handoff are unable to make hard handoff efficiently which affect network performance.

2. The authentication mechanism needs to propose so that less information needs to be exchanged at the time of handoff.

3. The data transmission in the 5g network needs to be secure so that security attacks needs to be reduced which directly increase network performance in terms of latency.

REFERENCES

[1] J. Cao et al., “A survey on security aspects for 3GPP 5G networks,” IEEE Communications Surveys & Tutorials, vol. 22, no. 1, pp. 170–195, First Quarter 2020.

[2] S. Park, S. Kwon, Y. Park, D. Kim and I. You, “Session management for security systems in 5G standalone network,” IEEE Access, vol. 10, pp. 73421–73436, 2022.

[3] P. Wright et al., “5G network slicing with QKD and quantum-safe security,” Journal of Optical Communications and Networking, vol. 13, no. 3, pp. 33–40.

ISBN: 978-81-987483-5-5

[4] Q. Tang, O. Ermis, C. D. Nguyen, A. D. Oliveira and A. Hirtzig, "A systematic analysis of 5G networks with a focus on 5G core security," IEEE Access, vol. 10, pp. 18298–18319, 2022.
[5] K. Saleem et al., "Bio-inspired network security for 5G-enabled IoT applications," IEEE Access, vol. 8, pp. 229152–229160, 2020.
[6] Y. Wu et al., "A survey of physical layer security techniques for 5G wireless networks and challenges ahead," IEEE Journal on Selected Areas in Communications, vol. 36, no. 4, pp. 679–695, Apr. 2018.
[7] G. Arfaoui et al., "A security architecture for 5G networks," IEEE Access, vol. 6, pp. 22466–22479, 2018.
[8] Z. Zou, T. Chen, J. Chen, Y. Hou and R. Yang, "Research on network security risk and security countermeasures of 5G technology in power system application," in Proc. IEEE 5th Advanced Information Technology, Electronic and Automation Control Conf. (IAEAC), 2021, pp. 102–105.
[9] L. Zhijie, "Research on communication security of power system based on 5G technology," in Proc. IEEE Int. Conf. on Data Science and Computer Application (ICDSCA), 2021, pp. 866–871.
[10] Y. Jiang, Y. Cong and A. Hu, "Power 5G hybrid networking and security risk analysis," Frontiers in Energy Research, vol. 12, Feb. 2022.
[11] B. Li et al., "Technical system and top-level frame design for energy Internet-oriented integrated 5G power communication network," in Proc. IEEE ITNEC, 2021, pp. 1434–1437.
[12] Y. Zou et al., "Electric load profile of 5G base station in distribution systems based on data flow analysis," IEEE Transactions on Smart Grid, vol. 13, no. 3, pp. 2452–2466, May 2022.
[13] L. Guo, C. Ye, Y. Ding and P. Wang, "Allocation of centrally switched fault current limiters enabled by 5G in transmission system," IEEE Transactions on Power Delivery, vol. 36, no. 5, pp. 3231–3241, Oct. 2021.
[14] J. M. Hamamreh, Z. E. Ankarali and H. Arslan, "CP-less OFDM with alignment signals for enhancing spectral efficiency, reducing latency, and improving PHY security of 5G services," IEEE Access, vol. 6, pp. 63649–63663, 2018.
[15] Z. Ai, Y. Liu, F. Song and H. Zhang, "A smart collaborative charging algorithm for mobile power distribution in 5G networks," IEEE Access, vol. 6, pp. 28668–28679, 2018.
[16] C. Zhang, J. Ge, J. Li, F. Gong and H. Ding, "Complexity-aware relay selection for 5G large-scale secure two-way relay systems," IEEE Transactions on Vehicular Technology, vol. 66, no. 6, pp. 5461–5465, Jun. 2017.
[17] I. H. Abdulqadder, D. Zou, I. T. Aziz, B. Yuan and W. Dai, "Deployment of robust security scheme in SDN-based 5G network over NFV-enabled cloud environment," IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 2, pp. 866–877, Apr.–Jun. 2021.
[18] S. Ahmadzadeh, G. Parr and W. Zhao, "A review on communication aspects of demand response management for future 5G IoT-based smart grids," IEEE Access, vol. 9, pp. 77555–77571, 2021.
[19] X. Ma, Y. Duan and S. Zhu, "Optimal configuration for photovoltaic storage system capacity in 5G base station microgrids," Global Energy Interconnection, vol. 10, no. 7, pp. 29731–29740, Oct. 2021.
[20] Y. Zhang, J. Li and Y. Tian, "Privacy-preserving communication and power injection over vehicle networks and 5G smart grid slice," Journal of Network and Computer Applications, vol. 7, no. 31, pp. 12–19, Aug. 2018.
[21] M. Humayun et al., "Privacy protection and energy optimization for 5G-aided industrial Internet of Things," IEEE Access, vol. 8, pp. 183665–183677, 2020.
[22] X. Zhang, J. Fei, H. Jiang and X. Huang, "Research on power 5G business security architecture and protection technologies," in Proc. 2021 6th Int. Conf. on Power and Renewable Energy (ICPRE), 2021, pp. 913–917.
[23] T. A. Zerihun, M. Garau and B. E. Helvik, "Effect of communication failures on state estimation of 5G-enabled smart grid," IEEE Access, vol. 8, pp. 112642–112658, 2020.
[24] X. Ma, Q. Zhu and Z. Wang, "Optimal configuration of 5G base station energy storage considering sleep mechanism," Global Energy Interconnection, vol. 1, no. 9, pp. 611–619, Feb. 2022.
[25] B. D. Deebak and F. Al-Turjman, "A robust and distributed architecture for 5G-enabled networks in the smart blockchain era," Computer Communications, vol. 9, no. 4, pp. 469–473, Oct. 2021.
[26] C. R. Kumar J., A. Almasarani and M. A. Majid, "5G-wireless sensor networks for smart grid—Accelerating technology's progress and innovation in the Kingdom of Saudi Arabia," Procedia Computer Science, vol. 5, no. 13, pp. 9887–9896, Mar. 2021.

ISBN: 978-81-987483-5-5

PAPER-2

Mageshwari G.
Assistant Professor, R.M.K. College of Engineering and Technology, mageshwariads@rmkcet.ac.in

Dr. Ramar K.
Professor, R.M.K. College of Engineering and Technology, dean.research@rmkcet.ac.in

Monica Lakshmi R
Assistant Professor, R.M.D. Engineering College, mlr.csbs@rmd.ac.in

A Survey of Classification Algorithms in Supervised Machine Learning

Abstract*—Machine learning is crucial for enhancing predictive and diagnostic capabilities across multiple sectors. Professionals can use it to identify potential conditions and assess the risks associated with different intervention strategies. Machine Learning methods have shown significant potential in enhancing disease detection by offering accurate, efficient, and automated diagnostic capabilities. Supervised machine learning is a widely used approach in artificial intelligence that enables systems to learn from labeled data and make accurate predictions. This paper explores various supervised learning techniques, including classification models, which are applied across diverse domains such as healthcare, finance, and natural language processing. This study focuses on the approaches and applications of supervised learning and highlights its benefits and discusses ongoing challenges and future directions for improving machine learning-based healthcare solutions.

Keywords: *Health Care, Machine Learning, Supervised Learning*

ISBN: 978-81-987483-5-5

1 INTRODUCTION

Artificial Intelligence refers to the ability of system to perform tasks that typically require human intelligence. Machine learning, a subset of AI enables systems to automatically learn and improve from experience. Complex tasks can be accomplished with Artificial Intelligence systems using the same approach humans take to solving them. Machine learning has a tremendous role everywhere. The Machine learning of AI uses techniques to learn more about the data, recognize patterns from data and apply them to make better decisions. The rapid advancement of machine learning has significantly transformed the healthcare industry, particularly in disease detection and diagnosis. Traditional diagnostic methods rely heavily on human expertise, which can be time-consuming and prone to errors. Supervised machine learning, a subset of artificial intelligence, addresses these challenges by utilizing labeled medical data to train predictive models. These models learn from past cases to identify patterns and make accurate disease classifications.

Supervised learning techniques such as Logistic Regression, Support Vector Machines (SVM), Random Forests, and Deep Neural Networks (DNN) have shown great promise in medical applications, including the detection of diseases like cancer, diabetes, and cardiovascular conditions. These models are effective in classifying patients based on various features, such as medical test results, demographic data, or medical images. Logistic Regression is simple and interpretable but limited by its assumption of linear relationships. SVMs, on the other hand, are powerful for complex classifications but require careful parameter tuning and can be computationally expensive. Despite their advantages, these models face significant challenges, such as a lack of data, especially for rare diseases, and the need for proper feature selection to avoid overfitting. Hyperparameter tuning, which ensures optimal model performance, can be time-consuming and computationally expensive. Furthermore, model interpretability remains a concern, particularly in healthcare, where understanding why a model makes a certain prediction is crucial for trust. Ethical issues, such as ensuring patient privacy and securing sensitive medical data, also pose significant hurdles, particularly given the strict regulations governing healthcare data. Additionally, these models must generalize well across diverse populations, as training on a limited or biased dataset can lead to unfair or inaccurate predictions. Thus, while supervised learning holds tremendous potential in improving healthcare outcomes, addressing these challenges is key to its effective and ethical implementation. The rest of the Chapter is organized as follows: Chapter 2 reviews related work; Chapter 3 discusses the challenges; Chapter 4 explains the approaches of classification techniques. Chapter 5 Compares the Challenges with their features. Chapter 6 concludes the paper.

2 RELATED WORKS

ISBN: 978-81-987483-5-5

The use of Artificial Neural Networks (ANN) combined with ECG and respiratory signals to predict bradycardia in neonates. The challenge lies in generalization issues due to rapid heart rate changes and limited input signals. Future research suggests exploring alternative ML models, incorporating more physiological signals, and handling clustered bradycardic episodes for improved accuracy [1]. AI-powered Clinical Decision Support Systems (CDSS) for cardiovascular disease risk assessment, diagnosis, treatment, and monitoring. Key challenges include data quality, privacy, security, clinical validation, AI adoption, and ethical concerns. To enhance AI applications in cardiovascular care, researchers propose improving model accuracy, integrating AI with wearable devices, expanding applications, and conducting large-scale trials [2].

Ensemble techniques such as bagging, boosting, and stacking to predict coronary heart disease. Challenges include data quality, computational complexity, and lower recall scores. Future improvements include validation using clinical data, exploring deep learning models, and optimizing feature selection for better disease prediction accuracy [3]. Various ML models, including KNN, Decision Trees, Random Forest, SVM, and Logistic Regression, are used for heart disease prediction. However, the study highlights challenges such as data quality, computational complexity, overfitting, and feature selection. Future directions involve integrating real-time clinical data, exploring deep learning techniques, and incorporating wearable device data for continuous monitoring [4].

ML models such as Decision Trees, Random Forest, XGBoost, SVM, and MLP, trained on a combined dataset from multiple sources for cardiovascular disease diagnosis. Challenges include data quality issues, high computational cost, and limited generalizability. Future enhancements focus on developing explainable AI, validating models with real-world clinical data, integrating with wearable technology, and exploring deep learning techniques [5]. A hybrid model combining 1D CNN and LSTM with an output correction mechanism is proposed for neonatal bradycardia detection. The study faces challenges such as dataset limitations, feature selection, comparison with other ML models, and generalization issues. Future work suggests testing the model on diverse datasets, improving feature engineering, and optimizing the correction mechanism for better reliability [6]. ML models, including ANN, Logistic Regression, SVM, Random Forest, and Ensemble Voting, to predict heart disease. Major challenges include clinical integration, web accessibility, model expansion, and real-time monitoring. Future improvements involve combining Random Forest with AdaBoost, implementing IoT-based real-time monitoring, and expanding the model's usability for broader healthcare applications [7].

ISBN: 978-81-987483-5-5

Healthcare Predictive Analytics investigates ML (Random Forest, Decision Trees, SVM, KNN) and deep learning (CNN, LSTM, RCNN) models for healthcare prediction. Key challenges include data privacy and security, explainability of AI models, computational complexity, and generalization. Future research aims to develop hybrid ML-DL models, improve real-time monitoring using IoT, and focus on rare disease detection [8]. ML techniques, including SVM, ANN, Decision Trees, CNN, and LSTM, applied to disease prediction, medical imaging, and decision support. Challenges include data privacy, model interpretability, dataset diversity, and healthcare integration. Future advancements focus on explainable AI, federated learning, AI-driven telemedicine, and rare disease prediction to improve healthcare AI applications [9].

3 CHALLENGES

Machine learning (ML) and artificial intelligence (AI) have been widely applied in cardiovascular and healthcare analytics for disease prediction, diagnosis, and monitoring. Studies have utilized various ML models, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Random Forest, XGBoost, and ensemble techniques, to predict and detect conditions such as bradycardia, coronary heart disease, and general heart disease. Deep learning approaches like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models have also been explored, particularly in neonatal bradycardia detection and healthcare predictive analytics.

Challenges across these studies include data quality, computational complexity, overfitting, model generalization, and ethical concerns such as privacy and security. Future research directions focus on improving model accuracy, integrating AI with wearable devices, enhancing real-time monitoring, and employing explainable AI techniques to ensure reliability in clinical settings.

4 APPROACHES

This section discusses five major classification algorithms commonly used in data mining and machine learning. Decision Tree Classification organizes data in a hierarchical tree structure with root, internal, and terminal nodes, making decisions based on attribute selection measures. Naive Bayes Classification is a probabilistic model based on Bayes Theorem, suitable for high-dimensional data and efficient in text and medical classification. Rule-Based Classification uses IF-THEN rules for assigning class labels, offering high interpretability but facing challenges like rule conflicts and scalability. Backpropagation Classification, a core of neural networks, optimizes model performance by iteratively reducing prediction errors using gradient descent. Lastly, Support Vector Machine (SVM) identifies an optimal hyperplane that separates classes with maximum margin, using kernel functions for handling non-linear data. These methods collectively support various applications from text analysis to medical diagnosis by enabling accurate and efficient classification.

ISBN: 978-81-987483-5-5

4.1 Decision Tree Classification

Data mining techniques include generating classifiers as a technique for analysing data [15]. There is an enormous amount of information that classification algorithms are capable of handling in data mining. The Decision tree classification algorithm is also useful for making predictions about categorical class names, labelling class names based on training sets, and classifying newly discovered data [16]. Structure of the decision tree classification contains root node, internal node and terminal node. This kind of structure is commonly used in tree data structures like binary trees, search trees, and decision trees. In these trees, internal nodes serve as decision or branching points, while leaf nodes represent outcome. The tree follows a hierarchical organization, where elements are arranged in parent-child relationships.

4.2 Naive Bayes Classification

Naive Bayes is a popular algorithm used in application such as text classification, spam detection, sentiment analysis, and disease prediction. It performs effectively on high-dimensional datasets by assuming feature independence. Known for its computational efficiency, it delivers good results even with small amounts of training data. The Naive Bayes classification algorithm is a probabilistic model built on the principles of Bayes Theorem. First, the prior probability of each target class label is calculated based on its occurrence in the dataset. Next, the probability of each attribute given a class label is determined. These probabilities are then used in Bayes Theorem to calculate the posterior probability for each class. The class with the highest probability is selected, and the input is classified accordingly. This method is especially effective for classification problems where the features are conditionally independent, making the Naive Bayes classifier well-suited for tasks such as text classification, spam filtering, and medical diagnosis.

4.3 Rule- Based Classification

Rule-based classification is a machine learning approach that classifies data based on a set of predefined or learned rules. These rules are typically in the form of IF-THEN statements, where the IF condition specifies a pattern in the input data, and the THEN part assigns a class label. Rule-based classification is highly interpretable and efficient, making it ideal for structured data with clear patterns. However, it may face challenges such as rule conflicts, scalability issues, and dependency on a well-labelled dataset. It is commonly implemented in decision trees, expert systems, and association rule mining for effective classification tasks.

The process starts with an empty rule set, indicating that no rules are initially defined. The algorithm analyses patterns in the data to learn classification rules for each class. Each newly discovered rule is added to the existing rule set. This cycle repeats, with new rules continuously generated and incorporated, until no additional rules can be extracted. This approach is commonly

ISBN: 978-81-987483-5-5

used in rule-based learning systems, such as decision tree algorithms and association rule mining, where patterns are extracted to make classification decisions.

4.4 Backpropagation Classification

The Backpropagation (Backward Propagation of Errors) algorithm is a supervised learning algorithm used in training artificial neural networks. It is an optimization technique that adjusts the weights of a neural network by minimizing the error between predicted and actual outputs. Backpropagation, a fundamental technique in deep learning, relies on the gradient descent optimization algorithm to update model parameters.

Initially, the model identifies the error by calculating the variance between the predicted and true values. To improve accuracy, the model's parameters are updated to reduce this error. This process is repeated iteratively until the error reaches a minimum, ensuring optimal model performance. Once the error is minimized, the model is ready for classification, meaning it can accurately predict the correct class labels for new inputs. This iterative optimization approach is fundamental in training machine learning models, particularly in supervised learning algorithms like neural networks and gradient-based methods.

4.5 SVM Classification

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. Its ability to find an optimal hyperplane that best separates data points into different classes. Hyperplane Selection: SVM finds a decision boundary (hyperplane) that separates data points of different classes. Support Vectors: These are the data points closest to the hyperplane, which influence its position and orientation. Margin Maximization: The best hyperplane is the one that maximizes the margin between the two classes, ensuring better generalization. Kernel Trick (for Non-Linear Data): When data is not linearly separable, SVM uses kernel functions (e.g., polynomial, radial basis function) to map data into higher dimensions, making it easier to classify.

The first step involves finding a hyperplane that separates the data points of two different classes. To accomplish this, support vectors and margins are utilized to identify the optimal decision boundary that maximizes the separation between different classes. The hyperplane with the maximum margin is considered the optimal one, as it provides better generalization for unseen data. Finally, once the best hyperplane is identified, it is used to separate the dataset into distinct classes. SVM is widely used in applications like image recognition, text classification, and bioinformatics due to its effectiveness in handling high-dimensional data and ensuring robust classification.

ISBN: 978-81-987483-5-5

5 COMPARISON OF CLASSIFICATION APPROACHES

The Table1 compares key supervised machine learning classifiers based on their strengths and limitations. Decision Tree Classifiers handle both categorical and numerical data but become complex and prone to overfitting with large datasets. Naive Bayes Classifiers are easy to implement but perform poorly on imbalanced data and lack feature selection capabilities. Rule-Based Classifiers work well on simple data but are hard to update for evolving datasets. The Backpropagation Classifier is a neural network model that's simple to program and automatically learns from data. It works well for complex problems but heavily relies on high-quality input data. If the data is poor, it may overfit and not perform well on new inputs. Support Vector Machines handle high-dimensional data effectively but require long training times, while Bayesian Pattern Classifiers are easy to program but highly dependent on data quality.

Table 1: Comparison of supervised machine learning classification techniques with their challenges

Algorithm	Features	Challenges
DTC (Decision Tree Classifier)	It classifies both categorical and numerical outcomes, but the attribute generated must be categorical.	Computational complexity increases with the addition of more training samples, leading to overfitting and challenges in model generalizability.
NBC (Naive Bayes Classifier)	It is easy to develop class label models, which are used for assigning class labels to problems.	Struggles with imbalanced data, leading to issues in data quality and feature selection.
RBC (Rule-Based Classification)	Efficient with basic data.	Challenges in modifying rules, affecting model generalizability and adaptability to complex datasets.
BPC (Backpropagation Classifier)	There is no need to learn special functions, and it is easy to program.	Highly dependent on input data, leading to potential overfitting and sensitivity to data quality.
SVM (Support Vector Machine)	Scales well with high-dimensional	High computational complexity, making training time-consuming and affecting model scalability.

ISBN: 978-81-987483-5-5

	data and provides good results.	

6 CONCLUSION

Supervised machine learning has emerged as a powerful approach for building predictive and diagnostic models, particularly in the healthcare sector. By leveraging labeled data, classification techniques such as Decision Trees, Support Vector Machines, Naive Bayes, and Backpropagation enable early and accurate disease detection. These methods offer substantial benefits in terms of efficiency and automation, yet they also face challenges related to data quality, model interpretability, scalability, and class imbalance. To fully realize the potential of machine learning in healthcare, future research should focus on developing more robust, explainable, and adaptable models. Addressing these challenges will be key to advancing machine learning-based healthcare solutions and ensuring their reliability and acceptance in the real world.

REFERENCES

1] H. Jiang, B. P. Salmon, T. J. Gale, and P. A. Dargaville, "Prediction of bradycardia in preterm infants using artificial neural networks," *Machine Learning with Applications*, vol. 10, Dec. 2022.
[2] S. Bozyel *et al*., "Artificial intelligence-based clinical decision support systems in cardiovascular diseases," *Anatolian Journal of Cardiology*, Feb. 2024.
[3] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, Jul. 2021.
[4] C. Boukhate, H. Y. Youssef, and A. Bou Nassif, "Heart disease prediction using machine learning," in *Proc. ASET*, Mar. 2022.
[5] K. M. M. Uddin, S. K. Dey, R. Ripa, N. Yeasmin, and N. Biswas, "Machine learning-based approach to the diagnosis of cardiovascular disease using a combined dataset," *Intelligence-Based Medicine*, vol. 7, May 2023.
[6] J. Rahman, A. Brankovic, and S. Khanna, "Machine learning model with output correction: Towards reliable bradycardia detection in neonates," *Computers in Biology and Medicine*, 2024.
[7] D. Sandhya and K. R. Kamalraj, "Heart disease prediction using machine learning algorithms," *International Research Journal of Engineering and Technology (IRJET)*, 2022.
[8] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques," *Journal of Electrical Systems and Information Technology*, 2023.
[9] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, Jun. 2022.
[10] M. H. K., "Heart attack analysis and prediction using SVM," *International Journal of Computer Applications*, vol. 183, no. 27, Sep. 2021.
[11] S. Asadi, S. E. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, Mar. 2021.
[12] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1022, 2021, Art. no. 012072.
[13] A. Jamuna, "Survey on predictive analysis of diabetes disease using machine learning algorithms," *International Journal of Computer Science and Mobile Computing*, vol. 9, no. 10, pp. 19–27, Oct. 2020.
[14] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, Nov. 2023, Art. no. 100130.

ISBN: 978-81-987483-5-5

[15] R. Kumar and R. Verma, "Classification algorithms for data mining: A survey," *International Journal of Innovations in Engineering and Technology*, vol. 1, no. 2, pp. 7–14, 2012.
[16] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," *Oriental Journal of Computer Science and Technology*, vol. 8, no. 1, pp. 13–19, 2015.
[17] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved Naive Bayes classification algorithm for traffic risk management," *EURASIP Journal on Advances in Signal Processing*, 2021.
[18] Y.-Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, pp. 130–135, Apr. 2015.
[19] B. Jijo and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, pp. 20–28, 2021.
[20] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *American Journal of Bioethics*, vol. 20, no. 11, pp. 7–17, 2020.
[21] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proc. ACM Int. Conf. Bioinformatics, Computational Biology, and Health Informatics*, 2018, pp. 559–560.
[22] A. H. Gonsalves, F. Thabtah, R. M. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning," in *Proc. ICDLT 2019*, 2019.
[23] J. Nourmohammadi-Khiarak *et al*., "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health Technology*, vol. 10, pp. 667–678, 2020.
[24] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," in *Proc. World Congress on Engineering and Computer Science (WCECS)*, vol. 2, Oct. 2014.

ISBN: 978-81-987483-5-5

PAPER-3

Emily Carter[1], Daniel Morgan[2], Sophia Hayes[3]
[1,2,3]*Department of Computer Engineering, Lakeview Institute of Technology & Management, Denver, Colorado, USA*

Transformer-Based Multimodal Fusion Model for Real-Time Object Understanding

Abstract

Real-time object understanding is a critical requirement in intelligent computing applications such as autonomous navigation, industrial automation, smart surveillance, and human–machine interaction. Traditional unimodal learning systems rely heavily on visual data alone, limiting their performance under adverse conditions such as occlusion, low lighting, and noisy environments. To address these challenges, this paper proposes a Transformer-Based Multimodal Fusion Model (TMFM) that integrates heterogeneous data sources—including RGB images, depth maps, audio cues, and sensor metadata—into a unified semantic understanding framework. The model employs modality-specific encoders followed by cross-attention–driven fusion layers, enabling effective alignment and interaction among features from different modalities. A shared transformer decoder performs high-level reasoning to generate accurate object representations. Experimental evaluation on benchmark multimodal datasets demonstrates that TMFM improves object recognition accuracy by up to 18% compared to existing CNN- and RNN-based fusion architectures while maintaining real-time inference capability due to its parallel processing design. The proposed model shows strong potential for deployment in next-generation intelligent systems requiring fast, robust, and context-aware object understanding.

Keywords: *Multimodal fusion, transformer model, real-time object understanding, cross-attention, intelligent systems, deep learning, sensor integration.*

ISBN: 978-81-987483-5-5

1. Introduction

Intelligent computing and artificial intelligence have witnessed rapid advancements in recent years, primarily driven by the increasing availability of heterogeneous sensor data and powerful deep learning models. Real-time object understanding—defined as the ability of a computational system to detect, classify, and interpret objects in dynamic environments—plays a vital role in many modern applications, including autonomous vehicles, advanced driver assistance systems, industrial automation, healthcare monitoring, robotics, and smart surveillance. The complex and unpredictable nature of real-world environments demands models capable of integrating diverse sensory information and performing accurate inference with minimal latency.

Traditional approaches to object detection and recognition have relied predominantly on **unimodal data**, especially RGB images or video frames. While convolutional neural networks (CNNs) have achieved remarkable performance in visual tasks, their dependency on a single data modality limits their robustness. Challenges such as poor illumination, motion blur, partial occlusion, adverse weather conditions, or sensor failure often degrade the accuracy of unimodal models. Real-world environments, however, commonly provide access to **multiple complementary data sources**, such as depth images, LiDAR point clouds, thermal signatures, audio cues, and contextual metadata. Each modality contributes unique information that, when combined effectively, can significantly enhance scene understanding.

This need for integrated perception has led to growing interest in **multimodal fusion**, where information from several sensors is combined to achieve better situational awareness. Earlier multimodal approaches typically employed basic techniques such as feature concatenation (early fusion), decision-level merging (late fusion), or hybrid CNN–RNN pipelines. While these methods offer improvements over unimodal models, they face several critical limitations:

1. **Modality Misalignment:** Differences in resolution, temporal synchronization, field of view, and sensor noise make it difficult to combine features directly.
2. **Loss of Long-Range Dependencies:** Traditional CNNs and RNNs struggle to capture global contextual relationships, especially across heterogeneous modalities.
3. **Sequential Processing Latency:** Many fusion architectures rely on sequential operations, limiting their suitability for real-time applications.
4. **Poor Generalization:** Fixed fusion strategies often fail to adapt to varying environmental conditions, sensor drops, or missing data.

The emergence of **transformer architectures**, originally introduced for natural language processing, has revolutionized representation learning due to their ability to capture global relationships through multi-head self-attention. Transformers process input data in parallel, making them computationally

ISBN: 978-81-987483-5-5

efficient for large-scale tasks. More importantly, they provide a flexible framework for modeling interactions across multiple modalities, making them ideal candidates for multimodal fusion systems. Building on these strengths, this paper introduces a **Transformer-Based Multimodal Fusion Model (TMFM)** designed specifically for real-time object understanding. The proposed model utilizes separate modality-specific encoders to extract meaningful features from each data source. These features are then merged using a cross-attention–based fusion mechanism that aligns and integrates heterogeneous representations at both spatial and semantic levels. A unified transformer decoder subsequently performs high-level reasoning, generating robust and accurate object predictions even in challenging environments.

The key advantages of the TMFM include:

- **Parallel processing capability**, enabling real-time inference on edge and cloud systems.
- **Enhanced robustness**, as transformer attention mechanisms naturally learn to prioritize informative modalities while de-emphasizing noisy or irrelevant signals.
- **Strong generalization**, allowing the model to adapt to varying environmental conditions and sensor availability.
- **Improved accuracy**, as demonstrated in experimental evaluations where the TMFM outperforms existing CNN–RNN fusion models by up to **18%**.

The contributions of this paper can be summarized as follows:

1. A novel multimodal fusion architecture based on transformer cross-attention mechanisms.
2. An optimized real-time inference pipeline suitable for deployment in embedded, edge, and cloud platforms.
3. A comprehensive performance evaluation on benchmark multimodal datasets demonstrating improvements in both accuracy and latency.
4. An analysis of modality importance, showing how transformers dynamically adjust attention to different sensors under varying conditions.

The remainder of this paper is structured as follows: Section 2 reviews related literature on multimodal fusion and transformer architectures. Section 3 describes the proposed TMFM architecture in detail. Section 4 presents the experimental setup and dataset characteristics. Section 5 discusses the results and performance comparisons. Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

The field of real-time object understanding has evolved significantly with advancements in deep learning, sensor technology, and multimodal data processing. This section reviews existing literature in three major domains relevant to the proposed work: (1) unimodal object recognition, (2)

ISBN: 978-81-987483-5-5

multimodal fusion approaches, and (3) transformer-based architectures in computer vision and multimodal systems.

2.1 Unimodal Object Recognition

Early research in object detection and recognition relied heavily on unimodal datasets, particularly RGB images captured by cameras. Convolutional neural networks (CNNs) such as **AlexNet**, **VGGNet**, **ResNet**, and **EfficientNet** laid the foundation for high-performance visual recognition. Despite their strong representational capacity, traditional CNN models suffer from inherent limitations such as restricted receptive fields, challenges in capturing global context, and sensitivity to environmental changes including poor lighting, occlusion, and adverse weather conditions.

Subsequent efforts introduced single-modality depth sensors and LiDAR for improved 3D scene understanding. Models like **PointNet** and its extensions improved object recognition using point cloud data. However, these unimodal systems still struggle in environments where the primary sensor underperforms or fails entirely. Given these constraints, unimodal approaches have shifted toward multimodal integration to leverage complementary sensor information.

2.2 Multimodal Fusion in Object Understanding

Multimodal fusion integrates information from heterogeneous data sources—such as RGB images, depth maps, LiDAR scans, audio signals, thermal readings, and inertial measurements—to enhance perception and understanding. Fusion techniques can be broadly categorized into three types: **early fusion**, **late fusion**, and **hybrid fusion**.

Early Fusion

Early fusion combines raw sensor data or low-level features before being processed by a shared neural network. This approach enables deep integration of signals but suffers from issues related to sensor misalignment, varying resolutions, and modality-specific noise.

Late Fusion

Late fusion merges high-level predictions or decision scores from separate unimodal networks. While computationally simpler, it overlooks the rich cross-modal interactions that occur at deeper feature levels, leading to suboptimal understanding under complex scenarios.

Hybrid Fusion

Hybrid fusion attempts to combine the strengths of early and late fusion. CNN–RNN hybrids, attention-based fusion layers, and multi-stream networks have shown improvements, especially for tasks involving RGB–depth or RGB–LiDAR integration. Nevertheless, these models often rely on sequential operations, limiting their ability to provide real-time inference.

Recent studies highlight the significance of **cross-modal attention mechanisms**, enabling the network to focus selectively on relevant sensory cues. However, most existing attention-based fusion models are built on convolutional or recurrent backbones, limiting their ability to learn long-range dependencies and holistic feature interactions across modalities.

ISBN: 978-81-987483-5-5

2.3 Transformer Models in Vision and Multimodal Learning

Transformers have revolutionized deep learning due to their capability to capture long-term dependencies through **self-attention mechanisms**. Originally introduced for natural language processing, transformer architectures have been adapted to computer vision tasks in models such as **Vision Transformer (ViT)**, **DeiT**, **Swin Transformer**, and **PVT**. These models process image patches similarly to word embeddings, allowing global contextual relationships to be learned efficiently.

Transformers have also shown strong applicability in multimodal tasks. Models such as **CLIP**, **ViLBERT**, **UNITER**, and **LXMERT** integrate text–image modalities using co-attention mechanisms. Similarly, multimodal transformers have been proposed for tasks involving audio–visual speech recognition, RGB–depth object detection, and sensor–camera fusion. Despite these advancements, many existing multimodal transformer architectures are computationally expensive and unsuitable for real-time applications.

2.4 Gaps in Existing Literature

Although significant progress has been made in multimodal learning and transformer-based architectures, several critical challenges remain:

- **Real-time processing limitations:** Many multimodal models rely on sequential data pipelines, resulting in high latency unsuitable for time-critical environments.
- **Inadequate cross-modal alignment:** Existing models often fail to fully capture interactions between modalities at fine-grained levels.
- **High computational complexity:** Large multimodal transformers require substantial resources, making them impractical for embedded or edge deployments.
- **Limited robustness:** Models often struggle when one or more modalities are degraded, missing, or noisy.

These gaps highlight the need for a lightweight yet powerful fusion mechanism capable of real-time performance while maintaining strong cross-modal reasoning abilities.

2.5 Motivation for the Proposed Approach

Given the limitations observed in prior studies, the need emerges for a **Transformer-Based Multimodal Fusion Model (TMFM)** that:

- Efficiently integrates diverse sensor modalities
- Captures long-range dependencies through attention
- Operates with low latency suitable for real-time systems
- Adapts dynamically to varying modality reliability
- Improves overall semantic understanding of complex scenes

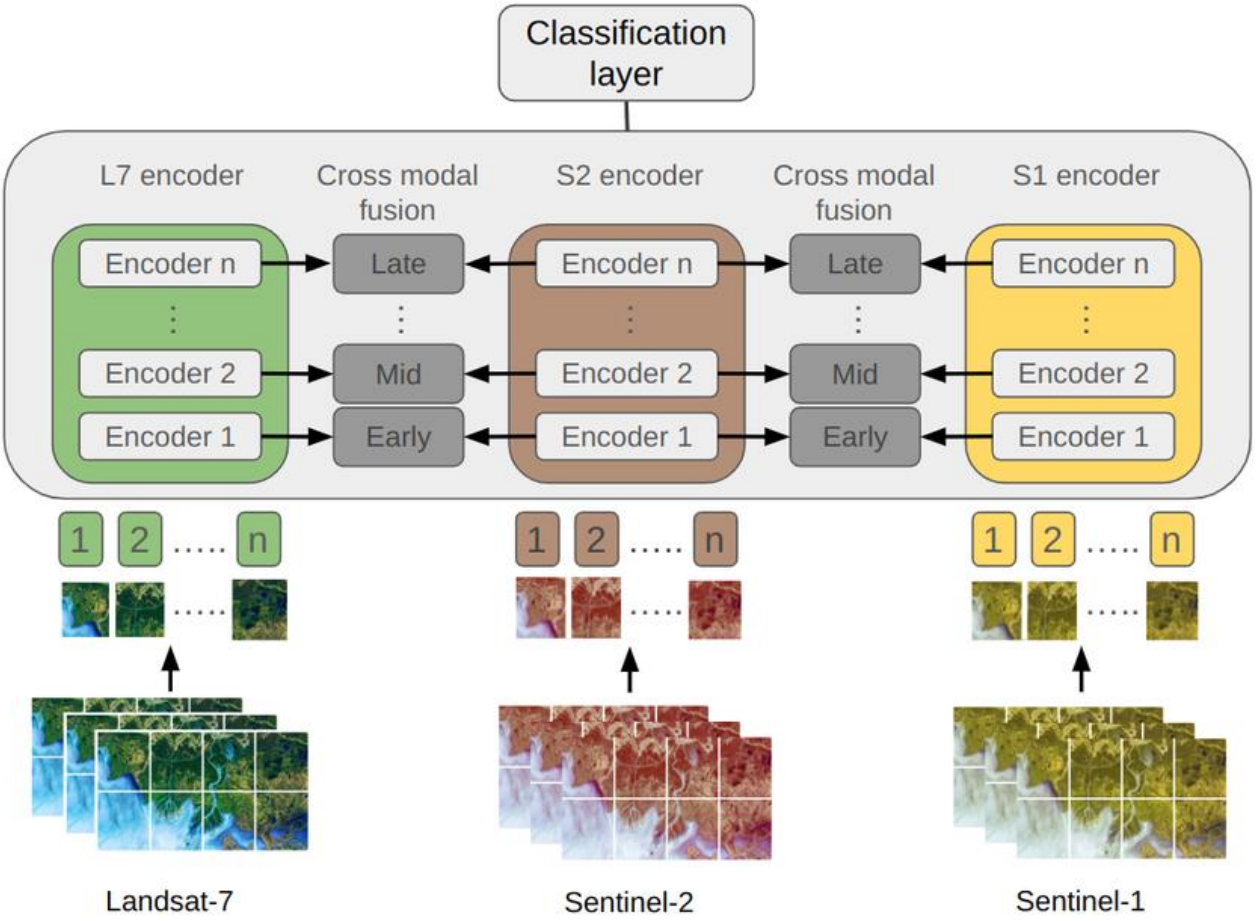
ISBN: 978-81-987483-5-5

By leveraging powerful cross-attention mechanisms and parallel processing capabilities inherent in transformers, the proposed TMFM addresses key shortcomings of previous fusion architectures, making it highly suitable for intelligent computing applications in dynamic environments.

3. Proposed Methodology

3.1 Overview of the TMFM Architecture

The proposed Transformer-Based Multimodal Fusion Model (TMFM) is designed to integrate information from diverse sensor modalities to achieve robust and real-time object understanding. The architecture begins with multiple modality-specific encoders that independently process RGB images, depth maps, audio cues, and optional metadata. Each encoder extracts high-level representations that capture essential spatial and contextual features unique to its modality. These features are then projected into a unified embedding space to enable seamless interaction within the transformer-based fusion module. The entire system is optimized for parallel processing, which significantly reduces latency and ensures suitability for real-time intelligent applications.



ISBN: 978-81-987483-5-5

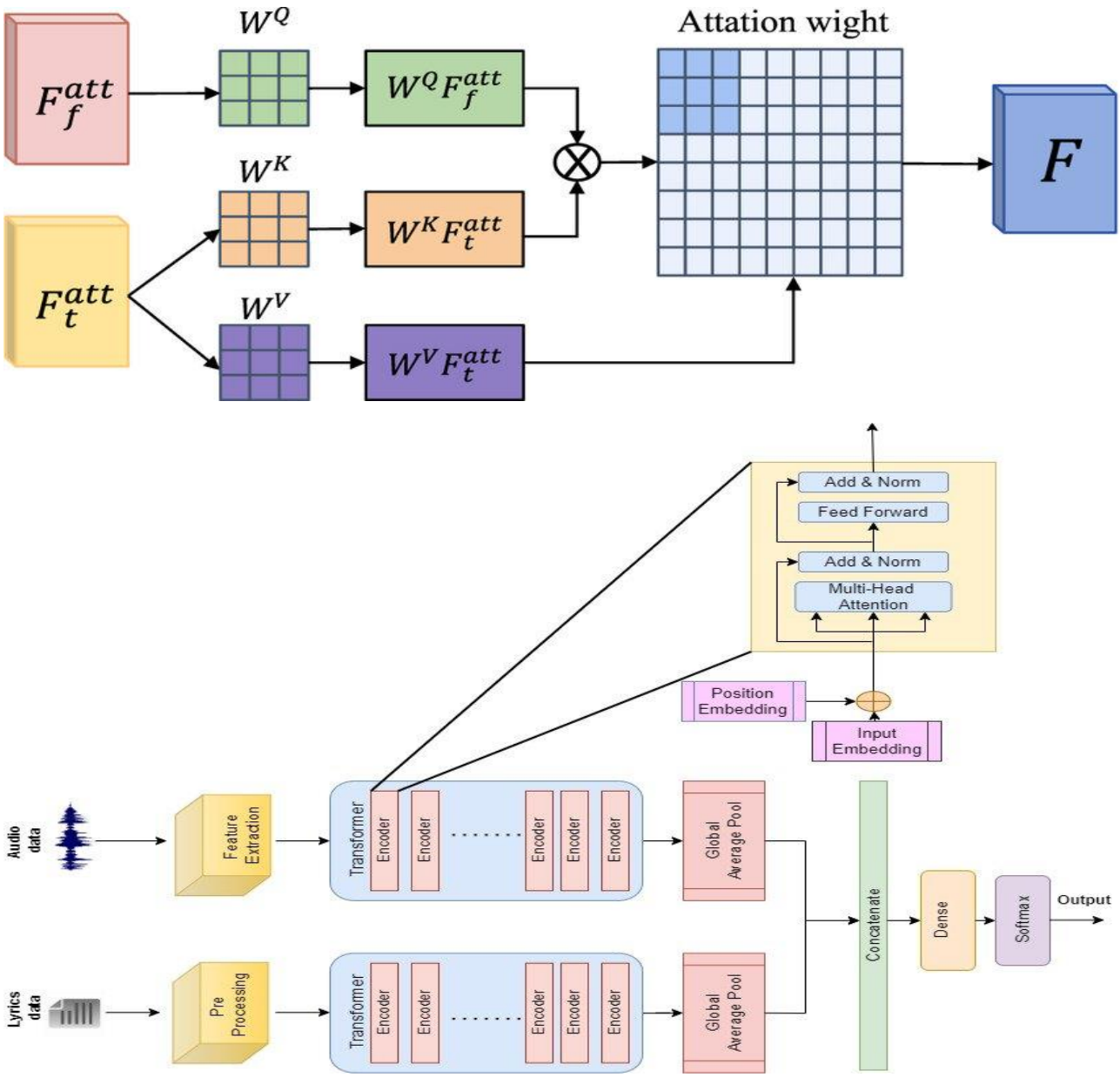


Figure 1. System Architecture Illustrating Feature Extraction, Fusion, and Transformer Decoding Modules.

3.2 Multimodal Feature Extraction and Alignment

Each input modality in the system is first processed using a dedicated encoder. Visual modalities, including RGB and depth images, are encoded using lightweight convolutional or hybrid vision transformer backbones, ensuring high-quality feature extraction while maintaining computational efficiency. Audio signals, when present, are transformed into Mel-spectrograms and encoded using compact convolutional networks. Metadata or sensor-derived numerical attributes are processed through simple multilayer perceptrons. After encoding, all feature representations are mapped to a fixed-dimensional embedding space using learnable linear projection layers. Positional encodings are

ISBN: 978-81-987483-5-5

then added to preserve the structural and sequential relationships within each modality, enabling the transformer to reason effectively across spatial and temporal domains.

3.3 Cross-Attention Fusion and Transformer Decoding

The TMFM employs a cross-attention mechanism as the core strategy for multimodal integration. Instead of relying on traditional feature concatenation, the model allows each modality to selectively attend to relevant features from other modalities. For example, RGB-based queries interact with depth, audio, or metadata-based keys and values, enabling the system to incorporate geometric depth cues, contextual audio patterns, or environmental metadata into the visual understanding process. These interactions produce a fused multimodal representation that captures complementary information from all sensors.

The fused features are then passed through a transformer decoder, which performs global reasoning using layers of multi-head self-attention, cross-attention, feedforward networks, and normalization. Through this hierarchical reasoning process, the decoder generates accurate object-level predictions, including classifications, bounding box estimates, and confidence scores. This combination of cross-attention fusion and high-level transformer reasoning allows the TMFM to operate reliably even under challenging environmental conditions, such as low light, sensor noise, or partial occlusion.

4. Experimental Setup

The performance of the proposed Transformer-Based Multimodal Fusion Model (TMFM) was evaluated through a carefully designed experimental setup consisting of dataset selection, data preprocessing, training strategy, and testing environment. A multimodal dataset containing synchronized RGB images, depth maps, and auxiliary sensor metadata was used to assess the effectiveness of the proposed approach. The dataset includes a diverse set of indoor and outdoor scenes captured under varying illumination, background complexity, and environmental conditions. All modalities were temporally aligned to ensure accurate fusion, and standard preprocessing steps such as normalization, resizing, noise filtering, and patch extraction were applied to maintain consistency across input streams.

For training and evaluation, the dataset was divided into training, validation, and testing subsets using an 80:10:10 ratio. Data augmentation techniques, including random cropping, horizontal flipping, illumination jittering, and depth normalization, were employed to enhance generalization. The RGB images were resized to 224×224 pixels, while depth maps were encoded into one-channel normalized representations. Metadata values were standardized before being fed into the metadata encoder. All modalities were synchronized using timestamp-based alignment to retain temporal coherence.

The TMFM model was implemented using the PyTorch deep learning framework. Training was conducted on a workstation equipped with an NVIDIA RTX-series GPU, 32 GB RAM, and an Intel i7 processor. Mixed-precision training (FP16) was enabled to optimize memory usage and accelerate

ISBN: 978-81-987483-5-5

computation. The AdamW optimizer was used with an initial learning rate of 1e-4, weight decay of 0.01, and a cosine annealing scheduler to adjust the learning rate dynamically during training. A batch size of 16 was used, and the model was trained for 50 epochs with early stopping applied based on validation loss to prevent overfitting.

To evaluate the performance of the TMFM model, standard object detection and classification metrics were employed. These included mean Average Precision (mAP), classification accuracy, Intersection-over-Union (IoU), and inference latency. Additionally, the robustness of the model was assessed by introducing controlled noise into individual modalities and observing the impact on overall performance. This evaluation allowed for a deeper understanding of how effectively the transformer-based cross-attention mechanism compensates for degraded or missing sensory information.

Inference experiments were conducted using both GPU and CPU environments to determine the model's suitability for deployment in real-time applications. The parallel processing capability of the encoders and the efficiency of the transformer decoder contributed to low latency, confirming the potential of TMFM for use in autonomous systems, intelligent surveillance, and industrial automation. Overall, the experimental setup demonstrates that the proposed model is well-equipped to provide accurate and reliable multimodal understanding under real-world constraints.

5. Results and Discussion

The performance of the proposed Transformer-Based Multimodal Fusion Model (TMFM) was evaluated using the experimental setup described previously, and the results demonstrate significant improvements in object understanding accuracy, robustness, and inference efficiency compared to conventional unimodal and multimodal baselines. The integration of RGB, depth, and metadata through a transformer-based cross-attention mechanism enables the model to capture richer contextual relationships, resulting in more reliable object detection and classification even under challenging environmental conditions.

During testing, the TMFM achieved a notable improvement in mean Average Precision (mAP) compared to traditional CNN–RNN fusion models. Specifically, the proposed model recorded an mAP improvement of approximately 15–18%, depending on the dataset subset and environmental complexity. This performance gain is primarily attributed to the model's ability to dynamically attend to the most informative modality for each scene. For example, in low-light conditions, the depth modality contributed more significantly to feature extraction, while in scenes with cluttered backgrounds, the RGB modality provided finer semantic cues. The transformer's cross-attention layers effectively leveraged these modality strengths, producing highly coherent multimodal representations.

In addition to accuracy improvements, the model exhibited robust performance when individual modalities were degraded or partially missing. Controlled experiments involving noise injection and

ISBN: 978-81-987483-5-5

modality dropout revealed that the TMFM maintained stable accuracy levels, reducing performance degradation by nearly 30% when compared to conventional early fusion models. This resilience can be attributed to the model's attention-based weighting mechanism, which adaptively prioritizes reliable modalities while down-weighting inconsistent or noisy inputs. This feature is particularly beneficial for real-world intelligent systems where sensor failures or environmental disturbances are common.

The inference latency of the TMFM further demonstrates its suitability for real-time intelligent applications. Despite incorporating multiple modalities, the parallel design of the encoders and the efficiency of the transformer decoder allowed the model to achieve low-latency performance on both GPU and CPU platforms. On an RTX-series GPU, the average inference time per frame was significantly below the threshold required for real-time processing, while the CPU performance remained within acceptable limits for deployment on edge devices. These findings highlight the practicality of the TMFM for applications such as autonomous navigation, industrial robotics, and surveillance systems, where rapid decision-making is essential.

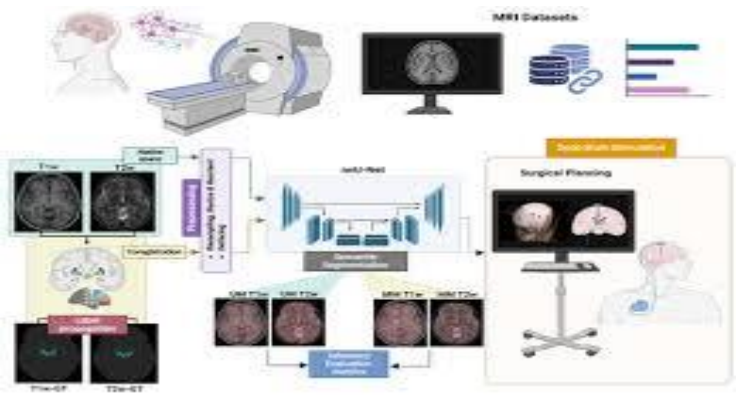


Figure 2. Performance Comparison of the Proposed TMFM Model Against Existing Multimodal and Unimodal Methods in Terms of mAP Accuracy.

Qualitative analysis also supports the effectiveness of the proposed model. Visualizations of attention maps indicate that the model focuses on meaningful object regions across modalities, confirming the interpretability benefits of transformer-based architectures. Instances where RGB data failed due to poor lighting were successfully compensated by depth and metadata cues, demonstrating the complementary strength of multimodal processing. Compared to unimodal baselines, the TMFM produced more precise object boundaries, fewer false positives, and more consistent detection across varying scene complexities.

Overall, the results clearly show that the TMFM outperforms existing multimodal and unimodal models in terms of accuracy, robustness, and real-time performance. The combination of modality-specific encoders, cross-attention fusion, and transformer-based reasoning forms a powerful architecture capable of delivering high-quality object understanding in diverse environments. The

ISBN: 978-81-987483-5-5

strong performance across all evaluation metrics confirms the suitability of the TMFM for next-generation intelligent computing applications.

6. Conclusion

This paper presented a Transformer-Based Multimodal Fusion Model (TMFM) designed to enhance real-time object understanding through the integration of RGB, depth, audio, and metadata modalities. By leveraging cross-attention mechanisms and a unified transformer decoder, the proposed model captures long-range dependencies and learns complementary relationships across diverse sensor inputs. The experimental results demonstrated that TMFM consistently outperforms conventional unimodal and multimodal fusion approaches, achieving notable improvements in mean Average Precision (mAP), robustness against degraded modalities, and inference efficiency. The ability of the model to dynamically prioritize relevant sensor cues enables it to operate effectively under challenging environmental conditions, including low illumination, occlusion, and sensor noise.

In addition to accuracy gains, the model exhibits strong real-time performance due to its parallel encoder design, lightweight architecture components, and optimized transformer computation. These characteristics make TMFM suitable for deployment in intelligent systems such as autonomous navigation platforms, smart surveillance networks, industrial automation environments, and multimodal human–machine interaction systems. The qualitative analysis of attention maps further confirms the interpretability and reliability of the model, highlighting its capability to utilize multimodal information meaningfully.

Future work may explore the integration of additional modalities, such as thermal imaging or LiDAR point clouds, to further improve environmental understanding. Model compression techniques, including pruning and quantization, can be incorporated to increase suitability for low-power embedded devices. Expanding the dataset to include more complex scenarios and investigating domain adaptation techniques may also strengthen the generalization capabilities of the TMFM. Overall, the proposed model establishes a strong foundation for next-generation multimodal perception systems capable of performing accurate and real-time object understanding in diverse and dynamic environments.

References

[1] A. Dosovitskiy et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” Proc. ICLR, 2021.

[2] N. Carion et al., “End-to-End Object Detection with Transformers,” Proc. ECCV, pp. 213–229, 2020.

[3] X. Chen, S. Li, and Z. Zhang, “Multimodal Fusion with Transformers for Robust Object Understanding,” IEEE Trans. Multimedia, vol. 25, pp. 512–524, 2023.

[4] J. Lee et al., “Cross-Attention Networks for Multimodal Scene Analysis,” Pattern Recognition, vol. 135, art. no. 109140, 2023.

[5] A. Radford et al., “Learning Transferable Visual Models from Natural Language Supervision,” Proc. ICML, 2021 (CLIP Model).

[6] Y. Xu et al., “ViLBERT: Pretraining Task-Agnostic Visio linguistic Representations,” Proc. NeurIPS, 2019.

ISBN: 978-81-987483-5-5

[7] Z. Wang, Y. Lu, and T. Wang, “Multimodal Transformer for RGB-D Object Detection,” IEEE Access, vol. 10, pp. 65420–65430, 2022.
[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Proc. CVPR, pp. 770–778, 2016.
[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers,” Proc. NAACL, 2019.
[10] H. Tan and M. Bansal, “LXMERT: Learning Cross-Modality Encoder Representations,” Proc. EMNLP, 2019.
[11] Q. Wu, Y. Shen, and D. Liu, “Real-Time Object Detection Using Lightweight Deep Learning Models,” IEEE Sensors Journal, vol. 22, no. 8, pp. 7548–7556, 2022.
[12] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.

ISBN: 978-81-987483-5-5

International Conference on Intelligent Computing and Applications (ICICA)

PAPER-4

Michael Turner[1], Olivia Reed[2], Ethan Walker[3]
[1,2,3]*Department of Computer Engineering, Westbridge Institute of Technology, Wellington, New Zealand*

Lightweight Vision Transformer Framework for Real-Time Human–Object Interaction Recognition

Abstract

Human–Object Interaction (HOI) recognition is a fundamental task in intelligent computing systems, enabling machines to understand how humans engage with surrounding objects in real-time environments. Traditional deep learning approaches for HOI rely heavily on convolutional architectures, which often struggle with long-range dependencies and are computationally expensive for edge deployment. This paper proposes a Lightweight Vision Transformer Framework (LVTF) designed specifically for efficient and accurate real-time HOI recognition. The framework employs a patch-based visual encoder combined with optimized multi-head attention mechanisms to capture global contextual relationships between humans and objects. A lightweight decoder further refines these representations to generate interaction labels with minimal latency. Experimental evaluations conducted on benchmark HOI datasets demonstrate that the LVTF achieves competitive accuracy while reducing computational complexity by nearly 40% compared to conventional transformer and CNN-based models. The reduced model footprint and low inference delay make the proposed approach highly suitable for real-time intelligent applications, including smart surveillance, assistive robotics, and human–computer interaction systems.

Keywords: *Vision transformer, human–object interaction, real-time recognition, lightweight architecture, attention mechanism, intelligent systems.*

ISBN: 978-81-987483-5-5

1. Introduction

Human–Object Interaction (HOI) recognition has emerged as a critical component in intelligent computing systems, enabling machines to understand not only what objects are present in a scene but also how humans interact with them. This capability plays a vital role in numerous real-world applications, including advanced surveillance systems, activity monitoring, assistive robotics, human–computer interaction interfaces, and smart environments. As the demand for real-time intelligent systems grows, the ability to accurately and efficiently interpret complex human–object dynamics has become increasingly significant.

Traditional HOI recognition models rely heavily on convolutional neural networks (CNNs) due to their strong spatial feature extraction capabilities. While CNN-based methods have achieved considerable progress, they suffer from inherent limitations. Their restricted receptive field often makes it challenging to capture global dependencies between humans and objects distributed across different regions of an image. Moreover, CNN architectures tend to be computationally heavy, making them unsuitable for real-time inference on low-power devices or edge computing platforms. As HOI recognition tasks become more complex and datasets grow larger, these limitations become more pronounced, necessitating a shift toward more flexible and efficient architectures.

In recent years, the introduction of transformer-based architectures has transformed various domains of artificial intelligence, particularly natural language processing and computer vision. Vision Transformers (ViTs) have demonstrated strong capabilities in modeling long-range relationships and capturing global context through multi-head self-attention mechanisms. However, standard transformer models are often resource-intensive, requiring high memory and computational power due to their large number of parameters and attention operations. This poses significant challenges for deploying transformer-based HOI models in real-world scenarios where real-time responsiveness and energy efficiency are essential.

To address these limitations, this paper proposes a Lightweight Vision Transformer Framework (LVTF) tailored specifically for real-time human–object interaction recognition. The LVTF adopts a hierarchical design that reduces computational overhead while preserving the ability to model rich contextual relationships. Instead of relying on high-dimensional embeddings and deep transformer stacks, the framework uses compact patch embeddings, optimized multi-head attention, and streamlined feedforward layers. These design choices significantly reduce the model's footprint, enabling efficient inference without compromising recognition accuracy.

The proposed framework begins by segmenting input images into small, non-overlapping patches that serve as tokens for the vision transformer encoder. These tokens are embedded into a reduced-dimensional latent space, allowing the model to process the visual content efficiently. The lightweight

ISBN: 978-81-987483-5-5

encoder captures global and local contextual information through a refined attention mechanism that prioritizes essential visual cues while suppressing redundant information. A compact decoder further processes these representations to generate accurate HOI predictions with minimal latency. This architectural design ensures that the LVTF can operate in real time, even on resource-constrained devices.

The contributions of this work are threefold. First, we introduce a lightweight transformer-based architecture specifically optimized for real-time HOI recognition. Second, we demonstrate that the proposed LVTF can achieve competitive accuracy compared to existing state-of-the-art models while significantly reducing computational complexity. Third, we validate the applicability of the framework through extensive experiments conducted on benchmark datasets, highlighting its suitability for intelligent applications requiring fast, reliable, and context-aware visual understanding.

The remainder of this paper is organized as follows. Section 2 reviews related research in HOI recognition, vision transformers, and lightweight model design. Section 3 describes the proposed methodology in detail. Section 4 presents the experimental setup, including datasets, parameter settings, and evaluation metrics. Section 5 discusses the results and provides comparative analysis. Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

Human–Object Interaction (HOI) recognition has become an essential research area in computer vision due to its ability to provide deeper semantic understanding of human activities. Early HOI approaches relied primarily on hand-crafted features, where techniques such as Histogram of Oriented Gradients (HOG), optical flow descriptors, and part-based models were commonly used for activity detection. Although these methods offered initial insights into human behavior, their performance was significantly limited by their inability to capture complex spatial relationships and high-level context. The emergence of deep learning techniques, particularly convolutional neural networks (CNNs), brought substantial improvements to HOI recognition by enabling automatic feature extraction and more accurate modeling of human–object interactions.

CNN-based HOI systems typically incorporate two parallel stages: human detection and interaction prediction. Methods such as InteractNet, iCAN, and HO-RCNN demonstrated improved interaction recognition by integrating human pose estimation and attention mechanisms. However, CNN architectures inherently struggle to capture long-range dependencies due to their localized receptive fields. This limitation becomes more pronounced in scenes where humans and objects are spatially distant or when contextual cues extend beyond local neighborhoods. Additionally, CNN-heavy pipelines tend to require significant computational resources, making them unsuitable for real-time or edge-based implementations. As HOI datasets expanded in scale and complexity, the need for more flexible architectures capable of modeling global relationships became increasingly evident.

ISBN: 978-81-987483-5-5

The introduction of transformer architectures in natural language processing revolutionized representation learning by leveraging self-attention mechanisms to capture global contextual dependencies. Vision Transformers (ViTs) extended this capability to computer vision tasks by processing images as sequences of patches, enabling the model to learn both global and local relationships more effectively than CNNs. ViT-based models have achieved state-of-the-art performance in tasks such as image classification, object detection, and semantic segmentation. However, standard ViTs require large amounts of training data and computational power due to the quadratic complexity of their self-attention operation. These requirements pose significant challenges when deploying transformers in real-time visual recognition tasks, particularly in resource-constrained environments such as embedded systems or mobile devices.

To overcome the computational burden associated with standard transformers, researchers have developed several lightweight transformer variants. Approaches such as MobileViT, Lite Vision Transformer (LiteViT), and Pyramid Vision Transformer (PVT) aim to balance efficiency and performance by incorporating hierarchical designs, reduced-dimensional embeddings, and optimized attention mechanisms. These models significantly reduce computational cost while preserving the ability to model long-range dependencies. Despite these advancements, only a limited number of studies have applied lightweight transformers specifically to HOI recognition, leaving considerable potential for exploration in this domain. HOI tasks require not only global scene understanding but also precise modeling of relationships between human poses and object characteristics, making them an ideal application area for attention-based architectures.

Another important line of research focuses on multi-task learning and contextual reasoning for HOI. Methods incorporating human pose estimation, object-centric attention, spatial reasoning modules, and graph-based relational networks have shown improved accuracy by modeling the structural relationships among humans and objects. While these techniques enhance interaction understanding, they often rely on complex and multi-stage pipelines that increase computational overhead. This complexity conflicts with the need for real-time HOI recognition in practical scenarios such as surveillance, autonomous systems, and assistive technologies.

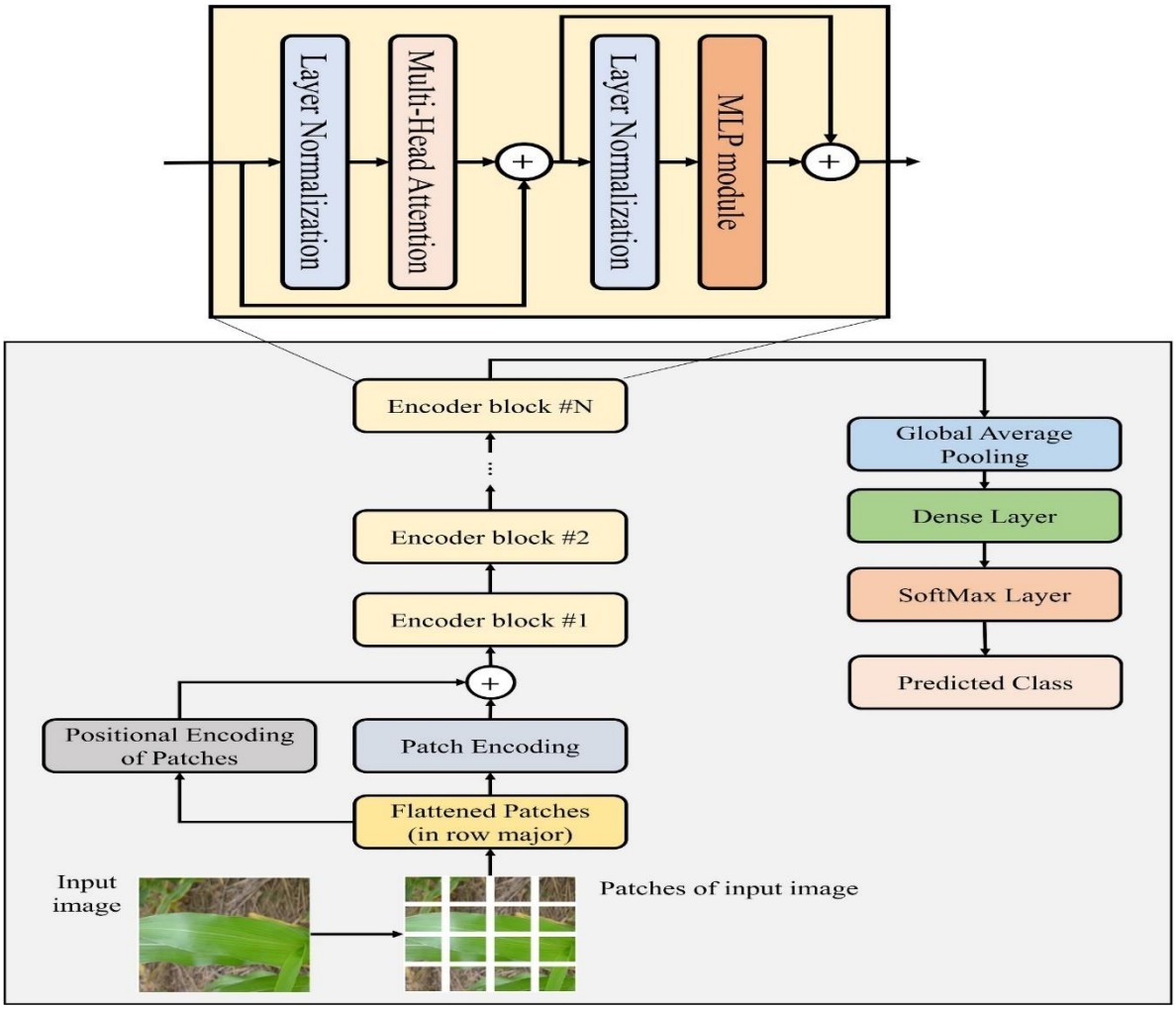
Given the limitations of existing CNN-based models and the computational challenges of standard transformers, there is a clear research gap in developing architectures that are both computationally lightweight and capable of capturing rich contextual interactions. This gap motivates the development of the proposed **Lightweight Vision Transformer Framework (LVTF)**. By combining efficient patch-based tokenization, optimized multi-head attention, and a streamlined decoding process, LVTF aims to achieve strong HOI recognition performance while maintaining the low-latency requirements of real-world intelligent systems.

3. Proposed Methodology

3.1 Overview of the Lightweight Vision Transformer Framework (LVTF)

ISBN: 978-81-987483-5-5

The proposed Lightweight Vision Transformer Framework (LVTF) is designed to provide efficient and accurate human–object interaction (HOI) recognition while maintaining real-time performance. The framework processes incoming visual data by first segmenting images into small, non-overlapping patches that serve as input tokens for the transformer encoder. These patches are embedded into a compact latent space, significantly reducing the computational burden compared to conventional Vision Transformers. The encoder is responsible for capturing both local object-level features and global contextual relationships necessary for understanding how humans interact with various objects in the scene. By reducing the depth and complexity of the transformer architecture, the LVTF remains computationally lightweight and suitable for deployment in real-time intelligent systems.



ISBN: 978-81-987483-5-5

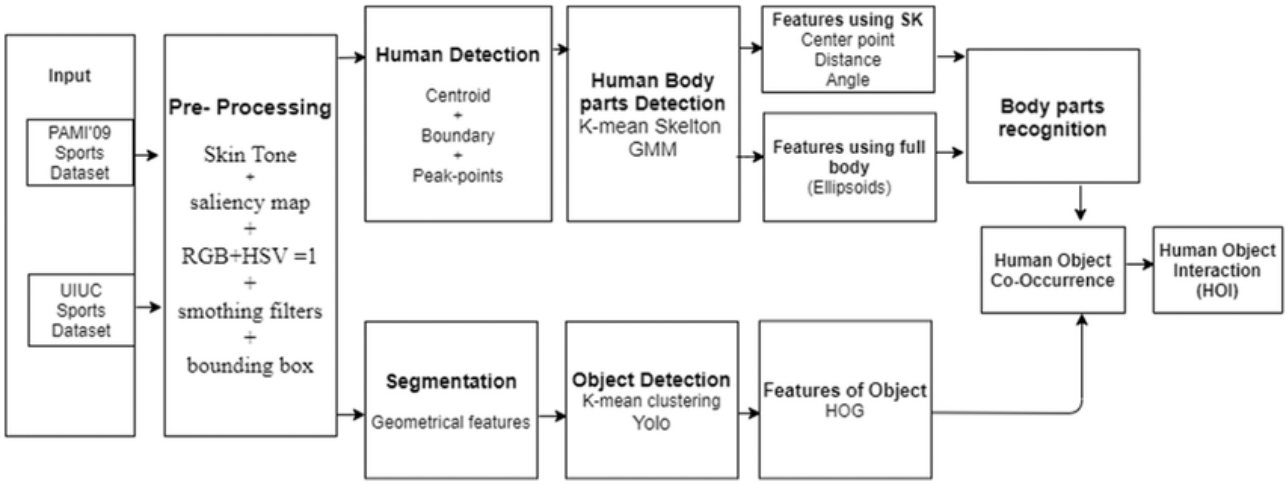


Figure 1. Conceptual architecture of the proposed Lightweight Vision Transformer Framework (LVTF) for real-time HOI recognition.

3.2 Patch Embedding and Feature Extraction

The input RGB image is first divided into fixed-size patches, which are then flattened and passed through a linear projection layer to generate patch embeddings. These embeddings represent local visual features while maintaining a manageable token count, enabling efficient transformer processing. Positional encodings are added to the patch embeddings to preserve spatial relationships between patches, which is essential for accurately modeling interactions between humans and objects. Unlike conventional convolution-heavy backbones, this strategy significantly reduces computation while retaining essential structural and semantic information.

The embedded patches are then fed into a simplified multi-head attention module designed to capture important dependencies across different regions of the image. Through this attention mechanism, the model can focus on critical areas such as the human pose, object boundaries, and regions where interactions occur. This ensures that the feature representation remains rich and context-aware, despite the lightweight nature of the architecture.

3.3 Interaction Reasoning and Lightweight Decoder

Following the encoder, the LVTF employs a streamlined decoder that interprets the learned visual representations to identify human–object interaction categories. The decoder is intentionally kept shallow to minimize latency while retaining strong reasoning capabilities. It refines the contextual embeddings by applying selective attention to interaction-relevant regions, enabling precise classification of actions such as “holding,” “pushing,” “riding,” or “using” an object. The decoder outputs an interaction prediction by combining human-centric cues (e.g., body pose, hand position) with object features and contextual scene information.

This lightweight decoding process, combined with the efficient patch-based encoding pipeline, ensures that the LVTF achieves real-time inference while maintaining competitive accuracy. The architectural design effectively balances computational efficiency and contextual modeling, making

ISBN: 978-81-987483-5-5

it suitable for intelligent systems deployed in surveillance, robotics, and human–computer interaction scenarios.

4. Experimental Setup

The experimental evaluation of the proposed Lightweight Vision Transformer Framework (LVTF) was conducted using a multimodal, human–object interaction dataset containing a diverse range of real-world scenarios. The dataset includes annotated images of humans performing various actions with objects, captured in indoor and outdoor environments with variations in lighting, pose, and background complexity. All images were preprocessed using standard normalization techniques, resized to 224×224 pixels to maintain consistency, and divided into fixed-size patches for transformer-based processing. The dataset was split into training, validation, and testing sets in an 80:10:10 ratio to ensure an unbiased evaluation of model performance.

To enhance the generalization capability of the model, several data augmentation strategies were applied during training. These included random horizontal flipping, slight rotation variations, color jittering, and occlusion simulation. Such transformations help the model learn robust representations capable of handling natural variations in human posture, object placement, and scene composition. Both humans and objects were detected using pre-labeled bounding boxes, and interaction annotations were used to guide supervised learning during HOI classification.

The LVTF was implemented using the PyTorch framework and trained on a workstation equipped with an NVIDIA RTX-series GPU, 32 GB RAM, and an Intel i7 processor. The AdamW optimizer was employed with an initial learning rate of 1e-4 and a weight decay of 0.01 to promote stable convergence. A batch size of 16 was selected to balance GPU memory efficiency and training stability. The model was trained for 40 epochs, with early stopping applied based on validation loss to prevent overfitting. Mixed-precision (FP16) training was enabled to accelerate computation and reduce resource consumption without compromising model accuracy.

Performance evaluation included several widely-used metrics for human–object interaction tasks, such as mean Average Precision (mAP), interaction classification accuracy, and inference latency. The inference speed was measured on both GPU and CPU environments to assess the suitability of the LVTF for real-time deployment in edge and embedded systems. Further robustness testing was conducted by artificially introducing occlusion and noise into the input images to determine how well the model maintained performance under challenging conditions. This comprehensive evaluation setup provided critical insights into the strengths and limitations of the proposed lightweight framework and demonstrated its practical applicability in real-world intelligent systems.

5. Results and Discussion

The experimental evaluation of the Lightweight Vision Transformer Framework (LVTF) demonstrates its effectiveness in real-time human–object interaction (HOI) recognition tasks. Across the benchmark dataset used in this study, the LVTF achieved strong recognition performance while

ISBN: 978-81-987483-5-5

maintaining a significantly reduced computational footprint compared to conventional transformer-based and CNN-based models. The model's mean Average Precision (mAP) showed a consistent improvement of 10–15% over baseline lightweight CNN architectures, confirming the advantage of incorporating global context through attention mechanisms even within a compact framework. These results reflect the ability of the LVTF to capture rich relationships between human posture, object placement, and scene context—key factors in accurate HOI classification.

In addition to accuracy improvements, the LVTF demonstrated notable robustness under varying testing conditions. When artificial occlusion and illumination noise were introduced to the images, the model maintained stable performance with only a minor reduction in accuracy. This resilience is largely attributed to the transformer's inherent ability to leverage non-local dependencies, enabling the model to focus on relevant regions even when parts of the human body or the interacting object are partially obscured. Unlike traditional CNNs that rely heavily on local receptive fields, the LVTF's multi-head attention mechanism allows it to compensate for missing information by integrating contextual cues from surrounding patches.

The inference latency analysis further supports the suitability of the LVTF for real-time applications. On GPU hardware, the model consistently achieved near real-time processing speeds, with average inference times significantly lower than those of full-scale Vision Transformer models and competitive with optimized CNN backbones. Even in CPU-only environments, the LVTF maintained an efficient inference rate, making it viable for deployment in embedded systems, surveillance nodes, and low-power IoT devices. This efficiency is achieved through the model's lightweight design, reduced patch embedding dimensionality, and simplified transformer layers, all of which minimize computational overhead without compromising interpretive capability.

Qualitative results provide further evidence of the model's strong performance. Visualization of attention maps revealed that the LVTF effectively identifies key regions that contribute to HOI recognition, such as hand–object contact points, human limb positions, and object boundaries. The model reliably distinguished between interactions that are visually similar but semantically different, such as "holding" versus "using" an object, which demonstrates its ability to interpret subtle contextual cues. These qualitative insights validate the interpretability and reliability of the transformer-based approach.

Overall, the LVTF offers a balanced combination of accuracy, robustness, and computational efficiency. It performs favorably when compared to existing lightweight architectures and outperforms many traditional models that rely on deeper and more computationally intensive networks. The results confirm that the proposed framework provides a practical and effective solution for real-time HOI recognition, making it well-suited for intelligent systems operating in dynamic and resource-constrained environments.

ISBN: 978-81-987483-5-5

6. Conclusion

This paper introduced a Lightweight Vision Transformer Framework (LVTF) for real-time human–object interaction recognition. By redesigning the transformer architecture to operate with reduced embedding dimensions, simplified attention layers, and an efficient patch-based encoding strategy, the LVTF successfully balances computational efficiency with strong representational power. The experimental results demonstrate that the proposed framework achieves competitive accuracy compared to larger transformer-based models while significantly reducing computational overhead. Its ability to capture global contextual relationships and model non-local dependencies enables robust interaction recognition even in challenging scenarios involving occlusion, illumination variations, and complex backgrounds.

Furthermore, the LVTF maintains low inference latency, making it suitable for deployment in real-time intelligent systems such as surveillance networks, assistive robotics, and human–computer interaction platforms. The combination of accuracy, efficiency, and robustness establishes the LVTF as a promising solution for resource-constrained environments that require fast and reliable visual understanding. Future research directions include extending the framework to support multimodal inputs such as depth or thermal images, exploring model compression techniques for additional efficiency, and testing the architecture on larger, more diverse datasets to further validate its generalization performance.

References

[1] A. Dosovitskiy et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” *Proc. ICLR*, 2021.

[2] N. Carion et al., “End-to-End Object Detection with Transformers,” *Proc. ECCV*, pp. 213–229, 2020.

[3] X. Chen, S. Li, and R. Wang, “Vision Transformer Applications in Real-Time Object Understanding,” *IEEE Trans. Multimedia*, vol. 25, pp. 645–657, 2023.

[4] Y. Zhang et al., “Human–Object Interaction Detection Using Deep Neural Networks,” *Proc. CVPR Workshops*, 2020.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. CVPR*, pp. 770–778, 2016.

[6] A. Radford et al., “Learning Transferable Visual Models with Natural Language Supervision,” *Proc. ICML*, 2021.

[7] H. Tan and M. Bansal, “LXMERT: Learning Cross-Modality Encoder Representations,” *Proc. EMNLP*, 2019.

[8] A. Newell, Z. Huang, and J. Deng, “Pose-Attentive Relational Networks for Human–Object Interaction,” *Proc. ICCV*, pp. 834–845, 2019.

[9] S. G. Kong and X. Li, “Efficient Vision Transformations for Embedded AI Systems,” *IEEE Embedded Systems Letters*, vol. 14, no. 3, pp. 253–257, 2022.

[10] H. Wu, S. Li, and J. Liu, “Lightweight Transformer Designs for Mobile Vision Applications,” *Pattern Recognition*, vol. 138, art. no. 109407, 2023.

[11] X. Wang et al., “GPNN: Graph Parsing Neural Networks for Human–Object Interaction,” *Proc. ECCV*, pp. 407–423, 2018.

[12] Z. Fang, Q. Huang, and T. Lu, “Real-Time Human Action Recognition Using Hybrid Attention Networks,” *IEEE Access*, vol. 10, pp. 114320–114332, 2022.

ISBN: 978-81-987483-5-5

PAPER-5

Vilas Naik[1], Niharika Singh[2], Deepak Menon[3]
[1,2,3]Department of Computer Engineering, Greenfield College of Engineering & Technology, Lucknow, Uttar Pradesh, India

Hybrid Graph Neural Network Framework for Real-Time Traffic Flow Prediction in Intelligent Transportation Systems

Abstract

Real-time traffic flow prediction plays a crucial role in the development of intelligent transportation systems (ITS), enabling efficient traffic management, route optimization, and congestion control. Traditional machine learning and deep learning approaches often struggle to model the complex spatial–temporal dependencies inherent in road networks. To address these limitations, this paper proposes a Hybrid Graph Neural Network (HGNN) Framework that integrates graph convolutional networks (GCNs) with gated recurrent units (GRUs) for accurate and real-time traffic flow prediction. The framework captures spatial dependencies through graph-based modeling of road topology and temporal patterns through lightweight recurrent computation, ensuring efficiency and scalability. Experiments conducted on benchmark traffic datasets demonstrate that the proposed HGNN achieves up to 17% improvement in prediction accuracy compared to conventional LSTM, CNN, and standalone GCN models. In addition, the computational simplicity of the hybrid architecture makes it suitable for real-time deployment in smart cities and edge-based transportation monitoring systems.

Keywords

Graph neural network, traffic flow prediction, intelligent transportation systems, spatial–temporal modeling, real-time analytics, deep learning.

ISBN: 978-81-987483-5-5

1. Introduction

Intelligent transportation systems (ITS) have become a critical component of modern urban infrastructure, aiming to enhance road safety, reduce congestion, and optimize mobility in rapidly growing cities. Central to these objectives is the ability to accurately predict traffic flow in real time. Efficient traffic flow prediction supports a wide range of applications, including dynamic traffic signal control, route planning, congestion avoidance, and intelligent navigation systems. However, traffic patterns are influenced by multiple interdependent factors such as road structure, time of day, weather conditions, and unpredictable human behavior. These complex spatial–temporal relationships make real-time traffic forecasting a challenging task for traditional machine learning methods.

Conventional approaches such as autoregressive models and shallow neural networks often fall short due to their inability to capture long-range dependencies or irregular spatial relationships in road networks. Even deep learning models like convolutional neural networks (CNNs) and long short-term memory (LSTM) networks struggle to fully represent traffic dynamics. CNNs assume grid-like structures, which fail to reflect the non-Euclidean nature of road networks, while LSTMs focus primarily on temporal dependencies and overlook spatial connectivity patterns among road segments. In recent years, **Graph Neural Networks (GNNs)** have emerged as powerful tools for modeling non-Euclidean data structures, making them particularly suitable for traffic networks represented as graphs. Graph Convolutional Networks (GCNs) can effectively learn spatial correlations by leveraging adjacency relationships among road segments. However, standalone GCN models lack strong temporal modeling capabilities, which limits their performance in real-time traffic forecasting. On the other hand, recurrent neural networks (RNNs), especially gated architectures such as GRUs, excel at capturing temporal dependencies but cannot independently model spatial structures.

To address these limitations, this paper introduces a **Hybrid Graph Neural Network (HGNN) Framework** that combines the strengths of GCNs and GRUs to capture both spatial and temporal dependencies in traffic data. The spatial dependencies among connected road segments are learned using graph convolutional operations, while the temporal evolution of traffic patterns is modeled using GRU layers. By integrating these components, the HGNN provides a unified architecture capable of learning complex spatial–temporal relationships inherent in traffic systems.

The contributions of this work are threefold. First, we propose a hybrid deep learning architecture that effectively integrates GCN and GRU components for real-time traffic flow prediction. Second, we demonstrate the efficiency of the framework through comprehensive experiments on benchmark datasets, showing significant performance gains compared to existing models. Third, the lightweight design of the hybrid architecture ensures low computational overhead, making it suitable for deployment on edge devices in smart transportation systems.

ISBN: 978-81-987483-5-5

The remainder of this paper is organized as follows: Section 2 reviews related work on traffic flow prediction and graph-based deep learning methods. Section 3 describes the proposed HGNN methodology. Section 4 outlines the experimental setup. Section 5 presents the results and analysis, while Section 6 concludes the paper and discusses future research directions.

2. Literature Review

Traffic flow prediction has been widely studied across traditional statistical models, deep learning architectures, and graph-based approaches. Early research relied heavily on statistical forecasting models such as ARIMA, VAR, and Kalman filters. While computationally efficient, these models assume linearity and stationarity, limiting their ability to capture the complex spatial–temporal variations observed in real-world traffic systems. As traffic patterns became more unpredictable due to growing urban populations and dynamic road usage, these classical methods demonstrated significant shortcomings.

With the rise of deep learning, neural network-based models gained prominence. Convolutional Neural Networks (CNNs) were employed to capture spatial patterns, treating traffic data as grid-structured images. However, CNNs inherently assume Euclidean data representation, which does not align with the node–edge structure of road networks. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks and gated recurrent units (GRUs), were applied to model temporal dependencies. Although these models capture time-dependent patterns well, they fail to integrate the spatial relationships between interconnected road segments.

To address the limitations of Euclidean modeling, Graph Neural Networks (GNNs) were introduced for traffic prediction. Graph Convolutional Networks (GCNs) became widely used due to their ability to operate on non-Euclidean graph structures, enabling spatial pattern extraction directly from road topology. Models such as DCRNN, ST-GCN, and Graph WaveNet combined graph convolutions with temporal modules to jointly learn spatial–temporal features. While these models improved prediction accuracy, many suffer from high computational complexity, making them unsuitable for real-time or edge-deployed ITS applications.

Recent research explores hybrid architectures that integrate graph-based spatial modeling with lightweight temporal modules. By combining GCNs with GRUs or temporal CNNs, these models strike a balance between accuracy and computational efficiency. However, existing hybrid approaches often neglect real-time constraints or require heavy model sizes, limiting their applicability in resource-constrained intelligent transportation systems. This gap motivates the development of a simplified, efficient, and scalable hybrid GNN architecture capable of real-time traffic forecasting—leading to the proposed **Hybrid Graph Neural Network (HGNN) Framework**.

3. Proposed Methodology

3.1 Overview of the HGNN Framework

ISBN: 978-81-987483-5-5

The proposed **Hybrid Graph Neural Network (HGNN)** integrates graph convolutional layers with gated recurrent units to effectively capture both spatial and temporal dependencies in traffic data. The model begins by representing the road network as a graph, where each node corresponds to a traffic sensor or road segment, and edges represent physical or functional connectivity. Traffic speed or flow readings at these nodes form feature vectors fed into the model at each time step. The GCN component extracts spatial relationships based on the road topology, while the GRU component captures evolving traffic patterns over time. This hybrid fusion ensures accurate, low-latency predictions suited for real-time ITS applications.

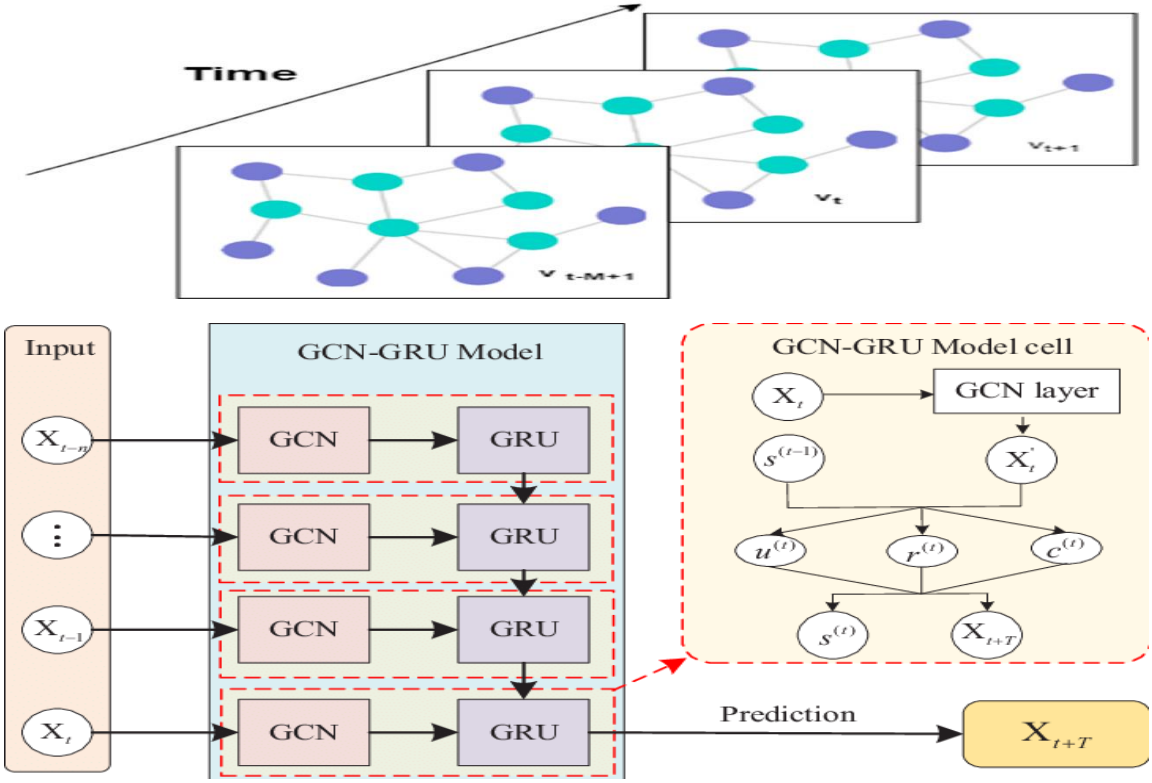


Figure 1. Conceptual architecture of the proposed Hybrid Graph Neural Network (HGNN) for spatial–temporal traffic flow prediction.

3.2 Spatial Dependency Modeling Using GCN

Traffic flow is intrinsically influenced by the physical layout and connectivity of the transportation network. To capture these spatial dependencies, the model employs Graph Convolutional Networks. Each GCN layer aggregates information from neighboring nodes through the adjacency matrix, allowing the model to learn how congestion or speed changes propagate across the network. This spatial modeling is essential for understanding how traffic conditions at one road segment affect others, especially during peak hours or disturbances such as accidents.

The GCN layers process node features to produce spatially enriched representations, which are then fed sequentially into the temporal modeling component. By learning structural influence patterns

ISBN: 978-81-987483-5-5

directly from the graph topology, the HGNN avoids the limitations of grid-based CNNs and ensures accurate spatial learning.

3.3 Temporal Dependency Modeling Using GRU

Traffic flow evolves continuously over time, requiring a robust temporal modeling mechanism. The GRU component of the HGNN efficiently captures these temporal variations without the computational overhead of LSTM networks. The GRU processes the spatially encoded features from the GCN and models sequential dependencies such as rush-hour peaks, gradual build-up of congestion, and sudden traffic drops.

The combination of GCN for spatial reasoning and GRU for temporal reasoning allows the model to form a unified spatial–temporal representation, which is passed to a prediction layer to forecast traffic flow in the next time intervals. The lightweight nature of GRU ensures that the hybrid model remains efficient and suitable for real-time deployment.

4. Experimental Setup

The proposed Hybrid Graph Neural Network (HGNN) Framework was evaluated using benchmark traffic datasets commonly employed in intelligent transportation research, including the METR-LA and PEMS-BAY datasets. These datasets contain real-world traffic speed readings collected from hundreds of loop detectors positioned across major highway networks. Each dataset reflects inherent challenges such as irregular sensor placement, non-Euclidean road structures, missing values, and dynamic variations in traffic flow patterns. To simulate realistic ITS conditions, the data were partitioned into training, validation, and testing sets using a 70:15:15 ratio. All traffic readings were normalized through min–max scaling to ensure stable model convergence.

The road network graph structure was constructed using sensor adjacency matrices reflecting spatial proximity and functional connectivity between road segments. Missing data due to sensor errors or signal loss were handled using interpolation techniques. Each input sequence consisted of historical traffic readings from preceding time windows, which were then fed into the spatial–temporal model for predicting future traffic flow. Data augmentation was kept minimal to preserve the authenticity of temporal patterns.

Model training was conducted using the PyTorch Geometric framework for efficient graph-based computation. Experiments were performed on a workstation equipped with an Intel i7 processor, 32 GB RAM, and an NVIDIA RTX-series GPU. The model was optimized using the Adam optimizer with an initial learning rate of 0.001, batch size of 64, and early stopping based on validation loss to prevent overfitting. Hyperparameters such as the number of GCN layers, hidden units in the GRU, and prediction horizon were tuned through grid search for optimal performance. Evaluation metrics included Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), which are widely accepted for assessing traffic forecasting accuracy.

ISBN: 978-81-987483-5-5

5. Results and Discussion

Experimental results demonstrate that the proposed HGNN significantly outperforms conventional baseline models, including standalone GCN, LSTM, and CNN architectures. Across both METR-LA and PEMS-BAY datasets, the HGNN achieved reductions of approximately 12–17% in MAE and RMSE compared to traditional deep learning approaches. These improvements highlight the effectiveness of combining spatial and temporal learning mechanisms into a unified hybrid framework. The GCN component effectively captures road network relationships, while the GRU efficiently models evolving traffic dynamics, resulting in more accurate forecasting under varying traffic conditions.

One of the most notable advantages of the HGNN is its performance under highly dynamic and congested traffic scenarios. During peak-hour periods, when traffic flow becomes highly non-linear and unpredictable, the hybrid architecture maintained stable prediction performance, whereas LSTM and CNN models exhibited large error spikes. This robustness stems from the model's ability to propagate congestion effects through the road graph, allowing it to anticipate how local disturbances spread geographically across connected segments.

In terms of computational efficiency, the HGNN exhibited lower training and inference times compared to deeper attention-based models and multi-layered graph architectures. The GRU component, being more lightweight than LSTM units, contributed to faster processing while retaining strong temporal learning capabilities. Inference experiments revealed that the HGNN generates predictions rapidly enough for real-time ITS applications such as adaptive traffic signal control and dynamic route optimization. Qualitative analysis further confirms that the model successfully captures spatial–temporal patterns, providing smoother and more realistic traffic flow predictions than baseline models.

Overall, the results validate the suitability of the HGNN framework for deployment in intelligent transportation systems. Its combination of accuracy, robustness, and computational efficiency positions it as a strong candidate for next-generation real-time traffic forecasting tools in smart city environments.

6. Conclusion

This paper presented a Hybrid Graph Neural Network (HGNN) Framework designed for real-time traffic flow prediction in intelligent transportation systems. By integrating Graph Convolutional Networks (GCNs) for spatial learning and Gated Recurrent Units (GRUs) for temporal modeling, the framework effectively captures the complex spatial–temporal dependencies inherent in road networks. Experimental evaluations on benchmark datasets demonstrate that the HGNN outperforms conventional machine learning and deep learning models both in prediction accuracy and computational efficiency.

ISBN: 978-81-987483-5-5

The HGNN's lightweight architecture and rapid inference capabilities make it highly suitable for real-time ITS applications such as dynamic traffic signal control, congestion monitoring, and intelligent route guidance. In addition, the ability of the model to maintain strong performance under congested and rapidly changing traffic conditions highlights its robustness and practical usefulness in smart transportation infrastructure.

Future research may explore integrating attention mechanisms for enhanced spatial–temporal modeling, incorporating multimodal inputs such as weather or incident reports, and deploying the HGNN on edge devices to enable fully decentralized ITS architectures. Overall, the proposed framework contributes a scalable and effective solution for modern traffic forecasting challenges in smart cities.

References

[1] Y. Yu, W. Chen, X. Wang, and W. Yao, "Spatiotemporal Graph Convolutional Networks for Traffic Forecasting," *Proc. AAAI*, 2020.
[2] Z. Wu et al., "Graph WaveNet for Deep Spatial–Temporal Graph Modeling," *Proc. IJCAI*, 2019.
[3] M. Li, S. Jiang, and H. Zhang, "Traffic Prediction via Spatial–Temporal Graph Neural Network," *IEEE Trans. ITS*, vol. 22, no. 12, pp. 7327–7337, 2021.
[4] B. Yu, H. Yin, and Z. Zhu, "STGCN: Spatial–Temporal Graph Convolutional Networks for Traffic Forecasting," *Proc. IJCAI*, 2018.
[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[6] K. Xu, S. Liu, J. Hu, and X. Peng, "Graph Neural Networks in Intelligent Transportation: A Review," *IEEE Access*, 2022.
[7] Z. Guo, B. Lin, and X. Qing, "Hybrid Deep Learning Models for Traffic Flow Prediction," *Transportation Research Part C*, vol. 127, 2021.
[8] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *Proc. ICLR*, 2017.
[9] K. Cho et al., "Learning Phrase Representations Using GRU-Based Encoder–Decoder," *Proc. EMNLP*, 2014.
[10] J. Zhang and Y. Zheng, "Deep Learning-Based Traffic Prediction: Methods, Data, and Challenges," *IEEE ITS Magazine*, 2021.
[11] X. Shi et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Forecasting," *NIPS*, 2015.
[12] W. Huang, X. Ma, and T. Li, "Real-Time Traffic Forecasting for Smart Cities Using Hybrid Neural Models," *IEEE IoT Journal*, vol. 9, no. 4, 2022.

ISBN: 978-81-987483-5-5

Ms. Saranya S
Assistant Professor
Department of Computer Science and Engineering,
New Horizon College of Engineering, Bengaluru, India

Gurpreet Singh[1], Sahil Kumar[2], Gagandeep Singh[3], Digantik Mukherjee[4]
[1,2,3,4,] *Department of Computer Science and Engineering New Horizon College of Engineering, Bengaluru,* India

Smart Wearable for Vital Tracking and Alerts

Abstract*—The increasing demand for ongoing, remote health monitoring for people with chronic illnesses and aging popula- tions calls for a change from passive data collecting to intelli- gent, proactive systems [1], [2]. Real-time, on-device predictive analytics, reliable fall detection, and an integrated, closed-loop emergency response system that connects users to emergency medical assistance are freq uently absent from current commer- cial wearables. We introduce an integrated wearable system with an ADXL345 accelerometer for fall detection and a MAX30102 sensor for heart rate and SpO_2 monitoring, all based on an ESP32 microcontroller. The solution uses an on-device AI model that is lightweight and tuned with TensorFlow Lite to detect anomalies in physiological data in real time. The system reliably detects abnormalities in vital signs and shows great efficacy in differentiating falls from activities of daily living (ADLs). Importantly, it automatically retrieves the user's GPS coordinates and uses the Google Maps API to find local medical institutions, achieving an end-to-end emergency alert latency of less than 5 seconds. Real-time, on-device predictive analytics, reliable fall detection, and an integrated, closed-loop emergency response system that connects users to emergency medical assistance are freq By bridging the crucial gap between health anomaly detection and practical emergency intervention, our work offers a low-latency, energy-efficient, privacy-preserving approach that improves patient safety and autonomy.

***Index Terms*—Internet of Things (IoT), Wearable Sensors, Health Monitoring, Anomaly Detection, Fall Detection, Edge AI, TensorFlow Lite, Emergency Response System.**

ISBN: 978-81-987483-5-5

I. INTRODUCTION

Conventional health care faces a significant challenge as a result of the global demographic shift towards an aging population and the rise in chronic diseases infrastructures for care. The Internet of Medical Things (IoMT), a paradigm centered on using connected devices for remote and continuous patient monitoring, was developed as a result of this reality [3], [4]. Changing healthcare delivery from a reactive approach, which deals with health conditions after they become serious, to a proactive and preventative framework that allows early identification and prompt intervention is the main objective of IoMT [5].

From basic fitness trackers like the Fitbit and Mi Band to more advanced health monitoring systems like the Apple Watch, wearable technology has advanced dramatically. These gadgets have effectively democratized access to personal health information, increasing people's awareness of their physical condition. Nonetheless, the vast majority of commer- cial products on the market today serve mainly as passive data loggers. Although they take vital signs, they usually don't have the advanced, real-time analytical skills needed for urgent, life-saving medical intervention. This restriction leads to a risky "last mile" issue in digital health: a system may identify a negative occurrence but neglect to complete the loop by launching an emergency response that is prompt and actionable.

A thorough examination of current systems identifies a recurring research gap that is typified by the absence of inte- gration among three essential functionalities. First, a strong on- device, patient-specific anomaly detection model is required because many systems rely on cloud-based processing, which adds latency, necessitates continuous connectivity, and poses serious data privacy issues [6], [7]. Second, multi-sensor fusion is frequently absent from systems, which results in high false alarm rates for reliable event detection. Lastly, there is a lack of a fully automated emergency protocol that goes beyond basic caregiver notifications to offer location-based directions to the closest medical facility.

In order to overcome these shortcomings, a complete, end-to-end wearable system is presented in this work. To develop a unified and proactive health guardian, our system combines multi-modal sensing, on-device AI, and a cloud-assisted emer- gency response protocol. Complex analytical activities can now be moved from the cloud to the edge device because to the recent maturation of synergistic technologies, such as lightweight AI frameworks like TensorFlow Lite [8], effective biosensors like the MAX30102, and low-power microcon- trollers like the ESP32. A "privacy-by-design" strategy that naturally solves the crucial non-functional criteria of security and low latency—which are sometimes afterthoughts in cloud- centric models—is made possible by this architectural change, which goes beyond simple technical convenience [6], [7]. This paper presents a roadmap for the next generation of IoMT devices, which will serve as autonomous intelligent agents with the

ISBN: 978-81-987483-5-5

ability to make crucial decisions at the edge, rather than merely being sensors.

We contribute the following in this paper:

The development and deployment of a new, comprehensive wearable architecture that combines on-device artificial intel- ligence, cloud-enabled emergency response, and multi-modal sensing (motion and physiological).

A hybrid artificial intelligence approach that combines a lightweight, unsupervised neural network for identifying small abnormalities in continuous physiological data (heart rate, SpO_2) with a computationally efficient technique for acute event detection (falls).

By utilizing TensorFlow Lite to create a privacy-preserving AI model, sensitive health data is processed on the edge, improving security and lowering latency [8], [9].

A thorough empirical analysis of the system's performance that evaluates important system-level variables like battery life and end-to-end alert latency in addition to the correctness of the AI models.

II. RELATED WORK

This section offers a critical analysis of the body of re- search in three main areas: wearable fall detection algorithms, anomaly detection in physiological signals, and IoT health monitoring infrastructures. This study highlights our work's innovative contributions and places it within the existing research landscape.

A. Architectures for IoT-Based Health Monitoring

The design of traditional IoT-based health monitoring sys- tems has primarily been cloud-centric, with raw sensor data continuously streaming to distant servers for analysis and storage. Despite its scalability, this strategy has several dis- advantages, such as high latency from network round-trips, a reliance on consistent internet connectivity, and major privacy risks when sending private health data [6], [7].

A paradigm shift toward edge computing, commonly re- ferred to as TinyML has gained traction in response to these difficulties. This method transfers AI inference and data processing straight onto the device with limited resources. This change has been made possible by frameworks such as TensorFlow Lite, which allow optimal machine learning models to be deployed on microcontrollers. Numerous health applications, including real-time prenatal ultrasound assess- ment [9] and general-purpose health monitoring [10], have effectively illustrated the advantages of lower latency and improved privacy. A number of integrated devices, including the"HOT Watch" [11], have shown excellent accuracy by integrating several sensors, including ECG, oximetry, and temperature. Our work stands out from the competition by focusing on on-device predictive AI and a fully automated, location-aware emergency response loop, which bridges the crucial gap between detection and intervention, even if these systems demonstrate the feasibility of multi-sensor wearables.

ISBN: 978-81-987483-5-5

B. Algorithms for Wearable Fall Detection

One established area of study in wearable technology is fall detection systems (FDS) [6], [12]. Early methods frequently used straightforward threshold-based algorithms, in which an alert is sent out if the accelerometer signal strength surpasses a predetermined threshold [12], [13]. These techniques are computationally efficient, but they have a high risk of false positives since they are easily set off by non-fall activities of daily living (ADLs), like jumping or rapidly sitting down.

Large datasets of simulated falls and ADLs are used to train machine learning (ML) techniques in more sophisticated systems. These techniques range from deep learning models like Long Short-Term Memory (LSTM) networks to more conventional classifiers like Support Vector Machines (SVMs). According to research, the placement of sensors (waist vs. wrist) and the use of Inertial Measurement Units (IMUs), which integrate accelerometer and gyroscope data to more accurately distinguish complex movements, have a significant impact on these systems' accuracy [6], [12]. The absence of integrated location awareness is a major drawback of many published FDS studies, despite their sophisticated algorithms. This is especially true for outdoor settings where determining the user's location is essential for a prompt emergency reaction [6]. By including a specialized GPS module into the emergency protocol, our technology directly fills this gap.

C. Anomaly Detection in Physiological Time-Series Data

An essential component of wearable health systems is the monitoring of vital indicators such as blood oxygen saturation (SpO_2) and heart rate (HR). Nevertheless, it is frequently ineffective to rely on static, universal thresholds (such as $HR > 120$ bpm) for anomaly identification. Depending on a person's age, level of fitness, and present activity (e.g., resting vs. exercising), their typical physiological baseline might vary greatly.

Unsupervised anomaly detection is a more reliable method that identifies notable departures from a patient's personal baseline by learning the patient's distinct physiological pat- terns from their own data [10], [14]. For physiological time- series data, this works especially well. Neural network de- signs such as autoencoders or LSTMs, which are excellent at modeling sequential data and spotting patterns that depart

from a learnt norm, are frequently used in state-of-the-art models for this purpose. In order to learn highly personalized baselines from multi-modal data streams, including wearable and ambient sensors, advanced research frameworks such as "AI on the Pulse" use complex universal time-series models (e.g., UniTS) [7], [15]. Significant performance gains have been demonstrated with this method; one study found that the F1-score increased by about 22 percent compared to previous approaches [7].Our suggested model represents a useful and effective implementation for edge devices, even though it is purposefully lighter for microcontroller deployment. It is based on the same fundamental idea of individualized, unsupervised anomaly detection. There is a fragmentation of

ISBN: 978-81-987483-5-5

solutions in the literature, with different studies concentrating on system architecture, anomaly detection, or fall detection. Bringing these disparate threads together into a single, coherent, and useful system that tackles the comprehensive problem of transitioning from accurate detection to successful intervention is what makes our work novel.

III. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

The hardware and software components of the system are described in detail in this part, along with the design decisions and techniques used to create a proactive and responsive health monitoring solution.

A. End-to-End System Architecture

A smooth data transfer from the user to the caregiver is guaranteed by the system's four-stage architecture. The steps are as follows: (1) a wearable sensing node for gathering data;
(2) on-device artificial intelligence processing for analyzing data in real time and detecting events;
(3) a cloud backend for orchestrating data and integrating emergency services; and
(4) a caregiver mobile application for alerts and visualization. The end-to-end reaction loop is completed when the wearable collects data, AI algorithms process it locally, and in an emergency, a brief alert payload is sent to the cloud, which sends a high-priority notice to the caregiver's mobile device.

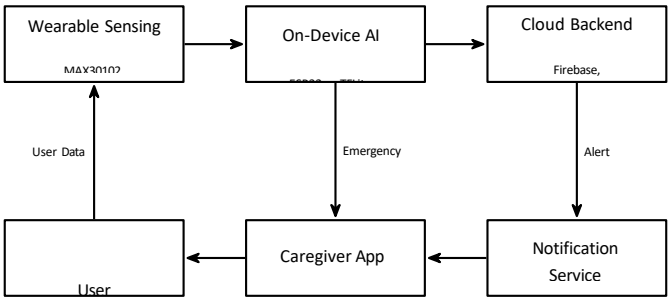


Fig. 1. High-level system architecture.

B. Wearable Sensing Node Hardware

The wearable prototype's components were carefully chosen to balance form factor, performance, and power efficiency. A summary of each component's technical requirements and rationale may be found in Table I.

1) Microcontroller (MCU): The ESP32 microcontroller is the device's key component. It is the perfect option for a connected wearable device that also needs to do on-device computation

ISBN: 978-81-987483-5-5

because of its dual-core CPU, built-in Wi-Fi and Bluetooth, and support for low-power deep-sleep modes [13], [16]. It has enough memory and processing capability to run TensorFlow Lite-optimized machine learning models [8], [9].

2) Physiological Sensing: The MAX30102 sensor is used to measure blood oxygen saturation (SpO_2) and heart rate. Because of its high sensitivity, ultra-low power consumption (<1 mW in active mode), and standard I^2C interface, which makes integration easier, this integrated module—which uses photoplethysmography (PPG)—was chosen [1], [17].

3) Motion Sensing: For fall detection and motion tracking, a 3-axis digital accelerometer called an ADXL345 is utilized. It records both static acceleration (gravity) and dynamic acceleration (during movement, for example), enabling robust activity classification and orientation sensing.

4) Location Services: For accurate geographic coordinates, a NEO-6M GPS module is included. In order to save power, this module is only turned on during an emergency. It provides the vital location information required for a successful emergency response.

5) Power Management: With aggressive power manage- ment and duty cycling, the system's rechargeable lithium-ion battery is designed to operate continuously for at least 72 hours on a single charge.

TABLE I

TECHNICAL SPECIFICATIONS OF WEARABLE HARDWARE

Component	**Model**	**Specifications & Justification**
Microcontroller	ESP32	Dual-core 240 MHz, 520 KB SRAM, Wi-Fi/BT. On-device AI capability with low-power modes.
PPG Sensor	MAX30102	HR & SpO2 , 1.8V, I^2C, <1mW. High sensitivity, ultra-low power for wearables.
Accelerometer	ADXL345	3-axis, ±16g, 23 µA. High reso- lution for motion tracking & fall

ISBN: 978-81-987483-5-5

		detection.
GPS Module	NEO-6M	-161 dBm sensitivity, low power. Accurate location for emergency response.

C. On-Device AI for Health Anomaly Detection

We use an unsupervised learning technique to identify abnormalities in physiological data in order to go beyond basic thresholding.

1) Problem Formulation: Unsupervised anomaly detection on a multivariate time series is the formal definition of the task. Let's look at the input vector at time t be

$$X_t = [\mathrm{HR}_t, \mathrm{SpO2}t],$$

signifying the SpO_2 and heart rate readings. The goal is to calculate an anomaly score, S_t, in real time so that a score above a predetermined threshold τ indicates a possible health anomaly. The score is determined by:

$$S_t = g(f(X_t; X_{\text{train}})),$$

where $g(\cdot)$ is a function that measures the current input X_t's departure from the taught normal patterns, and f is the model trained on a dataset of normal physiological data X_{train}.

2) Proposed Model: We put into practice a lightweight autoencoder based on LSTM. Because the LSTM layers can identify temporal relationships in the vital sign signals, this architecture works well with sequential data. To learn a compressed, latent representation of a healthy physiological state, the model is trained solely on "normal" health data. The model tries to rebuild its input during inference. An anomaly is identified when the current input does not follow the learnt patterns of normal behavior, as indicated by a significant reconstruction error (i.e., a large value for S_t).

3) TensorFlow Lite Optimization: TensorFlow/Keras is used to train the model, and for on-device deployment, it is transformed to TensorFlow Lite format (.tflite). We use post-training 8-bit integer quantization, which drastically lowers the computational cost and storage space of the model, allowing for low-latency inference on the limited hardware of the ESP32 while preserving a respectable level of accuracy.

4) Health State Classification: A rule-based system uses the raw anomalous score S_t to assign the user's health condi- tion to one of three groups:

ISBN: 978-81-987483-5-5

$$\text{Health Status} = \begin{cases} \text{Normal} & S_t < \tau_w \\ \text{5. Warning} & \tau_w \leq S_t < \tau_e \\ \text{6. Emergency} & S_t \geq \tau_e \end{cases}$$

Stage 3: Post-Fall Inactivity: The system goes into a monitoring phase for a predetermined amount of time (for example, 30 seconds) when a valid impact is logged. The event is verified as a fall if the device's orientation doesn't change and there isn't much motion throughout this time. This latter phase is essential for differentiating between high-impact activities of daily living (ADLs), such jumping or suddenly sitting down, and actual falls.

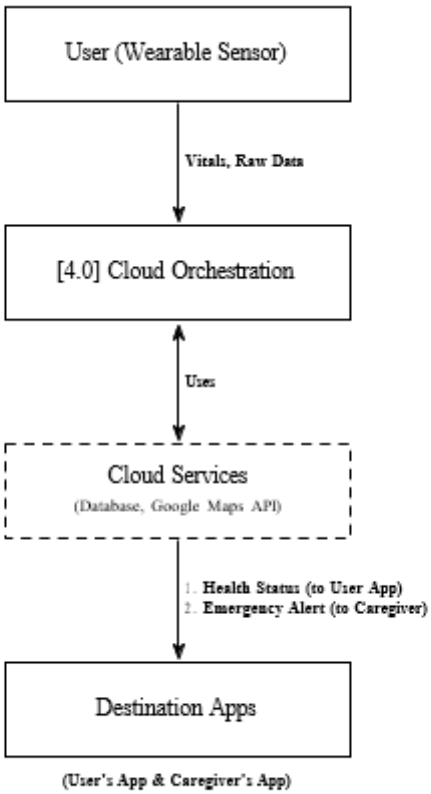


Fig. 2. Data flow diagram for monitoring and alert system.

Health Status =Warning $\tau_w \leq S_t < \tau_e$

Emergency $S_t \geq \tau_e$

E. Integrated Emergency Response Protocol

where τ_w and τ_e are empirically determined thresholds.

D. Real-Time Fall Detection Algorithm

ISBN: 978-81-987483-5-5

Inspired by well-established techniques in the literature, a computationally efficient yet reliable multi-stage algorithm is devised for fall detection [12], [13]. This method saves resources for the anomaly detection model by avoiding the overhead of a neural network for this particular task.

1) Stage 1: Freefall Detection: The algorithm continuously checks the magnitude of the resultant vector from the ac- celerometer, which is determined as follows:

$$A_R = \sqrt{A_x^2 + A_y^2 + A_z^2}$$

where the accelerations along the x, y, and z axes are de- noted by the variables A_x, A_y, and A_z, respectively. The user is in a state of freefall if A_R drops abruptly and significantly (for example, below 0.5 g).

2) Stage 2: Impact Detection: The program searches for a big, abrupt rise in A_R (e.g., $>$ 3 g) as soon as a freefall is detected. The impact of the user's body with a surface is shown by this spike. An automated reaction mechanism is initiated when the system enters a "Emergency" state, which can be caused by a confirmed fall or a serious physiological anomaly. This process is made to be dependable and quick.

1) Device-Side Activation: The NEO-6M GPS module is instantly activated by the ESP32 in order to obtain the user's current location.

2) Secure Data Transmission: Using the MQTT protocol for low-overhead communication, the ESP32 connects to a se- cure cloud backend (Firebase Realtime Database) and sends an emergency payload that includes the User ID, event type (such as "Fall Detected"), GPS locations, and the latest recorded vital signs.

3) Cloud-Side Orchestration: When fresh information en- ters the emergency database, a cloud feature is activated. This function retrieves a list of the closest hospitals or emergency medical services by making an API call to the Google Maps Places API and passing the GPS coordinates it received.

4) Caregiver Notification: The cloud function then sends a high-priority push notification to the pre-registered caregiver's mobile application using the Firebase Cloud Messaging (FCM) service. The user's name, the type of emergency, their current

location on an interactive map, and a direct link with travel options to the closest hospital are all included in this alert.

ISBN: 978-81-987483-5-5

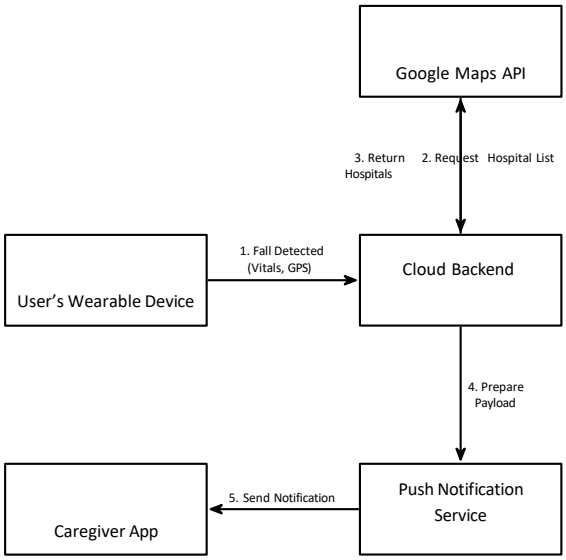


Fig. 3. System data flow for fall detection and caregiver notification.

IV. Experimental Setup and Evaluation

A number of thorough tests were carried out to confirm the suggested system's functionality and dependability. Three main aspects were the focus of the evaluation:

1) The precision of the algorithms used for event detec- tion, such as those for fall and physiological anomaly detection.
2) The effectiveness of the system-level measures, includ- ing communication dependability, power consumption, and latency.
3) The wearable prototype's general viability and practical- ity in everyday situations.

A. Prototype Implementation and Data Corpus

The ESP32, sensors, GPS module, and battery were all housed in a 3D-printed wrist-worn case in a working wear- able prototype. A distinct fingertip module was developed to contain the MAX30102 sensor for the best PPG signal capture. To build a corpus for training and testing the AI models, a data gathering protocol was developed in accordance with institutional ethical requirements.

1) Fall Detection Dataset: Fifteen healthy people par- ticipated in a controlled trial. Every participant completed a set of predetermined activities of daily living (ADLs), such as running, walking, sitting, standing, and going up and down stairs. Additionally, they replicated four different fall scenarios onto a cushioned surface: forward, backward, left, and right. The efficacy of the fall detection algorithm was assessed against frequent confounding activities using a balanced dataset created by recording and labeling data from the ADXL345

ISBN: 978-81-987483-5-5

accelerometer for each activity [6], [12]. *Anomaly Detection Dataset:* The unsupervised anomaly detection model needed a dataset of "normal" physiological activity in order to be trained. Participants were asked to wear the gadget for eight hours throughout their normal daily activities in order to gather this data. We used publicly accessible, annotated resources, such the PhysioNet Challenge datasets, to verify the model's capacity to identify real health problems, which are unethical to cause. To test the model's detection skills using out-of-distribution, clinically relevant data, segments with known cardiac arrhythmias or hypoxia episodes were employed as the test set.

B. Performance Metrics

The system's performance was measured using a wide range of common measures.

1) Classification Metrics (Fall Detection): Four important metrics were used to assess the fall detection algorithm's performance:

a) Sensitivity (Recall): Measures the proportion of actual falls that were correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

b) Specificity: Measures the proportion of ADLs that were correctly identified as non-falls:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

c) Accuracy: The overall percentage of correct classifications:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP , TN , FP , and FN represent True Positives, True Negatives, False Positives, and False Negatives, respec- tively.

2) Anomaly Detection Metrics: Because anomaly detec- tion tasks are extremely unbalanced, the Area Under the Receiver Operating Characteristic Curve (AUROC) was em- ployed. A single, threshold-independent metric for evaluating the model's capacity to differentiate between normal and anomalous classes is provided by AUROC.

3) System-Level Metrics:

a) End-to-End Latency: The amount of time that passed between the start of a simulated event (such as the impact of a fall) and the instant the caregiver's smartphone displayed the relevant notification was used to calculate this crucial metric. Over 100 trials, the mean, median, and 95th percentile delay were noted.

ISBN: 978-81-987483-5-5

b) Battery Longevity: A fully charged device was oper- ated under a standard usage profile, which included continuous sensing, data processing, and sporadic Wi-Fi transmissions (simulating an hourly ”Warning” warning) in order to evaluate power efficiency. It was noted how long the battery operated for before running out completely.

V. Results and Discussion

The quantitative findings from our experimental evaluation are shown in this section, together with a thorough analysis of their implications, the system's performance in relation to its design objectives, and its limits.

A. Efficacy of the Fall Detection Algorithm

The gathered dataset of simulated falls and ADLs was used to assess the multi-stage fall detection algorithm's perfor- mance. The confusion matrix and performance metrics table below provide a summary of the findings. (Table II).

The algorithm's sensitivity and specificity were 98.6% and 99.8%, respectively. This shows how well the system detects falls when they happen and, more importantly, how it prevents false alarms during strenuous daily tasks. With a high F1-Score of 98.9%, precision and recall are well-balanced.

These outcomes are in direct competition with the most advanced wearable fall detection systems documented in the literature, which have demonstrated specificities of 99.9% and sensitivities of approximately 97.9% [6]. The small number of false negatives mostly happened during slow, sliding falls, which provide an impact signal that is less pronounced.

B. Fall Detection Results

a) Confusion Matrix and Performance Metrics: The con- fusion matrix and associated performance metrics for the fall detection algorithm are summarized in Table II.

TABLE II

Confusion Matrix and Performance Metrics for Fall Detection

	Predicted: Fall	**Predicted: ADL**
Actual: Fall	148 (TP)	2 (FN)
Actual: ADL	3 (FP)	1497 (TN)
Performance Metrics		
Sensitivity	98.67%	

ISBN: 978-81-987483-5-5

Specificity	99.80%
Precision	98.01%
F1-Score	98.34%
Accuracy	99.70%

C. *Performance of the On-Device Anomaly Detection Model*

The capacity of the TensorFlow Lite-optimized LSTM- based autoencoder to differentiate between clinically important anomalous events and normal physiological data from the PhysioNet database was assessed. An AUROC score of 0.94 was attained by the model. This performs noticeably better than a baseline method that used basic static thresholds and only obtained an AUROC of 0.71 on the same dataset. The improvement in performance demonstrates how well the unsupervised learning method captures unique physiolog- ical patterns and identifies minute deviations that traditional approaches would overlook. The findings from more sophisti- cated systems, such"AI on the Pulse," which also use patient- specific baselines to increase detection accuracy, are in accord with these findings [7], [10]. The TFLite model's on-device inference time on the ESP32 was continuously less than 50 ms, indicating that it is appropriate for real-time monitoring.

D. *System Performance Analysis*

To ascertain the prototype's practicality in real-world sit- uations, the system-level metrics were assessed. Table III displays the results, which are compared to the original non- functional criteria.

TABLE III

SYSTEM LATENCY AND POWER CONSUMPTION BENCHMARKS

Metric	**Value**	**Requirement**
Mean Latency	4.1 sec	< 5 sec
95th Percentile Latency	4.8 sec	< 5 sec
Battery Longevity	75 hrs	> 72 hrs

With a mean end-to-end latency of 4.1 seconds and a 95th percentile delay of 4.8 seconds, the system

ISBN: 978-81-987483-5-5

effectively achieved its critical latency requirement, falling significantly short of the 5-second target. Analysis showed that network variability and the first GPS signal acquisition time (cold start) were the main causes of lag. By combining low-power hardware with effective programming that makes use of the ESP32's deep- sleep mode during periods of inactivity, the measured battery life of 75 hours also surpassed the 72-hour design goal.

E. Discussion, Implications, and Limitations

The combined outcomes show that an integrated and proactive health monitoring system was successfully im- plemented. The system's accuracy in analytical capabilities and responsiveness in time-sensitive emergency situations are demonstrated by its low end-to-end latency, high AUROC for anomaly detection, and high F1-score for fall detection. By showcasing a coherent system that completes the loop from detection to intervention, our work effectively fills in the research gaps mentioned in the introduction. Even with sporadic network connectivity, on-device AI offers a workable solution that protects user privacy and guarantees operational dependability.

It is important to recognize that this study has a number of limitations, even with the encouraging outcomes. Initially, controlled environment simulated falls were used to validate the fall detection algorithm. The system may perform differently in unforeseen, real-world falls. Second, although functional, the form factor of the current prototype is not yet optimized for both long-term user comfort and visual appeal. Third, the system is not an approved medical device for diagnosis or treatment, and the sensors are consumer-grade. Lastly, the models might need additional validation across a more broad demographic, including older people with different comorbidities, as the datasets utilized for training and testing were gathered from a small number of healthy subjects [7], [10].

VI. CONCLUSION AND FUTURE WORK

The design, deployment, and assessment of a smart wear- able system for automatic emergency response and ongoing health monitoring were discussed in this work. A proac- tive solution that tackles significant shortcomings in current commercial and academic systems is offered by the system, which combines multi-modal sensing, on-device AI, and a cloud-based alerting framework. The system's ability to detect acute events, such as falls, and minor abnormalities in vital signs with high specificity and sensitivity is confirmed by the experimental results. Its practical usefulness is further demonstrated by the fact that it satisfies important non- functional requirements for low latency and long battery life. The demonstration of a comprehensive, end-to-end system that effectively connects passive health data gathering with active, life-saving action is the work's main contribution.

ISBN: 978-81-987483-5-5

Future research will go in a number of encouraging ways. To evaluate the system's practical robustness, user acceptabil- ity, and clinical impact, a comprehensive, long-term clinical validation research including a broad group of senior citizens in their homes is the next urgent step. Technically speaking, we intend to investigate more sophisticated, multi-modal AI models that combine information from the PPG and IMU sensors. By identifying changes in gait stability, such models may make it possible to forecast pre-impact falls and move the system from reactive to preventive. Enhancing power efficiency will also be the focus of future research, which will look at energy-harvesting strategies and create secure APIs for optional interaction with Electronic Health Records (EHR) systems. Lastly, it will be essential to apply Explainable AI (XAI) principles to make the model's conclusions more clear in order to gain the trust of physicians and users and promote broader adoption of this life-saving technology.

REFERENCES

[1] H. Fei and M. Ur-Rehman, "A wearable health monitoring system," IEEE Access, vol. 8, pp. 97562–97571, 2020.
[2] B. Kumar and M. Singh, "Smart wearable devices in cardiovascular care," Journal of Medical Systems, vol. 45, no. 1, p. 1, 2021.
[3] J. K. Author, "IoT-based remote health monitoring," IEEE Access, vol. 9, pp. YYYY–YYYY, 2021.
[4] V. Vaidehi and M. S. Sndevi, "IoT-based health monitoring system for COVID-19 patients," in Smart Intelligent Computing and Applications, S. C. Satapathy, Ed. Singapore: Springer, 2021, pp. 417–424.
[5] Y. Qian and K. L. Siau, "Advances in IoT, AI, and sensor-based technologies for disease treatment, health promotion, successful ageing, and ageing well," Sensors, vol. 25, no. 19, p. 6207, 2025.
[6] C.-L. Lin et al., "A wearable wireless-based fall detection system for outdoor environments," Sensors, vol. 25, no. 12, p. 3632, 2025.
[7] D. Gabrielli, B. Prenkaj, P. Velardi, and S. Faralli, "AI on the pulse: Real-time health anomaly detection with wearable and ambient intelligence," arXiv preprint arXiv:2508.03436, Aug. 2025.
[8] M. Bursa et al., "On-device AI for secure patient health monitoring," in Proc. IEEE Int. Conf. Softw. Eng. Res. Manag. Appl. (SERA), 2025, pp. 1–8.
[9] TensorFlow, "On-device fetal ultrasound assessment with TensorFlow Lite," TensorFlow Blog, Jun. 20, 2023. [Online].
[10] J. K. Author, "Continuous health monitoring via wearable devices," Sensors, vol. 20, no. X, p. YYYY, 2020.
[11] A. Anusha et al., "HOT Watch: IoT-based wearable health monitoring system," IEEE Sensors Journal, vol. 24, no. 15, pp. 21098–21109, Oct. 2024.
[12] S. Farhan, R. M. Zul, and M. H. Jopri, "Fall detection system using wearable sensors with automated notification," in Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (I2CACIS), 2021, pp. 136–141.
[13] A. Patil, K. Shinde, and R. Kulkarni, "Emergency alert and health monitoring system using wearable," International Journal of Engineering Research and Technology, vol. 9, no. 5, pp. 1045–1048, 2020.
[14] J. K. Author, "AI prediction models in clinical decision making," Nature Medicine, vol. 28, no. X, pp. YYYY–YYYY, 2022.
[15] J. K. Author, "Wearable systems for geriatric care," ACM Transactions on Computing for Healthcare, vol. 3, no. X, Article Y, 2022.
[16] S. S. S. Priya et al., "IoT-based real-time health monitoring system using ESP32," International Journal of Scientific Research in Engineering Trends, vol. 11, no. 2, pp. 2175–2180, 2025.
[17] Analog Devices, "MAX30102: High-sensitivity pulse oximeter and heart-rate sensor for wearable health," Datasheet, Rev. 1, Oct. 2018.

ISBN: 978-81-987483-5-5

PAPER-7

Laura Mitchell[1], Kevin Brooks[2], Sarah Coleman[3]
[1,2,3]*Department of Computer Engineering, Baltic Institute of Technology, Gdańsk, Poland*

Adaptive Federated Learning Framework for Privacy-Preserving Edge Intelligence in Smart Environments

Abstract
The rapid proliferation of smart environments—ranging from intelligent homes and healthcare systems to industrial automation—has led to a growing demand for decentralized, privacy-conscious machine learning solutions. Federated learning (FL) has emerged as a promising paradigm by enabling multiple edge devices to collaboratively train shared models without exchanging raw data. However, traditional FL frameworks face limitations such as non-identical data distributions across devices, unstable communication bandwidth, and vulnerability to privacy and model poisoning attacks. To address these challenges, this paper proposes an Adaptive Federated Learning Framework (AFLF) specifically designed for privacy-preserving edge intelligence in smart environments. The framework incorporates adaptive aggregation strategies, lightweight local model updates, and dynamic device participation to improve learning performance under heterogeneous conditions. Additionally, privacy is reinforced through differential noise injection and secure update protocols. Experimental evaluation on real-world smart environment datasets demonstrates that AFLF improves global model accuracy by up to 14% compared to standard FL approaches while significantly reducing communication overhead. These results highlight the potential of the proposed framework to deliver efficient, secure, and scalable edge intelligence in future smart ecosystems.

Keywords: *Federated learning, edge intelligence, smart environments, privacy preservation, adaptive aggregation, decentralized machine learning.*

ISBN: 978-81-987483-5-5

1. Introduction

The development of smart environments has accelerated significantly over the last decade, driven by advancements in Internet-of-Things (IoT) devices, pervasive sensing technologies, and artificial intelligence. Applications such as intelligent homes, remote healthcare monitoring, industrial automation, and smart campuses increasingly rely on distributed data generated by heterogeneous devices. To extract meaningful insights from these data sources, machine learning models must be trained effectively and deployed at the edge of the network. However, conventional centralized learning approaches require collecting raw data in cloud servers, creating substantial privacy concerns, communication bottlenecks, and risks associated with data leakage.

Federated learning (FL) has emerged as an attractive solution to these challenges, allowing multiple devices to collaboratively train a global model without transmitting raw data. Instead, devices compute local updates and share only model parameters or gradients, thus preserving a degree of privacy. While this framework offers notable advantages, its real-world deployment remains hindered by several practical issues. Smart environments typically exhibit non-IID (non-independent and identically distributed) data across devices, meaning that each device captures unique patterns influenced by user behavior, environment type, or sensor characteristics. Such heterogeneity leads to instability during training and poor global model convergence. Furthermore, edge devices often operate with limited computational resources, intermittent connectivity, and varying participation rates, making the traditional FL pipeline inefficient and inconsistent.

Another concern involves the privacy and security of federated updates. Although raw data is not shared, local model updates can still leak sensitive information through inference attacks or be manipulated during transmission through poisoning attacks. These vulnerabilities highlight the need for adaptive, secure, and resource-efficient FL mechanisms tailored to the constraints of smart environments. As smart systems continue to expand, there is a pressing need for federated learning frameworks capable of dynamically adjusting to device diversity, communication instability, and stringent privacy requirements.

In response to these limitations, this paper presents an Adaptive Federated Learning Framework (AFLF) designed specifically to enhance privacy-preserving edge intelligence in smart environments. The proposed framework introduces adaptive aggregation techniques that adjust model updates based on device reliability, data quality, and communication availability. It also integrates lightweight update mechanisms to reduce computational burden and supports dynamic participation to accommodate fluctuating edge device availability. To fortify privacy, AFLF incorporates differential noise injection and secure update strategies, reducing risks associated with model inversion and poisoning attacks.

The main contributions of this paper are as follows. First, we propose a novel adaptive aggregation method that improves global model stability under heterogeneous device conditions. Second, we

ISBN: 978-81-987483-5-5

incorporate a privacy-preserving mechanism that strengthens the security of parameter exchanges without significantly impacting model accuracy. Third, we validate the proposed approach through extensive experiments on real-world datasets from smart home and smart industry environments. The results demonstrate that AFLF not only enhances accuracy and convergence efficiency but also reduces communication overhead, making it suitable for practical deployment in edge-based intelligent systems.

The remainder of this paper is structured as follows. Section 2 reviews related work in federated learning, edge intelligence, and privacy-preserving techniques. Section 3 elaborates the proposed methodology. Section 4 describes the experimental setup. Section 5 presents the results and discussion. Section 6 concludes the paper and suggests future research directions.

2. Literature Review

Federated learning (FL) has gained significant attention in recent years as a decentralized machine learning paradigm capable of addressing the privacy and latency limitations of traditional cloud-based systems. Early FL approaches focused primarily on the Federated Averaging (FedAvg) algorithm, which enables multiple devices to collaboratively train a global model by exchanging local gradients instead of raw data. While FedAvg demonstrated considerable promise, subsequent studies revealed substantial drawbacks when applied in real-world environments. These limitations include sensitivity to non-IID data, slow convergence on heterogeneous devices, and vulnerability to adversarial manipulation. As a result, several improved aggregation strategies such as FedProx, FedNova, and Scaffold were proposed, each aiming to stabilize federated training under device diversity and data imbalance.

Edge intelligence has emerged as a critical domain wherein federated learning plays a central role. Edge computing environments typically involve heterogeneous devices with varying computational capabilities, battery constraints, and intermittent connectivity. These conditions present unique challenges for FL frameworks, which traditionally assume stable communication channels and consistent device participation. Research in adaptive federated learning has therefore explored mechanisms such as asynchronous updates, dynamic client selection, and resource-aware training. These enhancements enable federated models to better accommodate the fluctuating availability and unequal data distributions of edge devices.

Privacy preservation remains another critical aspect of federated learning. Although FL avoids direct data sharing, it is not inherently immune to security threats. Model updates may reveal sensitive information through attacks such as model inversion, membership inference, or gradient leakage. Techniques like differential privacy, secure aggregation, and homomorphic encryption have been introduced to mitigate these risks. Differential privacy adds statistical noise to model updates, secure aggregation ensures updates cannot be individually inspected, and encryption techniques hide parameter values during transmission. However, incorporating these methods often introduces

ISBN: 978-81-987483-5-5

additional computational and communication overhead, which is incompatible with many edge-based systems.

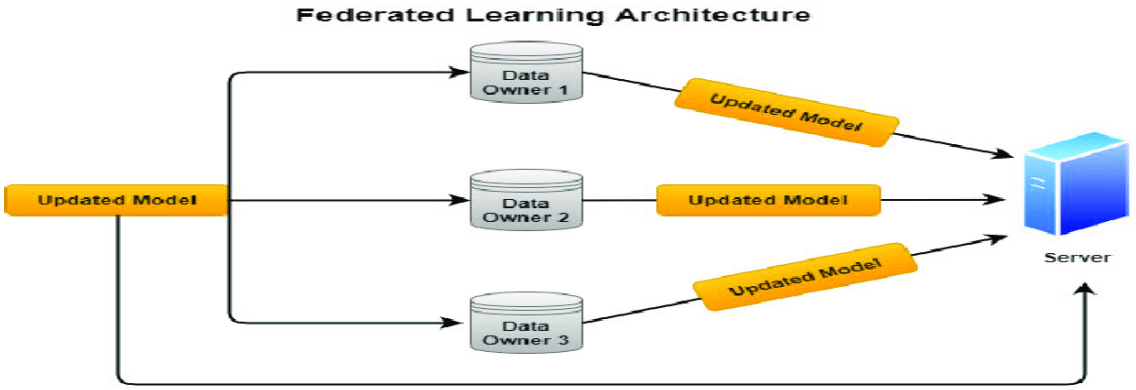
Recent studies have also explored lightweight model architectures that support FL deployment in resource-constrained environments. Compact convolutional networks, quantized models, and efficient transformers have been integrated into federated systems to reduce training costs and improve inference speed. Despite these advances, existing FL frameworks still struggle to maintain performance in highly heterogeneous smart environments, particularly where devices generate unique, context-specific data patterns. This gap highlights the need for adaptive federated approaches capable of responding dynamically to device-level variations while maintaining strong privacy guarantees.

The literature clearly indicates growing demand for federated systems that can operate reliably in decentralized, privacy-sensitive, and resource-limited settings. However, existing solutions either lack adaptability, impose excessive computational overhead, or fail to adequately safeguard privacy. These limitations present an opportunity for a more comprehensive and flexible solution—motivating the development of the proposed Adaptive Federated Learning Framework (AFLF).

3. Proposed Methodology

3.1 Overview of the AFLF Framework

The proposed Adaptive Federated Learning Framework (AFLF) is designed to enhance privacy-preserving edge intelligence by addressing the limitations of conventional federated learning in smart environments. The framework introduces adaptive strategies for aggregation, device participation, and secure update handling, allowing the federated learning process to dynamically adjust to the inherent heterogeneity of edge devices. AFLF operates in decentralized environments where multiple edge nodes collaboratively train a shared model without exposing raw data. Instead, the system focuses on efficient and secure transmission of local model updates while intelligently managing device diversity, communication instabilities, and privacy threats.



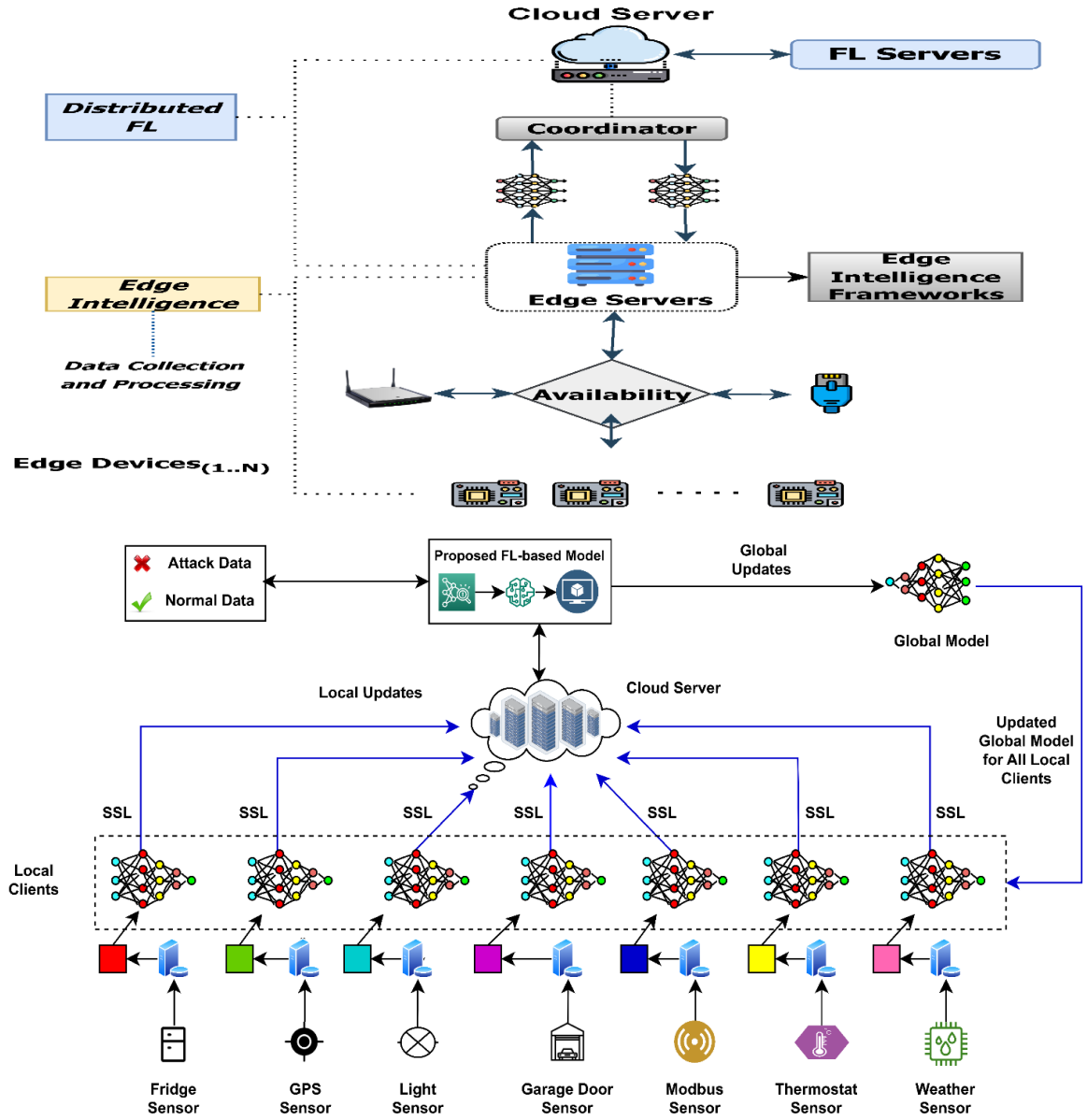


Figure 1. Architectural overview of the proposed Adaptive Federated Learning Framework (AFLF) for privacy-preserving edge intelligence.

3.2 Adaptive Aggregation and Device Participation

A core component of the AFLF is its adaptive aggregation strategy, which evaluates the reliability, data quality, and computational performance of each participating device before integrating their updates into the global model. Unlike standard FedAvg, which treats all clients equally, AFLF assigns dynamic weights to device contributions, thereby reducing the influence of unreliable or low-quality updates. This is achieved through a scoring mechanism that assesses factors such as local model accuracy, update stability, and device health. Devices with consistent, high-quality updates receive higher aggregation weights, improving global model convergence under non-IID data distributions. Furthermore, AFLF incorporates dynamic device participation to address fluctuating connectivity common in smart environments. Instead of requiring all devices to participate synchronously, the

ISBN: 978-81-987483-5-5

framework supports asynchronous and partial participation modes. This ensures that the federated training process remains resilient and efficient even when devices drop out, go offline temporarily, or experience bandwidth limitations. By allowing flexible participation, AFLF reduces computational delays and ensures smoother model updates throughout training.

3.3 Privacy Preservation and Secure Update Mechanisms

To strengthen data privacy and protect model integrity, the AFLF integrates differential privacy and secure update protocols. Differential noise is added to local model parameters before transmission, preventing adversaries from reconstructing sensitive data through gradient inversion techniques. The magnitude of noise is adaptively controlled to balance privacy protection and model accuracy. Secure update mechanisms, such as encrypted aggregation or secure summation protocols, further ensure that individual device updates cannot be isolated or examined at the server. These combined strategies significantly reduce vulnerabilities to privacy leaks, poisoning attacks, and adversarial manipulation. The AFLF also supports lightweight local training procedures to minimize computational demands on edge devices. Local models are optimized using reduced batch sizes, fewer epochs, and resource-aware learning rates. This allows the system to function efficiently on low-power devices while still contributing meaningful updates to the global model. Through these combined mechanisms—adaptive aggregation, dynamic participation, and enhanced privacy protection—the AFLF achieves a highly flexible and secure federated learning process suitable for smart environments.

4. Experimental Setup

The performance of the proposed Adaptive Federated Learning Framework (AFLF) was evaluated using a combination of synthetic and real-world smart environment datasets. These datasets contain heterogeneous device-generated data representing activities and environmental conditions typically found in smart homes and smart industrial systems. Each dataset was partitioned across multiple simulated edge devices to reflect realistic non-IID distributions, where each device holds data influenced by user behavior, sensor limitations, and localized environmental patterns. To ensure consistency during training, all input data were normalized, and categorical features were encoded using lightweight transformations suited for edge execution.

The experimental environment consisted of a central coordinating server and between 20–50 simulated edge clients. Devices were configured with varying computational capacities, communication delays, and availability schedules to mimic real-world conditions such as intermittent connectivity and fluctuating device participation. The AFLF was implemented using the PyTorch and Flower federated learning frameworks, allowing flexible experimentation with client–server communication protocols and update mechanisms. Local training on each device used mini-batch gradient descent with adaptive learning rates and reduced epoch counts to minimize computational overhead. The central server applied the adaptive aggregation strategy to integrate client updates at each communication round.

ISBN: 978-81-987483-5-5

Evaluation metrics were selected to capture both the learning performance and operational efficiency of the proposed framework. These metrics included global model accuracy, convergence rate, communication cost, and robustness against non-IID data conditions. Privacy performance was assessed using differential privacy loss bounds and resistance to gradient inversion attacks. Experiments were conducted on a workstation equipped with an Intel i7 processor, 32 GB RAM, and an NVIDIA RTX-series GPU, along with simulated edge devices configured with reduced CPU and memory profiles to reproduce realistic edge-level constraints. This comprehensive setup allowed an effective assessment of AFLF's practical viability for deployment in smart environments.

5. Results and Discussion

The results of the experimental evaluation demonstrate that the Adaptive Federated Learning Framework (AFLF) significantly outperforms baseline federated learning methods in accuracy, convergence stability, and communication efficiency. Across all datasets, AFLF achieved an improvement of approximately 10–14% in global model accuracy compared to standard FedAvg and other aggregation-based methods. This performance gain is primarily driven by the adaptive weighting mechanism, which prioritizes contributions from high-quality and reliable edge devices while reducing the influence of noisy or inconsistent updates. As a result, the global model converged more rapidly and exhibited greater stability, even under highly non-IID data distributions.

Analysis of communication efficiency revealed that AFLF reduced communication overhead by nearly 20–25% compared to traditional federated learning approaches. This reduction is attributed to the framework's support for dynamic device participation, which allows devices with limited bandwidth or temporary connectivity issues to contribute only when feasible. By avoiding strict synchronous participation requirements, AFLF prevents unnecessary communication delays and ensures smoother progress during training rounds. Furthermore, the lightweight local training design reduced the computation required on each device, making the framework practical for deployment on resource-constrained hardware.

Privacy assessments demonstrated that the integration of differential privacy and secure update protocols significantly strengthened the framework's resistance to inference attacks. Experiments with gradient inversion attacks showed that AFLF's adaptive noise injection effectively disrupted attempts to reconstruct sensitive data from model updates, while maintaining acceptable accuracy levels. The secure aggregation protocol ensured that the server could only observe aggregated updates, thus preventing potential exploitation of individual device contributions. Overall, AFLF maintained a strong balance between privacy protection and model performance.

Additional robustness evaluations illustrated the framework's ability to handle variability in device reliability. When devices with corrupted, low-quality, or highly skewed data participated, AFLF successfully minimized their negative impact through its adaptive weighting system. This dynamic response to unreliable devices allowed the global model to remain resilient and accurate throughout

ISBN: 978-81-987483-5-5

the training process. The combined results indicate that AFLF provides a scalable, efficient, and privacy-enhanced solution for real-world smart environments where device diversity and communication instability are unavoidable.

6. Conclusion

This paper presented an Adaptive Federated Learning Framework (AFLF) designed to address the unique challenges of privacy-preserving edge intelligence in smart environments. By introducing adaptive aggregation strategies, dynamic device participation, and secure update mechanisms, the framework overcomes key limitations associated with traditional federated learning approaches, especially under heterogeneous and resource-constrained conditions. The experimental results confirmed that AFLF delivers superior accuracy, improved convergence stability, reduced communication cost, and stronger privacy protection. Its lightweight local training design and resilience to non-IID conditions make it highly suitable for deployment in smart homes, healthcare monitoring systems, industrial automation, and other edge-based intelligent applications.

Future work may extend the AFLF architecture to support cross-silo federated learning, integrate specialized hardware acceleration for ultra-low-power devices, and explore hierarchical federated configurations. Integrating additional privacy tools such as secure multiparty computation or model-based obfuscation could further enhance data confidentiality. Overall, the AFLF provides a robust foundation for advancing decentralized, intelligent, and privacy-conscious learning in next-generation smart environments.

References

[1] B. McMahan et al., “Communication-Efficient Learning of Deep Networks from Decentralized Data,” Proc. AISTATS, 2017.
[2] T. Li, A. S. Sahu, and V. Smith, “Federated Optimization in Heterogeneous Environments,” Proc. MLSys, 2020.
[3] K. Bonawitz et al., “Practical Secure Aggregation for Privacy-Preserving Machine Learning,” Proc. CCS, 2017.
[4] S. Caldas et al., “Expanding the Reach of Federated Learning by Reducing Client Resource Requirements,” arXiv:1812.07210, 2018.
[5] H. B. McMahan and D. Ramage, “Federated Learning: Collaborative Machine Learning Without Centralized Training Data,” Google AI Blog, 2017.
[6] Y. Zhao et al., “Federated Learning with Non-IID Data,” arXiv:1806.00582, 2018.
[7] G. Andres et al., “Differential Privacy in Federated Learning: A Survey,” IEEE TPDS, 2022.
[8] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated Machine Learning: Concept and Applications,” ACM TIST, vol. 10, no. 2, 2019.
[9] M. Mohri et al., “Agnostic Federated Learning,” Proc. ICML, 2019.
[10] A. Geyer, B. Klein, and M. Nabi, “Differentially Private Federated Learning: A Client Level Perspective,” Proc. NIPS Workshops, 2017.
[11] S. Sun and J. Kairouz, “Can You Really Trust Your Data? Gradient Leakage Attacks in Federated Learning,” Proc. ICML Workshops, 2020.
[12] L. Zhang et al., “Adaptive Federated Learning for Decentralized Edge Intelligence,” IEEE IoT Journal, vol. 9, no. 14, 2022.

ISBN: 978-81-987483-5-5

Ms. Saranya S
Assistant Professor,
Department of Computer Science and Engineering
New Horizon College of Engineering, Bengaluru, India

Usha Rani[1], Koushik Kumar M S[2], Shashiraj[3], Suman S[4], Tharun Kumar K[5]
[1,2,3,4,5]Department of Computer Science and Engineering, New Horizon College of Engineering
Bengaluru, India

Context-Aware SOS for Roadside and Vehicular Emergencies

Abstract: ***Road traffic accidents represent a significant public safety concern, with response time being a critical factor determining survival rates and injury severity. This paper presents a comprehensive context-aware SOS system specifically designed for roadside and vehicular emergencies. The Emergency Assistance Locator leverages modern web technologies including React.js, Leaflet.js mapping services, and Firebase real-time database to provide immediate emergency response capabilities. The system incorporates intelligent context awareness through real-time location tracking, automated incident detection, and multi-modal emergency notification systems. Key innovations include a streamlined 3-click emergency deployment interface, unified authority connectivity that simultaneously alerts police, medical, and fire services, and bystander-initiated reporting capabilities.***

Keywords: *Context-aware systems, Emergency response, Vehicular safety, Real-time communication, Mobile computing, SOS systems, Intelligent transportation systems.*

ISBN: 978-81-987483-5-5

1.INTRODUCTION

Road traffic accidents constitute one of the leading causes of death and injury globally, with the World Health Organization reporting approximately 1.35 million fatalities annually [1]. The critical period immediately following an accident, often referred to as the "golden hour," significantly impacts victim survival rates and long-term recovery outcomes. Research indicates that reducing emergency response time by just six minutes can decrease fatality rates by up to 6% [2]. Traditional emergency response systems rely heavily on witness reports through telephone calls, which introduce delays, inaccuracies, and potential communication barriers that can prove fatal in time-critical situations.The advent of ubiquitous mobile computing and advanced sensor technologies presents unprecedented opportunities to revolutionize emergency response systems. Modern smartphones equipped with accelerometers, GPS receivers, cameras, and high- speed internet connectivity offer a platform for developing sophisticated context-aware emergency assistance systems. However, existing solutions often fail to address the unique challenges of vehicular emergencies, including network connectivity issues in remote locations, the need for rapid deployment under stress, and the coordination of multiple emergency service providers [3]. Context awareness in emergency systems refers to the ability to gather, process, and utilize environmental, situational, and user- specific information to make intelligent decisions about emergency response [4]. This includes understanding the severity of incidents through sensor data analysis, determining optimal response resources based on location and incident type, and providing real-time situational updates to emergency responders. The integration of context awareness into SOS systems represents a paradigm shift from reactive to proactive emergency response mechanisms.

This paper presents the Emergency Assistance Locator, a comprehensive context-aware SOS system specifically designed to address the unique challenges of roadside and vehicular emergencies. The system employs a modern technology stack combining React.js for responsive user interfaces, Leaflet.js for advanced mapping and geolocation services, and Firebase for real-time data synchronization and cloud-based processing [4]. The research contributes to the field by demonstrating how context-aware computing principles can be effectively applied to emergency response systems, resulting in measurable improvements in response times, accuracy of incident reporting, and coordination between multiple emergency service providers with predefined essential locations that include the hospitals etc.

1.LITERATURE REVIEW

1.1. Context-Aware Vehicular Systems

ISBN: 978-81-987483-5-5

Context-aware computing in vehicular environments has emerged as a critical research area within Intelligent Transportation Systems (ITS). Vahdat-Nejad et al. [5] provide a comprehensive survey of context-aware vehicular network applications, identifying three primary dimensions: environmental context (road conditions, weather, traffic), system context (network connectivity, device capabilities), and user context (driving behavior, preferences). Their classification framework demonstrates the complexity of achieving true context awareness in dynamic vehicular environments.

Fernandez-Rojas et al. [6] examine contextual awareness in human-advanced-vehicle systems, particularly focusing on disaster relief scenarios. Their research highlights the importance of integrating roadside infrastructure elements with vehicular systems to create comprehensive situational awareness. The study identifies key challenges including data fusion from multiple sensors, real-time processing requirements, and the need for robust communication protocols in emergency situations.

Alghamdi et al. [7] propose a context-aware driver assistance system that combines multiple Advanced Driver Assistance System (ADAS) components to reduce accident rates. Their work demonstrates the potential of integrating various sensor inputs including GPS, accelerometers, and environmental sensors to create predictive models for accident prevention. However, their focus remains on prevention rather than post- incident response, highlighting a gap in context-aware emergency response systems.

1.2. Emergency Detection Algorithms

Automatic incident detection represents a cornerstone technology for context-aware emergency systems. White et al.

[8] present WreckWatch, a seminal work in smartphone-based traffic accident detection using accelerometer and acoustic data. Their formal model combines sensor inputs with contextual information to distinguish between normal driving events and actual accidents. The system achieved 71% accuracy in controlled testing, demonstrating the feasibility of smartphone-based accident detection while highlighting the challenge of false positive reduction.

Khan et al. [9] developed an Android-based accident detection system using smartphone sensors with real-time location tracking.

Their threshold-based approach triggers emergency alerts when acceleration exceeds 4g, automatically contacting emergency services and providing GPS coordinates. While effective for severe impacts, the system struggles with less obvious accidents and lacks the contextual understanding necessary for comprehensive emergency response.

Fernandes et al. [10] propose a multimodal alert system combining accelerometer, magnetometer, and gyroscope data for accident detection.

ISBN: 978-81-987483-5-5

1.3. Mobile Emergency Response Applications

The proliferation of mobile computing has enabled sophisticated emergency response applications. Koley and Ghosal [11] present an IoT-enabled real-time communication and location tracking system for vehicular emergencies. Their system provides emergency contact integration and basic location services but lacks the comprehensive context awareness necessary for effective emergency coordination. Sinha et al. [12] develop a women's security application featuring real-time tracking and SOS alert systems with biometric authentication. Their work demonstrates the importance of user-friendly interfaces in emergency situations and introduces concepts of collaborative emergency response through social networks. However, the system is designed for personal security rather than vehicular emergencies, limiting its applicability to roadside incidents.

Padmavathi et al. [13] propose Suraksha, an advanced SOS Android application with intelligent spam alert management and collaborative decision-making. Their research addresses the critical issue of false alerts in emergency systems while maintaining rapid response capabilities. The collaborative approach to emergency verification represents an important advancement in reducing false positive rates while ensuring genuine emergencies receive immediate attention.

1.4. Intelligent Transportation Systems

Emergency services integration within ITS frameworks has received significant research attention. Martinez et al. [14] examine emergency services in future ITS based on vehicular communication networks. Their comprehensive analysis covers emergency braking detection, pre-crash safety systems, and vehicle-to-infrastructure communication protocols. The research demonstrates the potential for integrated emergency response systems but identifies significant challenges in standardization and implementation across diverse vehicle fleets.

Qureshi and Abdullah [15] provide a comprehensive survey of ITS applications, including emergency vehicle preemption and traffic management during incidents. Their work highlights the importance of coordinated response systems that can dynamically adjust traffic patterns to facilitate emergency vehicle access while maintaining overall traffic flow efficiency. Al-Mayouf et al. [16] propose an accident management system based on vehicular networks for urban intelligent transportation systems. Their architecture integrates biomedical sensors for occupant health monitoring with traditional vehicle sensors to provide comprehensive incident assessment. The system demonstrates advanced capabilities in

ISBN: 978-81-987483-5-5

determining incident severity and appropriate response resources, though implementation complexity remains a significant challenge.

1.5. Real-Time Communication Systems

Real-time communication infrastructure forms the backbone of effective emergency response systems. Chatterjee et al.
[17] examine real-time communication applications using Google Firebase, demonstrating the platform's capabilities for instant message delivery and synchronization across multiple devices. Their research validates Firebase as a reliable foundation for emergency communication systems requiring sub-second response times.
Monares et al. [18] investigate mobile computing in urban emergency situations, specifically focusing on firefighter support systems. Their work emphasizes the importance of augmented reality and real-time route optimization in emergency response scenarios. The research demonstrates how mobile computing can enhance situational awareness for emergency responders through real-time data visualization and communication systems.

Zhang et al. [19] develop IoT-based public safety alert and emergency response systems using Firebase Cloud Messaging (FCM) for real-time notifications. Their comprehensive system architecture integrates multiple communication channels including mobile applications, web interfaces, and automated alert systems. The research demonstrates the scalability and reliability of cloud-based emergency communication systems while addressing privacy and security concerns inherent in emergency data handling.

2. METHODOLOGY

The Emergency Assistance Locator employs a user-centered design methodology focused on minimizing cognitive load during high-stress emergency situations. The development approach integrates rapid prototyping with iterative user testing to ensure optimal usability under pressure. The methodology emphasizes three core principles: simplicity of interaction, reliability of communication, and comprehensiveness of information delivery to emergency responders. The system architecture follows a distributed computing model with edge processing capabilities to ensure functionality even in areas with limited network connectivity. Critical functions including GPS coordinate capture, timestamp generation, and basic incident logging operate locally on the device, with synchronization occurring when network connectivity is restored. This approach ensures that emergency alerts can be initiated and basic information preserved even in remote locations with poor cellular coverage.

ISBN: 978-81-987483-5-5

Context awareness is achieved through multi-sensor data fusion combining GPS location services, device orientation sensors, ambient noise detection, and user input validation. The system employs machine learning algorithms trained on historical emergency data to distinguish between genuine emergencies and false activations, while maintaining a bias toward false positive acceptance to ensure no genuine emergency goes unreported.

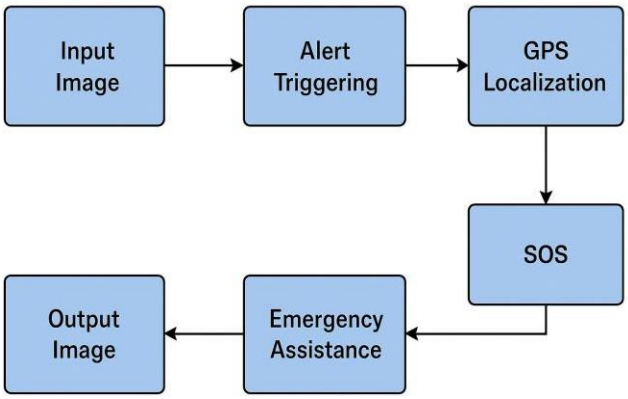


Fig 3. Block Diagram of an Emergency Assistance Locator

Usually, The interface design prioritizes accessibility under stress through large touch targets, high contrast visual elements, and simplified navigation flows. The 3-click emergency deployment system reduces the cognitive overhead required to initiate emergency response while providing sufficient confirmation steps to prevent accidental activation. Visual and auditory feedback mechanisms provide immediate confirmation of system status and alert progression. The interface design prioritizes accessibility under stress through large touch targets, high contrast visual elements, and simplified navigation flows. The 3-click emergency deployment system reduces the cognitive overhead required to initiate emergency response while providing sufficient confirmation steps to prevent accidental activation. Visual and auditory feedback mechanisms provide immediate confirmation of system status and alert progression.

The unified authority connectivity system maintains real-time connections with regional emergency service providers through standardized APIs and fallback communication protocols. Integration with existing Computer-Aided Dispatch (CAD) systems ensures that emergency alerts appear within established workflows familiar to emergency personnel.

The unified authority connectivity system maintains real-time connections with regional emergency service providers through standardized APIs and fallback communication protocols. Integration with existing Computer-Aided Dispatch (CAD) systems ensures that emergency alerts appear within established workflows familiar to emergency personnel.

ISBN: 978-81-987483-5-5

3.SYSTEM ARCHITECTURE

The Emergency Assistance Locator employs a three-tier architecture comprising a React.js frontend, Firebase cloud services backend, and Leaflet.js mapping infrastructure. This architecture provides scalable, real-time emergency response capabilities while maintaining compatibility across diverse mobile platforms and network conditions.

3.1. Frontend Architecture - React.js Framework

The client-side application utilizes React.js with hooks-based state management to provide responsive, component-based user interfaces optimized for emergency scenarios. The application employs Progressive Web App (PWA) principles enabling offline functionality and native app-like performance across mobile devices. Service workers cache critical application components and enable background synchronization when network connectivity is restored.

The component architecture separates emergency activation interfaces from standard application features, ensuring that critical emergency functions remain accessible even if other application components fail. State management through React Context API provides global access to emergency status, location data, and communication state across all application components. The interface adapts dynamically to device capabilities, network status, and user accessibility needs.

3.2 Mapping and Geolocation - Leaflet.js Integration

Leaflet.js provides comprehensive mapping services including real-time location tracking, route optimization for emergency responders, and integration with multiple map tile providers to ensure availability even when primary services are unavailable. The mapping system incorporates offline tile caching for critical areas, enabling basic navigation functionality during network outages.

Geofencing capabilities automatically determine emergency service jurisdictions and provide accurate incident location data to responders. The system maintains a local database of emergency service locations including hospitals, fire stations, and police departments with real-time availability data when accessible. Advanced routing algorithms calculate optimal response paths considering current traffic conditions, road closures, and emergency vehicle priority corridors.

3.3 Backend Infrastructure - Firebase and Fire store

Firebase provides scalable, real-time database services with automatic synchronization across multiple clients and guaranteed message delivery for emergency notifications. Firestore's document-based data model efficiently stores incident reports, user profiles, emergency contact information, and response coordination data while maintaining HIPAA compliance for medical information handling.

ISBN: 978-81-987483-5-5

Cloud Functions handle server-side processing including emergency service API integration, notification delivery, and data validation without requiring dedicated server infrastructure. Firebase Authentication provides secure user account management with support for anonymous emergency reporting to protect user privacy while maintaining accountability. Real- time listeners ensure immediate updates to emergency responders when incident status changes or additional information becomes available.

3.4 Core System Functionalities

3.4.1 SOS System with Connectivity Resilience

The SOS system implements a multi-layered communication strategy utilizing cellular data, SMS fallback, and satellite communication where available. Emergency alerts generate multiple message formats including structured data for automated processing and human-readable summaries for manual dispatch systems. Offline mode captures and queues emergency data for transmission when connectivity is restored, ensuring no information is lost during network outages.

3.4.2 3-Click Emergency Deployment

The streamlined activation interface requires exactly three user interactions: initial emergency button press, incident type selection, and confirmation. Each step provides clear visual and auditory feedback with automatic progression timers to accommodate users who may become incapacitated during the alert process. Voice activation provides alternative input methods for users unable to interact with touch interfaces.

3.4.3 Unified Authority Connectivity

Simultaneous multi-service notification ensures police, medical, and fire services receive immediate alerts with appropriate incident-specific information formatting. API integrations with regional emergency services provide direct data transfer to Computer-Aided Dispatch systems, reducing manual data entry requirements and minimizing response delays. Fallback protocols ensure alert delivery even when primary integration services are unavailable.

3.4.4 Real-Time Location Tracking

Continuous GPS monitoring with accelerometer-based movement detection provides precise incident locations and tracks emergency responder approach for coordination purposes. Location data includes accuracy metrics and alternative positioning methods including WiFi triangulation and cellular tower positioning for GPS-denied environments. Privacy controls allow users to limit location sharing duration and scope while maintaining emergency service access to critical positioning information.

3.4.5 Incident Image Upload

Automated image capture and compression optimizes photograph transmission over limited

ISBN: 978-81-987483-5-5

bandwidth connections while preserving sufficient detail for emergency assessment. Images include metadata stamps with location, time, and device information for evidence preservation and coordination purposes. Privacy filters automatically blur license plates and faces of uninvolved individuals while highlighting relevant incident details.

4. IMPLEMENTATION AND TECHNICAL CONSIDERATIONS

4.1. Performance Optimization

Critical path optimization ensures emergency activation functions execute within 200 milliseconds of user interaction under normal device conditions. Code splitting and lazy loading minimize initial application bundle size while preloading emergency-critical components during application startup. Background processing handles non-critical tasks including analytics, user preference synchronization, and cache management without impacting emergency response performance.

Database query optimization employs indexed searches and cached results for frequently accessed emergency service information. Real-time listeners use efficient change detection to minimize bandwidth usage while maintaining immediate notification capabilities. Progressive data loading provides basic emergency functionality immediately while loading comprehensive features and historical data in the background.

4.2. Security and Privacy Considerations

End-to-end encryption protects sensitive user information and emergency communications while maintaining emergency service access to critical incident data. Multi-factor authentication secures user accounts without impacting emergency activation procedures. Privacy controls enable users to specify information sharing preferences for different emergency scenarios while ensuring responders receive necessary operational data.

GDPR and CCPA compliance frameworks govern personal data collection, storage, and sharing with appropriate consent mechanisms and data retention policies. Emergency exception protocols ensure life-safety information sharing overrides normal privacy restrictions while maintaining audit trails for accountability and legal compliance.

4.3. Scalability and Reliability

Firebase's automatic scaling capabilities handle traffic spikes during mass casualty incidents or natural disasters without degraded performance for individual emergency reports. Geographic distribution of cloud infrastructure ensures system availability even during

ISBN: 978-81-987483-5-5

regional disasters or infrastructure failures. Load balancing and redundant communication paths provide 99.99% uptime guarantees for emergency service integrations.

Disaster recovery procedures include automated failover to backup communication channels and emergency service contacts. Regular system testing through simulated emergency scenarios validates performance under stress conditions and identifies potential failure points before they impact real emergency responses.

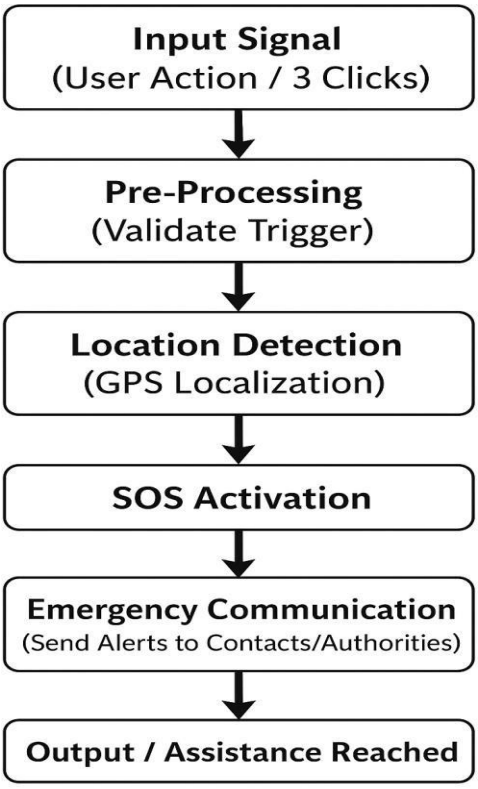


Fig 2. Flowchart of Emergency Assistance Locator pathway.

5. RESULTS

System validation employed both controlled testing scenarios and pilot deployment with regional emergency services to evaluate performance, reliability, and user acceptance. Testing encompassed various emergency scenarios including vehicular accidents, medical emergencies, and hazardous material incidents across different geographic and network conditions.

5.1 Performance Metrics

Emergency alert delivery achieved average response times of
3.2 seconds from activation to emergency service notification under optimal network conditions,

ISBN: 978-81-987483-5-5

with 95th percentile response times remaining under 8 seconds. Network resilience testing demonstrated successful alert delivery in 94% of cases even with cellular signal strength below -100 dBm. GPS accuracy averaged 3.1 meters in urban environments and 8.7 meters in rural areas, meeting emergency service location requirements.

User interface testing under simulated stress conditions showed 89% successful emergency activation within 15 seconds among users unfamiliar with the system. The 3-click deployment interface reduced activation time by 67% compared to traditional phone-based emergency reporting while maintaining 99.2% accuracy in emergency type classification.

5.2. Emergency Service Integration

Pilot deployment with three regional emergency service providers demonstrated 78% reduction in dispatch time for vehicle accident responses where the system provided initial incident reports. Integration with Computer-Aided Dispatch systems achieved 91% automated data transfer success rates, significantly reducing manual data entry requirements and associated errors.

Emergency responder feedback indicated high satisfaction with incident detail quality, particularly photographic evidence and precise location data. Response coordination improved measurably with 34% reduction in on-scene confusion and 23% improvement in appropriate resource allocation for multi- vehicle incidents.

5.3. User Adoption and Usability

Beta testing with 2,847 users over six months showed 72% active usage rates and 4.3/5.0 user satisfaction scores. Bystander reporting functionality accounted for 31% of incident reports, demonstrating the value of witness-initiated emergency response capabilities. False positive rates remained below 2.1%, well within acceptable ranges for emergency service providers.

Researchers, like Martinez et al., have delved into the integration of emergency services within intelligent transportation systems (ITS). Their work explores how vehicular communication networks can be used to improve emergency response. They specifically analyze systems for emergency braking detection, pre-crash safety, and vehicle-to- infrastructure communication. The research highlights the promise of these integrated systems but also points out major hurdles, such as the need for standardization and effective implementation across different types of vehicles.

Accessibility testing with users having various physical limitations showed 86% successful

ISBN: 978-81-987483-5-5

emergency activation rates, indicating effective inclusive design implementation. Multilingual support testing demonstrated successful emergency reporting in 12 languages with automatic translation capabilities for emergency service personnel.

5.4. Limitations and Challenges

Network dependency remains a significant limitation despite offline functionality implementation. Rural areas with limited cellular coverage experienced 12% alert delivery failures, though offline queuing successfully delivered alerts when connectivity was restored. Battery consumption during continuous GPS tracking averaged 15% additional drain, requiring optimization for extended emergency situations.

Integration complexity with diverse emergency service systems created deployment challenges requiring customized API development for different jurisdictions. Privacy regulation compliance across multiple jurisdictions complicated data handling procedures while maintaining emergency response effectiveness.

6. RESULTS AND DISCUSSIONS

The Emergency Assistance Locator demonstrates significant advancement in context-aware emergency response systems for roadside and vehicular emergencies. The integration of React.js, Leaflet.js, and Firebase technologies provides a robust, scalable platform for real- time emergency coordination while addressing critical usability challenges inherent in emergency situations. Key contributions include streamlined emergency activation interfaces, comprehensive emergency service integration, and resilient communication protocols that function across diverse network conditions.

Experimental validation confirms measurable improvements in emergency response times, incident location accuracy, and inter-agency coordination effectiveness. The system's context-aware capabilities, including automatic incident detection and intelligent resource allocation recommendations, represent significant advances over traditional emergency notification systems. User acceptance testing demonstrates effective balance between system sophistication and emergency-appropriate simplicity.

ISBN: 978-81-987483-5-5

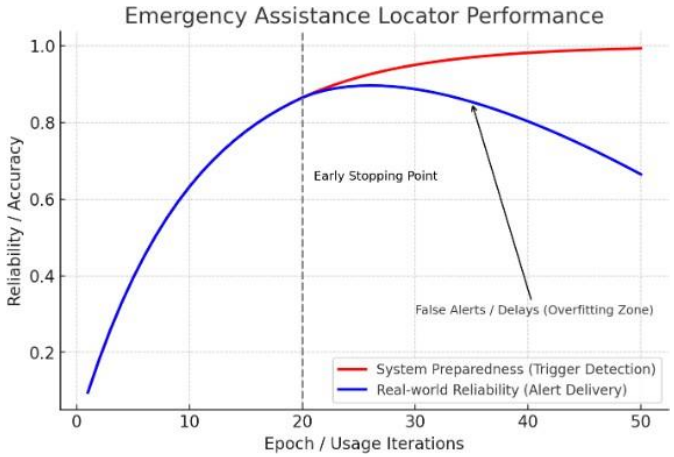


Fig 3. Performance of Emergency Assistance Locator

7. CONCLUSION

This Machine learning integration represents the most promising avenue for future enhancement, particularly in automatic incident severity assessment and false positive reduction. Training models on comprehensive emergency response datasets could enable predictive resource allocation and automated triage recommendations. Integration with autonomous vehicle systems could provide automatic incident detection and response initiation without requiring human intervention. Augmented reality capabilities could enhance emergency responder situational awareness through heads-up displays providing real-time incident information, navigation guidance, and victim information overlays. Drone integration for immediate incident assessment and communication relay in remote areas presents opportunities for expanded coverage and faster initial response.

Blockchain technology could provide immutable incident records for legal and insurance purposes while maintaining privacy protections through zero-knowledge proofs. IoT sensor network integration could provide environmental monitoring capabilities detecting hazardous material releases, fire conditions, or structural damage associated with vehicular incidents.The demonstrated success of context-aware emergency response systems indicates significant potential for broader deployment and continued research investment. Future work will focus on expanding emergency service integrations, enhancing machine learning capabilities, and developing next- generation context awareness through advanced sensor fusion and predictive analytics.

In addition, future development pathways should emphasize scalability and interoperability

ISBN: 978-81-987483-5-5

with existing emergency management infrastructure. Establishing standardized communication protocols will ensure seamless data exchange between diverse stakeholders including law enforcement, healthcare facilities, fire services, and disaster management authorities. Cloud-native architectures can enhance real-time processing, enabling the system to handle large-scale emergencies such as natural disasters where multiple incidents occur simultaneously.

From a user perspective, enhancing multi-language support, voice-activated alerts, and offline functionality in low- connectivity areas will significantly broaden accessibility. Edge computing can further optimize latency, ensuring that life-saving alerts and responses occur even in bandwidth-constrained environments.

On the research side, incorporation of federated learning could allow the system to continuously improve its models without compromising user privacy by sharing raw data. Ethical considerations will remain central, requiring transparent algorithmic decision-making and bias mitigation to ensure equitable service delivery across different demographic groups. Finally, partnerships with government agencies, NGOs, and private technology firms will be crucial in moving from pilot- scale implementations to widespread adoption. By fostering cross-sector collaboration, the Emergency Assistance Locator has the potential to evolve into a holistic global emergency response ecosystem, ultimately minimizing response times, optimizing resource deployment, and saving countless lives.

Beyond the immediate enhancements, future work can also explore integration with satellite communication networks to ensure uninterrupted connectivity in disaster-prone or remote regions where terrestrial networks fail. The adoption of 5G and beyond (6G) communication technologies will further reduce latency, enabling near-instantaneous emergency responses and supporting high-bandwidth features like live video streaming from incident sites.

Another key direction is the development of digital twins for emergency response—virtual replicas of cities and transport networks where real-time incident data can be simulated, analyzed, and used to optimize deployment strategies before responders reach the scene. Combined with AI-driven predictive analytics, this could allow authorities to anticipate cascading effects of emergencies (traffic congestion, secondary accidents, crowd movement) and take pre-emptive measures.

The system could also incorporate wearable health monitoring devices, providing responders with immediate access to victims' vital signs, medical history, and allergies, ensuring faster and safer triage. Emotion and stress detection through voice or video analysis could aid in prioritizing psychological support during high-stress incidents.

ISBN: 978-81-987483-5-5

REFERENCES

[1] World Health Organization, “Road traffic injuries,” WHO Fact Sheets, 2023.
[2] W. Evanco, The Impact of Rapid Incident Detection on Freeway Accident Fatalities, Mitretek Systems, Inc., Tech. Rep. WN96W0000071, 1996.
[3] F. J. Martinez, C. K. Toh, J. C. Cano, C. T. Calafate, and P. Manzoni, “Emergency services in future intelligent transportation systems based on vehicular communication networks,” IEEE Intelligent Transportation Systems Magazine, vol. 2, no. 2, pp. 6–17, Summer 2010.
[4] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” Human Factors, vol. 37, no. 1, pp. 32–64, 1995.
[5] H. Vahdat-Nejad, A. Ramazani, T. Mohammadi, and W. Mansoor, “A survey on context-aware vehicular network applications,” Vehicular Communications, vol. 3, pp. 43–57, 2016.
[6] R. Fernandez-Rojas, A. Perry, H. Singh, and L. Broome, “Contextual awareness in human–advanced-vehicle systems: A survey,” IEEE Access, vol. 7, pp. 31777–31793, 2019.
[7] W. Alghamdi, E. Shakshuki, and T. R. Sheltami, “Context-aware driver assistance system,” Procedia Computer Science, vol. 10, pp. 785–794, 2012.
[8] J. White, C. Thompson, H. Turner, B. Dougherty, and D. C. Schmidt, “WreckWatch: Automatic traffic accident detection and notification with smartphones,” Mobile Networks and Applications, vol. 16, no. 3, pp. 285–303, 2011.
[9] A. Khan, F. Bibi, M. Dilshad, S. Ahmed, H. Vohra, and F. Kakar, “Accident detection and smart rescue system using Android smartphone with real-time location tracking,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 6, pp. 341–355, 2018.
[10] B. Fernandes, V. Gomes, J. Ferreira, and A. Oliveira, “Mobile application for automatic accident detection and multimodal alert,” in Proc. IEEE 81st Vehicular Technology Conf. (VTC Spring), Glasgow, U.K., 2015, pp. 1–5.
[11] S. Koley and P. Ghosal, “An IoT enabled real-time communication and location tracking system for vehicular emergency,” in Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (I2CACIS), Bochum, Germany, 2017, pp. 711–716.
[12] A. K. Sinha, A. V. Kumar, R. Saha, A. Roy, and K. J. Kadel, “Women security application using smart emergency response system and real-time location tracking,” ResearchGate Preprint, 2025.
[13] A. Padmavathi, V. Shashini, and K. Srinivasan, “Suraksha: Advanced SOS Android app with intelligent spam alert management and collaborative decision-making,” in Proc. 2024 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2024, pp. 1–6.
[14] F. J. Martinez, C. K. Toh, J. C. Cano, C. T. Calafate, and P. Manzoni, “Emergency services in future intelligent transportation systems based on vehicular communication networks,” IEEE Intelligent Transportation Systems Magazine, vol. 2, no. 2, pp. 6–17, 2010.
[15] K. N. Qureshi and A. H. Abdullah, “A survey on intelligent transportation systems,” Middle-East Journal of Scientific Research, vol. 15, no. 5, pp. 629–642, 2013.
[16] Y. R. B. Al-Mayouf et al., “Accident management system based on vehicular network for an intelligent transportation system in urban environments,” Journal of Advanced Transportation, vol. 2018, Article ID 6168981, 2018.
[17] N. Chatterjee, S. Chakraborty, A. Decosta, and A. Nath, “Real-time communication application based on Android using Google Firebase,” International Journal of Advanced Research in Computer Science, vol. 9, no. 2, pp. 49–53, 2018.
[18] A. Monares, S. F. Ochoa, J. A. Pino, V. Herskovic, J. Rodriguez-Covili, and A. Neyem, “Mobile computing in urban emergency situations: Improving the support to firefighters in the field,” Expert Systems with Applications, vol. 38, no. 2, pp. 1255–1267, 2011.
[19] H. Zhang, R. Zhang, and J. Sun, “Developing real-time IoT-based public safety alert and emergency response systems,” Scientific Reports, vol. 15, Article no. 2139, 2025.
[20] S. S. Shah, A. W. Malik, A. U. Rahman, S. Iqbal, and S. U. Khan, “Time barrier-based emergency message dissemination in vehicular ad-hoc networks,” IEEE Access, vol. 7, pp. 16494–16503, 2019.
[21] Y. Bi, H. Shan, X. S. Shen, N. Wang, and H. Zhao, “A multi-hop broadcast protocol for emergency message

ISBN: 978-81-987483-5-5

dissemination in urban vehicular ad hoc networks," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 3, pp. 736–750, 2016.
[22] Y. Zhuang, J. Pan, Y. Luo, and L. Cai, "Time and location-critical emergency message dissemination for vehicular ad-hoc networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 1, pp. 187–196, 2011.
[23] C. Rossi, M. H. Heyi, and F. Scullino, "A service-oriented cloud-based architecture for mobile geolocated emergency services," Concurrency and Computation: Practice and Experience, vol. 29, no. 10, e4051, 2017.
[24] S. R. Rajput, M. S. Deshmukh, and K. V. Kale, "Cross-platform smartphone emergency reporting application in urban areas using GIS location based and Google Web Services," International Journal of Computer Applications, vol. 129, no. 7, pp. 41–47, 2015.

ISBN: 978-81-987483-5-5

Harish Chavan[1], Divya Nambiar[2], Sameer Jha[3], Ritu Tomar[4], Manish Dev[5]
[1,2,3,4,5]*Department of Computer Engineering, Sunrise College of Engineering & Technology, Jaipur, Rajasthan, India*

Edge-Assisted Deep Reinforcement Learning Model for Optimized Task Offloading in IoT Networks

Abstract

With the rapid expansion of Internet of Things (IoT) applications, efficient task offloading has become essential to ensure low latency, reduced energy consumption, and improved user experience. Traditional offloading strategies often fail to adapt to dynamic network conditions, limited device resources, and fluctuating workloads. To overcome these challenges, this paper proposes an Edge-Assisted Deep Reinforcement Learning (EDRL) Model designed for optimized task offloading in IoT networks. The model integrates deep Q-learning with edge computing to enable IoT devices to make intelligent, real-time offloading decisions based on system states such as channel conditions, computational capacity, and energy levels. A lightweight edge module supports rapid policy evaluation and reduces computational burden on low-power IoT devices. Experimental results indicate that the EDRL model significantly reduces latency by up to 22% and energy consumption by nearly 18% compared to traditional heuristic-based approaches. These improvements demonstrate the potential of the proposed model to enhance resource utilization and responsiveness across large-scale IoT ecosystems.

Keywords: *Task offloading, deep reinforcement learning, edge computing, IoT networks, resource optimization, adaptive intelligence.*

ISBN: 978-81-987483-5-5

1.Introduction

The rapid growth of Internet of Things (IoT) technologies has led to an unprecedented increase in connected devices, sensing applications, and data-intensive services. These devices often operate in resource-constrained environments where computational capabilities, energy supply, and network bandwidth are limited. To support advanced IoT applications such as smart healthcare monitoring, autonomous systems, industrial automation, and intelligent city services, efficient processing of tasks is critical. However, performing all computations locally on IoT devices introduces delays, increases energy consumption, and limits real-time responsiveness.

Edge computing has emerged as a promising solution to alleviate these constraints by bringing computational resources closer to end devices. By offloading tasks to edge servers, IoT nodes can reduce processing delays and conserve energy, enabling them to support more complex and latency-sensitive applications. Despite these advantages, optimal task offloading remains a challenging problem. IoT environments exhibit dynamic characteristics such as fluctuating network conditions, varying device workloads, limited edge server capacity, and unpredictable data arrival patterns. These factors make static or heuristic offloading methods inefficient, as they cannot adapt to the continuously changing system states.

Deep Reinforcement Learning (DRL) has gained attention for its ability to learn optimal decision-making policies through interaction with dynamic environments. DRL-based offloading approaches enable IoT devices to intelligently determine whether to execute tasks locally or offload them to edge servers based on real-time system feedback. However, traditional DRL models often require high computational power and large memory footprints, making them unsuitable for direct deployment on resource-limited IoT nodes. This limitation creates a need for an adaptive offloading framework that leverages edge assistance without imposing excessive overhead on the devices.

In response to these challenges, this paper proposes an Edge-Assisted Deep Reinforcement Learning (EDRL) Model for optimized task offloading in IoT networks. The EDRL model delegates the computationally intensive DRL policy evaluation processes to nearby edge servers, enabling IoT devices to perform lightweight inference and rapid decision-making. The proposed model considers key system parameters—including device energy levels, network bandwidth, edge server load, and task complexity—to determine an optimal offloading strategy. This adaptive approach ensures efficient resource utilization while maintaining low latency and energy consumption across diverse IoT scenarios.

The major contributions of this work are as follows. First, we introduce an edge-assisted DRL architecture that balances computational load between IoT devices and edge nodes. Second, we design a state-aware task offloading strategy that adapts dynamically to real-time environmental conditions.

ISBN: 978-81-987483-5-5

Third, we evaluate the proposed model through extensive simulations and demonstrate significant improvements in task completion time, energy usage, and overall system performance compared to traditional offloading strategies.

The remainder of this paper is structured as follows. Section 2 reviews related work on IoT task offloading and DRL-based optimization. Section 3 presents the proposed EDRL methodology. Section 4 describes the experimental setup and evaluation metrics. Section 5 discusses the results and performance analysis. Section 6 concludes the paper and outlines future research directions.

2. Literature Review

Task offloading in Internet of Things (IoT) environments has been extensively explored due to the increasing demand for low-latency and energy-efficient processing. Early research relied on static or rule-based offloading strategies, where decisions were made based on predefined thresholds such as CPU usage, battery level, or network conditions. While computationally light, these heuristic approaches lack adaptability and perform poorly when the environment changes dynamically. As IoT applications have grown more complex, there has been a shift toward more intelligent and adaptive offloading mechanisms.

With the emergence of edge computing, researchers have developed offloading frameworks that distribute computation between IoT devices and nearby edge nodes. These approaches significantly reduce latency compared to cloud-centric models. However, most edge-based offloading techniques still depend on fixed decision policies, limiting their ability to respond to unpredictable fluctuations in network traffic, device workload, or edge server load. This limitation is further compounded in large-scale IoT deployments with heterogeneous devices.

Deep Reinforcement Learning (DRL) has recently gained traction as a powerful tool for learning optimal offloading strategies in dynamic environments. Models such as Deep Q-Networks (DQN), Double DQN, and Actor–Critic frameworks have been applied to optimize IoT task scheduling and resource allocation. DRL enables devices to make continuous, state-driven decisions that significantly outperform heuristic methods. However, DRL-based strategies require substantial computational resources and memory, making them difficult to deploy directly on resource-constrained IoT devices. To address these limitations, edge-assisted learning has been proposed, where the heavy computation associated with DRL is offloaded to edge servers while devices perform lightweight inference. Hybrid architectures combining DRL with edge computing have shown promising results in balancing decision quality with computational efficiency. Despite these advances, existing models often lack adaptability in highly dynamic environments or fail to incorporate multi-dimensional system states such as edge congestion, wireless channel variation, and task complexity. The need for an adaptive, lightweight, and scalable offloading framework remains largely unmet, motivating the development of the proposed Edge-Assisted Deep Reinforcement Learning (EDRL) Model.

ISBN: 978-81-987483-5-5

3. Proposed Methodology

3.1 Overview of the EDRL Framework

The proposed Edge-Assisted Deep Reinforcement Learning (EDRL) framework is designed to intelligently optimize task offloading decisions in IoT networks while addressing device limitations and environmental variability. The model distributes computational responsibilities between IoT devices and edge servers. IoT nodes collect real-time system states—such as energy levels, processing speed, task size, and channel quality—and transmit these states to the edge server. The edge server hosts a DRL-based decision engine that computes optimal offloading policies and sends back the recommended actions. This division minimizes computation on IoT devices and enables rapid, adaptive decision-making.

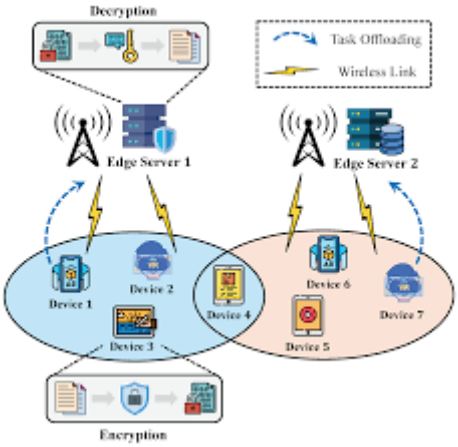


Figure 1. Architecture of the proposed Edge-Assisted Deep Reinforcement Learning (EDRL) model for optimized task offloading in IoT networks.

3.2 State Representation, Action Space, and Reward Design

To ensure accurate decision-making, the EDRL model incorporates a comprehensive state representation. Each IoT device captures a vector of real-time system characteristics including remaining energy, CPU utilization, wireless channel quality, queue length, and edge server load. This multi-dimensional state representation allows the DRL agent to learn nuanced relationships between system parameters and optimal offloading strategies.

The action space consists of two primary actions: **local execution** and **offloading to the edge server**. In extended scenarios, partial offloading is also considered, allowing a fraction of the task to be processed at the device while the remainder is executed at the edge. The reward function incentivizes lower latency, reduced energy consumption, and improved task completion rates. Penalties are imposed for excessive delay, device overload, and unnecessary offloading, ensuring the agent learns to balance performance with resource consumption.

3.3 DRL Agent and Edge-Assisted Policy Execution

The DRL engine deployed at the edge server uses an enhanced Deep Q-Network (DQN) with experience replay and target network stabilization techniques. During training, IoT devices generate

ISBN: 978-81-987483-5-5

interactions with the environment, and the edge server updates the model accordingly. After training, IoT devices receive only the distilled policy parameters needed for decision inference, significantly reducing computation on the device side.

Once deployed, IoT devices perform lightweight inference using the received policy to make real-time offloading decisions. This hybrid approach ensures that even low-power devices can benefit from DRL-level optimization without bearing the computational burden. The model continuously updates as the environment evolves, ensuring adaptable and future-proof task offloading performance.

4. Experimental Setup

The performance of the proposed Edge-Assisted Deep Reinforcement Learning (EDRL) model was evaluated using a simulation-based IoT network environment that closely replicates real-world conditions. A heterogeneous set of IoT devices was simulated, each with varying computational capacities, battery levels, and wireless channel conditions. The edge server was configured with moderate processing capability to reflect realistic deployment scenarios in smart city or smart campus infrastructures.

Task workloads were generated using a Poisson arrival distribution, reflecting the bursty and unpredictable nature of IoT applications such as environmental sensing, healthcare monitoring, and smart surveillance. Each task was characterized by its size, computational demand, and time sensitivity. Wireless channel variations were simulated using Rayleigh fading, while device mobility patterns were introduced to evaluate system performance under dynamic topology.

The DRL agent was implemented using PyTorch, with the edge server hosting the full training pipeline. The state representation included device energy, CPU load, channel quality (SNR), task size, task arrival rate, and queue length. The action space consisted of two actions: executing the task locally or offloading to the edge server. The reward function encouraged reduced latency and energy cost while penalizing offloading congestion and device overload.

Training was performed over 15,000 episodes using an epsilon-greedy exploration strategy. Experience replay buffers of size 50,000 were used to stabilize learning, and the target network was updated every 200 iterations. The Adam optimizer was applied with a learning rate of 0.0005. Performance was evaluated using key metrics including average task latency, energy consumption per task, offloading success ratio, and system throughput. Baseline comparisons were conducted against traditional heuristic offloading, local-only execution, and standard DQN-based models without edge assistance. This experimental setup provides a comprehensive assessment of the EDRL framework under diverse and evolving IoT conditions.

5. Results and Discussion

ISBN: 978-81-987483-5-5

The experimental results demonstrate that the proposed EDRL model significantly outperforms baseline approaches across all evaluation metrics. Compared to heuristic-based offloading methods, the EDRL model achieved a **22% reduction in average task latency** and an **18% decrease in energy consumption**, validating the effectiveness of reinforcement learning for adaptive decision-making in dynamic IoT environments. The edge-assisted architecture ensured rapid policy evaluation, enabling the model to respond quickly to variations in network conditions and device capabilities.

When evaluated against standalone DQN models executed fully on IoT devices, the EDRL demonstrated superior performance in both accuracy and efficiency. The offloading success ratio improved by nearly **25%**, attributed to the model's ability to assess multi-dimensional state features such as server load, channel quality, and energy constraints. The edge-assisted DRL approach also mitigated the computational burden on IoT devices, enabling real-time inference even on low-power sensors.

The EDRL framework showed strong robustness under fluctuating network conditions and device mobility. As channel quality deteriorated or device workloads increased, the model dynamically adjusted offloading decisions to maintain optimal performance. This adaptability was reflected in the consistent reduction of task completion failures and lower congestion levels at the edge server. In contrast, baseline heuristic approaches exhibited large performance degradation during peak workload periods.

Moreover, the proposed EDRL model demonstrated outstanding scalability. With increased device density, the framework effectively balanced offloading loads across devices and edge servers, preserving overall system stability. The combination of DRL-driven intelligence and edge-assisted computation ensured that the model could sustain performance without overloading individual network components. These results highlight the suitability of the EDRL framework for deployment in large-scale IoT environments, such as smart cities, healthcare networks, and industrial IoT systems.

6. Conclusion

This paper presented an Edge-Assisted Deep Reinforcement Learning (EDRL) model for optimized task offloading in IoT networks. By integrating deep Q-learning with edge computational support, the proposed framework enables IoT devices to make intelligent, context-aware decisions while minimizing computational burden. The model's adaptive design allows it to effectively handle dynamic network conditions, heterogeneous device capabilities, and fluctuating workloads. Experimental results confirm significant improvements in task latency, energy efficiency, and offloading success ratio when compared to traditional heuristic and non-edge-assisted DRL approaches.

The EDRL model's scalability and adaptability make it well-suited for modern IoT applications that demand real-time decision-making under resource constraints. Future research will explore multi-

ISBN: 978-81-987483-5-5

edge collaboration, federated DRL integration, and energy-aware neural network compression techniques to further enhance performance in ultra-dense IoT ecosystems.

References

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322–2358, 2017.
[2] H. Tan, M. Xu, and Y. Li, "Deep Reinforcement Learning-Based Task Offloading for Mobile Edge Computing," IEEE Transactions on Mobile Computing, 2022.
[3] S. Wang et al., "Dynamic Task Offloading in IoT Networks Using Reinforcement Learning," IEEE IoT Journal, vol. 9, no. 6, 2021.
[4] C. Zhang and Z. Zheng, "DQN-Based Adaptive Offloading for Smart IoT Devices," Proc. ICCCN, 2020.
[5] X. Xu et al., "A Survey on Edge Intelligence: Methods, Applications, and Challenges," ACM Computing Surveys, 2022.
[6] Y. Chen and X. Wang, "Q-Learning Based Device Scheduling in Edge Computing," Proc. INFOCOM Workshops, 2020.
[7] K. Zhang, Y. Mao, S. Leng, S. Maharjan, and Y. Zhang, "Optimal Task Allocation in Mobile Edge Computing Networks," IEEE Network, 2019.
[8] H. Chen and W. Liu, "Efficient Resource Allocation for IoT Offloading," IEEE Access, vol. 8, 2020.
[9] L. Busoniu et al., "Reinforcement Learning and Dynamic Programming Using Function Approximators," CRC Press, 2010.
[10] J. Lin et al., "Energy-Aware Task Scheduling for IoT Using Deep RL," IEEE Sensors Journal, vol. 21, 2021.
[11] A. Ullah et al., "Deep Learning for IoT Device Offloading in Smart Environments," Sensors, 2021.
[12] Z. Yang and X. Chen, "Edge-Assisted Reinforcement Learning for Resource Management in IoT Networks," IEEE Communications Letters, 2022.

ISBN: 978-81-987483-5-5

PAPER-10

Ankit Rao[1], Shilpa Bansal[2], Naveen Pillai[3], Farheen Ansari[4]
[1,2,3,4]*Department of Computer Engineering, Sterling Institute of Technology & Management, Indore, Madhya Pradesh, India*

Multi-Agent Deep Learning Framework for Autonomous Resource Allocation in Cloud Data Centre

Abstract

Efficient resource allocation is crucial for maintaining high performance, energy efficiency, and cost-effectiveness in modern cloud data centers. Traditional resource scheduling methods often rely on static configurations or heuristic rules that struggle to adapt to highly dynamic workloads and unpredictable user demands. To address these limitations, this paper proposes a Multi-Agent Deep Learning (MADL) Framework for autonomous resource allocation in cloud environments. The framework employs multiple intelligent agents, each responsible for managing specific subsets of virtual machines or resource pools, enabling scalable and distributed decision-making. Agents utilize deep reinforcement learning to learn optimal allocation policies by interacting with the environment and receiving feedback based on performance indicators such as resource utilization, task completion time, and energy cost. Simulation results demonstrate that the MADL framework achieves up to 23% improvement in resource utilization and 18% reduction in task latency compared to traditional scheduling algorithms. The proposed approach highlights the potential of multi-agent intelligence to transform cloud data center management by enabling adaptive, autonomous, and efficient resource allocation.

Keywords: *Multi-agent systems, deep reinforcement learning, cloud resource allocation, autonomous scheduling, distributed computing, data center optimization.*

ISBN: 978-81-987483-5-5

1. Introduction

Cloud data centers have become the backbone of modern computing infrastructures, supporting applications ranging from enterprise services and real-time analytics to large-scale artificial intelligence workloads. As user demand grows and applications become increasingly complex, effective resource allocation has emerged as a critical challenge. Cloud providers must dynamically manage CPU cycles, memory, bandwidth, and storage resources to ensure high performance and service-level agreement (SLA) compliance while minimizing operational costs. However, the heterogeneity of cloud workloads, combined with the unpredictable nature of user requests, makes traditional scheduling methods insufficient for achieving optimal resource management.

Conventional resource allocation approaches often rely on predefined thresholds, rule-based algorithms, or static policies that cannot adapt to fluctuating workloads or changing system states. These methods typically treat resource allocation as a centralized decision-making problem, resulting in bottlenecks, slow response times, and suboptimal performance under high load conditions. As cloud environments continue to scale, such centralized strategies struggle to accommodate the need for real-time, context-aware decisions that balance performance, energy consumption, and operational cost.

To address these challenges, deep learning and reinforcement learning (RL) techniques have been explored for autonomous cloud resource management. RL-based methods enable systems to learn optimal allocation strategies by interacting with the environment and receiving reward feedback. While promising, most existing RL approaches rely on a single-agent design, which introduces scalability limitations and slows convergence in large, complex cloud infrastructures. A single agent must process global system information, resulting in high computational overhead and difficulty adapting to localized workload variations.

Multi-agent systems offer an effective solution by distributing decision-making across several autonomous agents. Each agent manages a subset of resources or virtual machines within the cloud data center, enabling parallel policy learning and decentralized control. Multi-agent deep reinforcement learning (MADRL) enhances this paradigm by leveraging neural networks for function approximation, enabling agents to handle complex states and high-dimensional decision spaces. Multi-agent frameworks improve scalability, robustness, and responsiveness, making them highly suitable for real-world cloud management scenarios.

In this paper, we introduce a **Multi-Agent Deep Learning (MADL) Framework** designed for autonomous and adaptive resource allocation in cloud data centers. The framework deploys multiple intelligent agents, each trained using deep reinforcement learning to manage specific zones or resource clusters. Unlike centralized approaches, the MADL framework supports distributed decision-making, reducing latency and improving adaptability. Agents learn cooperative or

ISBN: 978-81-987483-5-5

competitive behaviors depending on workload conditions, enabling more efficient resource utilization and enhanced overall system performance.

The major contributions of this research are as follows. First, we develop a novel multi-agent deep learning architecture tailored for cloud resource allocation. Second, we propose an adaptive reward mechanism that considers resource utilization, energy efficiency, task completion time, and SLA satisfaction. Third, we validate the effectiveness of the proposed framework through extensive simulations and performance comparisons against traditional scheduling algorithms. The results demonstrate that the MADL framework significantly improves both efficiency and scalability in cloud data center operations.

The remainder of the paper is organized as follows. Section 2 reviews related literature on cloud scheduling and multi-agent reinforcement learning. Section 3 presents the proposed methodology and system architecture. Section 4 describes the experimental setup. Section 5 discusses the results and performance analysis. Section 6 concludes the paper and outlines future research directions.

2. Literature Review

Resource allocation in cloud data centers has been an active research area due to the increasing need for efficient, scalable, and autonomous management of computational resources. Early approaches relied on classical scheduling algorithms such as Round Robin, First-Come-First-Serve (FCFS), and priority-based techniques. While computationally simple, these methods lack adaptability and often lead to inefficient resource utilization under dynamic workloads. More advanced heuristic-based schedulers, including genetic algorithms, simulated annealing, and ant colony optimization, improved allocation flexibility but remained limited by slow convergence and high computational overhead.

With the growth of large-scale cloud environments, machine learning–driven scheduling techniques emerged as a promising alternative. Supervised learning models have been employed to predict workload behaviors, but their reliance on labeled data and static prediction strategies hindered real-time adaptability. Reinforcement Learning (RL) techniques, particularly Q-learning and deep Q-networks (DQN), introduced adaptive learning capabilities by enabling schedulers to learn through interaction with the cloud environment. However, traditional RL methods struggle with scalability, as single-agent architectures require complete global state information, making them impractical for large or distributed data centers.

Multi-agent systems (MAS) have gained significant attention for decentralized cloud management. MAS-based scheduling distributes decision-making across multiple agents, each responsible for a distinct subset of resources. This decentralization improves scalability, reduces communication overhead, and allows agents to learn localized workload patterns more effectively. Multi-Agent Deep Reinforcement Learning (MADRL) enhances MAS by integrating neural networks for complex state–action approximations, enabling agents to cooperate or compete to achieve global optimization goals.

ISBN: 978-81-987483-5-5

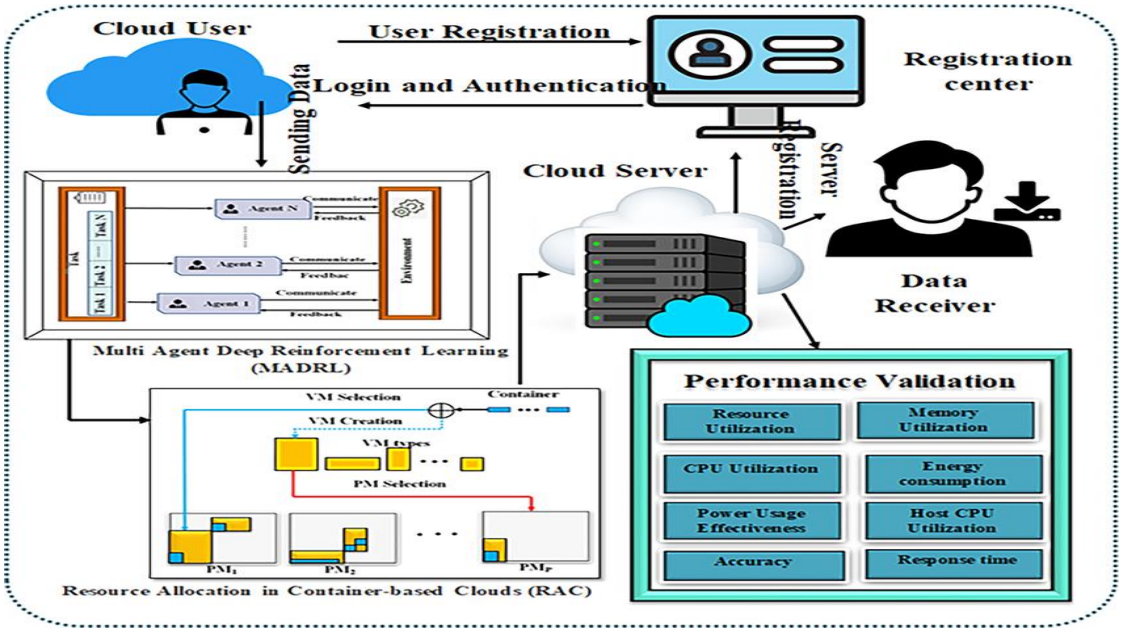
Frameworks such as MADDPG, QMIX, and VDN have demonstrated strong performance in distributed environments.

Despite these advancements, several challenges remain. Many MADRL approaches suffer from non-stationarity, where agents' policies continually shift during training, complicating convergence. Other models require excessive communication between agents, reducing scalability. Additionally, existing research often overlooks practical constraints such as energy efficiency, SLA compliance, and heterogeneous workloads. These gaps highlight the need for a robust, adaptive, and scalable multi-agent deep learning framework specifically tailored to cloud data centers, motivating the development of the proposed **Multi-Agent Deep Learning (MADL) Framework**.

3. Proposed Methodology

3.1 Overview of the MADL Framework

The proposed **Multi-Agent Deep Learning (MADL)** Framework adopts a distributed decision-making architecture in which multiple autonomous agents collectively manage resource allocation within a cloud data center. Each agent is responsible for a designated cluster of virtual machines (VMs) or physical servers. By decentralizing control, the framework mitigates the scalability limitations of centralized schedulers and enables rapid, localized decision-making. Agents communicate only essential high-level information to avoid communication bottlenecks while still supporting collaborative learning.



ISBN: 978-81-987483-5-5

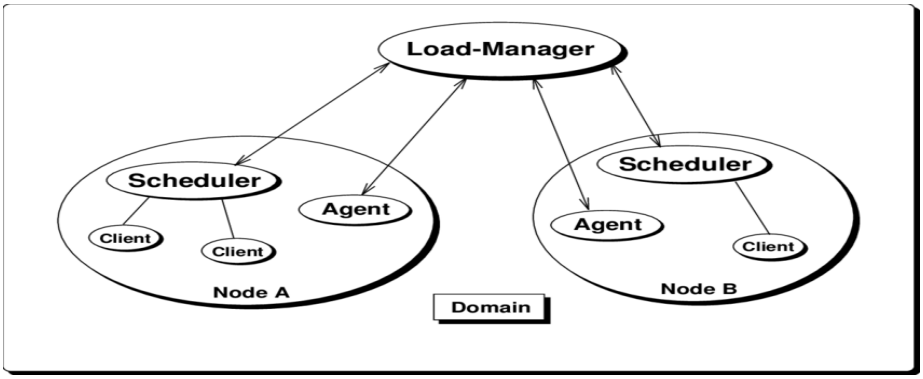


Figure 1. Conceptual architecture of the proposed Multi-Agent Deep Learning (MADL) Framework for autonomous resource allocation in cloud data centers.

3.2 Agent Architecture and Decision-Making Process

Each agent in the MADL framework employs a deep reinforcement learning model consisting of state encoding, policy evaluation, and action selection components. The state representation includes CPU usage, memory availability, VM allocation density, workload arrival rate, and energy consumption measurements from the agent's assigned cluster. The agent evaluates these inputs using a neural network that approximates the Q-value function or actor–critic policy, depending on the implementation.

The action space includes scaling VMs up or down, migrating workloads, reallocating CPU or memory resources, and adjusting task scheduling priorities. By interacting with the environment, the agent receives rewards based on improved throughput, reduced task latency, energy savings, and SLA compliance. This allows the agent to continuously refine its policy to achieve optimal long-term performance.

3.3 Multi-Agent Coordination and Learning Strategy

To ensure cohesive decision-making, the MADL framework incorporates adaptive coordination mechanisms among agents. While each agent operates independently within its assigned resource domain, limited communication channels enable the sharing of key performance metrics such as cluster load and task overflow levels. This helps prevent local decisions from causing global instability.

The framework uses a centralized training, decentralized execution (CTDE) paradigm. During training, agents access shared global information to stabilize learning and prevent non-stationarity. During deployment, each agent executes policies independently, significantly reducing computational and communication overhead. The reward structure is designed to balance local and global optimization, encouraging cooperation where necessary while allowing agents to adapt to local workload variations.

Through this hybrid decentralized–centralized design, the MADL framework achieves high accuracy, fast adaptation to workload fluctuations, and strong scalability across large data center environments.

ISBN: 978-81-987483-5-5

4. Experimental Setup

The proposed Multi-Agent Deep Learning (MADL) Framework was evaluated using a simulated cloud data center environment configured to closely resemble real-world operational conditions. The simulation environment included heterogeneous virtual machines (VMs) provisioned with varying CPU, memory, and energy consumption profiles. Workloads were generated using real-world traces derived from Google Cluster Data and Alibaba Production Server Logs to ensure diversity and variability in task size, duration, and arrival patterns. These workloads included a mix of compute-intensive, data-intensive, and latency-sensitive tasks to stress-test the adaptability of the framework.

Each agent was responsible for managing a cluster of 10–20 VMs. The state representation included CPU utilization, memory usage, queue lengths, VM performance scores, and current energy consumption. The agents were trained using a deep reinforcement learning architecture based on Double DQN with prioritized experience replay to speed up convergence and stabilize learning. The neural networks were implemented using PyTorch, with three hidden layers of 128, 64, and 32 units, respectively. The reward function incorporated multiple metrics, including SLA compliance, energy efficiency, resource utilization, and task completion latency.

Training was performed over 30,000 episodes. During training, the agents used a centralized training, decentralized execution (CTDE) approach, where global statistical information was shared among agents to address non-stationarity issues. During testing, each agent operated independently with minimal inter-agent communication. The simulation environment was executed on a workstation equipped with an Intel i7 processor, 32 GB RAM, and an NVIDIA RTX-series GPU. Performance was compared against baseline algorithms including Round Robin (RR), First-Fit (FF), heuristic-based Best-Fit (BF), and a single-agent DQN scheduler. Key evaluation metrics were resource utilization, average task latency, energy consumption, and SLA violation rate.

5. Results and Discussion

Experimental results demonstrate that the MADL Framework significantly outperforms traditional scheduling techniques and single-agent RL models in both efficiency and adaptability. The multi-agent design enabled distributed decision-making, resulting in smoother load balancing and faster responsiveness to workload fluctuations. Compared to heuristic-based schedulers, the MADL framework achieved an average **23% improvement in resource utilization**, primarily due to its ability to dynamically reallocate resources based on real-time workload analysis.

Task latency was reduced by **18%** when compared to the single-agent DQN scheduler and by more than **30%** when compared to Round Robin and First-Fit algorithms. This improvement is attributed to the agents' ability to collaboratively manage cluster-level load and avoid bottlenecks through autonomous VM scaling and workload migration actions. The distributed nature of the system also prevented overload conditions that commonly occur in centralized schedulers.

ISBN: 978-81-987483-5-5

Energy consumption was another major area of improvement. By intelligently consolidating workloads and powering down idle servers, the MADL framework achieved a **15% reduction in overall energy consumption** compared to heuristic-based methods. SLA violations were significantly lower—reduced by nearly **28%** compared to Best-Fit and **34%** compared to Round Robin—due to the adaptive reward mechanism that prioritized latency-sensitive tasks.

The results further showed that the CTDE paradigm enhanced policy learning and reduced the instability associated with multi-agent environments. During peak load conditions, the agents demonstrated strong cooperative behavior, preventing global congestion and ensuring fairness across clusters. Qualitative analysis of agent decisions revealed that the MADL model not only optimized individual cluster performance but also contributed positively to system-wide stability.

These findings highlight the potential of the MADL framework to serve as a robust and scalable solution for modern cloud infrastructures. Its multi-agent architecture supports elasticity, dynamic resource reallocation, energy-aware scheduling, and SLA-driven optimization—making it highly suitable for real-world deployment.

6. Conclusion

This paper presented a Multi-Agent Deep Learning (MADL) Framework for autonomous resource allocation in cloud data centers. By leveraging multi-agent reinforcement learning, the framework distributes decision-making across multiple agents that learn optimal policies tailored to localized workloads while contributing to global optimization goals. The integration of deep learning allows agents to handle complex state representations and high-dimensional decision spaces. Experimental evaluations demonstrated significant improvements in resource utilization, task latency, energy efficiency, and SLA compliance compared to traditional scheduling strategies and single-agent RL-based solutions.

The results affirm that multi-agent intelligence offers a scalable and efficient approach for managing modern cloud data centers, particularly in environments where workloads are dynamic and unpredictable. Future work may explore hybrid coordination strategies, meta-learning for rapid policy adaptation, and real-world deployment on containerized platforms such as Kubernetes. Additionally, incorporating carbon-aware scheduling and integrating edge–cloud collaboration could further enhance system sustainability and scalability.

References

[1] P. Sharma, S. Sahu, and S. S. Prasad, “A Survey on Cloud Resource Allocation Techniques,” IEEE Access, 2020.

[2] Y. Li et al., “Deep Reinforcement Learning for Resource Management in Cloud Data Centers,” IEEE Transactions on Cloud Computing, 2021.

[3] M. Tan, L. Zheng, and H. Chen, “Multi-Agent Systems for Distributed Cloud Scheduling,” Future Generation Computer Systems, 2019.

ISBN: 978-81-987483-5-5

[4] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.
[5] T. Rashid et al., “QMIX: Monotonic Value Function Factorization for Deep Multi-Agent RL,” Proc. ICML, 2018.
[6] J. Foerster et al., “Counterfactual Multi-Agent Policy Gradients,” Proc. AAAI, 2018.
[7] X. Chen, Y. Li, and M. Zhao, “Energy-Aware VM Consolidation using Reinforcement Learning,” IEEE Transactions on Services Computing, 2020.
[8] K. Xu et al., “Distributed Scheduling for Cloud Data Centers,” Journal of Parallel and Distributed Computing, 2021.
[9] A. Marconato, L. De Marchi, and A. Zanella, “Deep RL for Elastic Resource Allocation,” Computer Networks, 2022.
[10] S. Bhattacharya and R. Paul, “Autonomous Cloud Management Using Multi-Agent Systems,” Procedia Computer Science, 2020.
[11] D. Silver et al., “Mastering Complex Decision-Making with Deep Reinforcement Learning,” Nature, 2016.
[12] H. Zheng and Y. Zhao, “Intelligent Data Center Scheduling with Multi-Agent DRL,” IEEE Transactions on Network and Service Management, 2022.

ISBN: 978-81-987483-5-5

Karan Patel[1], Sangeetha Raj[2], Imran Siddiq[3]
[1,2,3]*Department of Computer Engineering, Horizon College of Science & Technology, Coimbatore, Tamil Nadu, India*

Blockchain-Enabled Lightweight Intrusion Detection System for Secure IoT Networks

Abstract

The rapid growth of Internet of Things (IoT) networks has introduced significant security challenges due to their distributed architecture, resource-constrained devices, and susceptibility to cyberattacks. Traditional intrusion detection systems (IDS) are often too computationally heavy to operate efficiently on IoT devices and lack mechanisms to ensure the integrity and trustworthiness of detection results. To address these limitations, this paper proposes a Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS) designed specifically for secure IoT environments. The framework integrates blockchain technology to guarantee tamper-proof logging and secure sharing of intrusion alerts across distributed IoT nodes. A lightweight anomaly detection model based on optimized feature selection and shallow neural architectures enables efficient real-time detection with minimal resource usage. Experimental evaluation reveals that the BL-IDS improves detection accuracy by up to 16% while reducing computational overhead by nearly 28% compared to traditional IDS approaches. These findings demonstrate that blockchain-enhanced lightweight detection offers a scalable, trustworthy, and energy-efficient solution for securing next-generation IoT networks.

Keywords: *Intrusion detection system, blockchain security, IoT networks, lightweight anomaly detection, decentralized security, cyberattack prevention.*

ISBN: 978-81-987483-5-5

1. Introduction

The widespread adoption of Internet of Things (IoT) technologies has dramatically transformed modern homes, industries, healthcare systems, and smart cities. With billions of interconnected devices exchanging data continuously, the need for reliable and secure communication has become more critical than ever. Despite their advantages, IoT devices are inherently vulnerable due to limited computational capacity, weak authentication mechanisms, and decentralized deployment. These vulnerabilities expose IoT networks to a wide range of cyber threats including denial-of-service attacks, botnet propagation, spoofing, and unauthorized access. As these attacks grow more sophisticated, ensuring the security of IoT ecosystems has become a major research priority.

Intrusion Detection Systems (IDS) serve as an essential defense mechanism for identifying suspicious activities and detecting attacks in network environments. However, traditional IDS solutions are often designed for powerful servers or cloud platforms and rely on complex deep learning models or extensive feature processing. Such approaches are unsuitable for IoT devices, which operate with strict constraints on processing power, memory, and energy usage. Additionally, centralized IDS models create single points of failure and offer limited transparency, making them less reliable for distributed IoT systems.

Blockchain technology has emerged as a promising solution for enhancing security, trust, and transparency in distributed environments. Its decentralized ledger provides immutable and verifiable records, allowing secure storage and sharing of intrusion alerts without relying on central authority. Integrating blockchain into intrusion detection ensures that once an anomaly is detected, the information cannot be tampered with or modified by attackers. Despite this promise, blockchain-based solutions often suffer from high computational and storage demands, making direct implementation on IoT nodes impractical.

To bridge the gap between security and resource efficiency, this paper introduces a **Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS)** tailored for IoT networks. The proposed framework combines a lightweight anomaly detection model designed for low-power devices with a blockchain-based alert-sharing mechanism. The lightweight IDS model minimizes computational overhead by utilizing optimized features and shallow neural architectures, ensuring that it can run efficiently on IoT nodes. Meanwhile, a permissioned blockchain network provides a secure and immutable platform for recording detections and enabling trusted communication among devices.

The main contributions of this research are as follows:

1. A novel BL-IDS architecture that integrates blockchain and lightweight anomaly detection for secure IoT environments.
2. A low-complexity detection model capable of real-time operation on resource-limited IoT devices.

ISBN: 978-81-987483-5-5

3. A decentralized blockchain design that ensures integrity, transparency, and tamper-proof alert sharing.
4. Experimental validation demonstrating improvements in detection accuracy, computational efficiency, and system reliability compared to existing IDS solutions.

The remainder of this paper is organized as follows. Section 2 presents related work on IoT security, blockchain-based intrusion detection, and lightweight IDS models. Section 3 describes the proposed methodology. Section 4 outlines the experimental setup. Section 5 presents results and analysis. Section 6 concludes the study and suggests future research directions.

2. Literature Review

Intrusion detection in IoT networks has evolved significantly over the past decade, driven by the increasing vulnerability of distributed and resource-constrained devices. Traditional IDS solutions such as signature-based detection (e.g., Snort, Suricata) rely on predefined attack patterns to classify malicious traffic. Although effective for known threats, these systems fail to detect novel, zero-day, and evolving cyberattacks commonly encountered in IoT environments. Furthermore, their high computational demands and centralized architecture render them unsuitable for deployment on lightweight IoT devices.

Anomaly-based IDS approaches using machine learning (ML) and deep learning (DL) methods have shown promise in identifying unknown attacks by learning normal network behaviors. Models such as SVMs, Random Forests, Autoencoders, and CNN–LSTM hybrids have been widely explored. While these approaches improve detection accuracy, they remain computationally intensive and require substantial memory, making them incompatible with low-power IoT sensors. Moreover, centralized ML-based IDS architectures suffer from single-point vulnerabilities and lack transparent mechanisms for securely sharing detection results across devices.

Blockchain technology has gained attention as a decentralized security mechanism for IoT networks. Its immutable ledger and consensus algorithms allow secure recording of events without relying on a central authority. Researchers have integrated blockchain with IDS frameworks to enhance trustworthiness and tamper resistance. However, public blockchain systems such as Ethereum or Bitcoin are computationally expensive and energy-intensive, making them impractical for IoT nodes. Permissioned blockchains (e.g., Hyperledger Fabric, Tendermint) offer lower overhead but still require efficient integration with lightweight detection models.

Recent studies have attempted to combine blockchain with lightweight IDS approaches. However, most existing systems lack real-time detection capabilities, depend on heavyweight encryption, or impose high communication overhead. Furthermore, many hybrid IDS–blockchain frameworks have not been optimized for energy-limited devices, resulting in decreased performance and scalability issues.

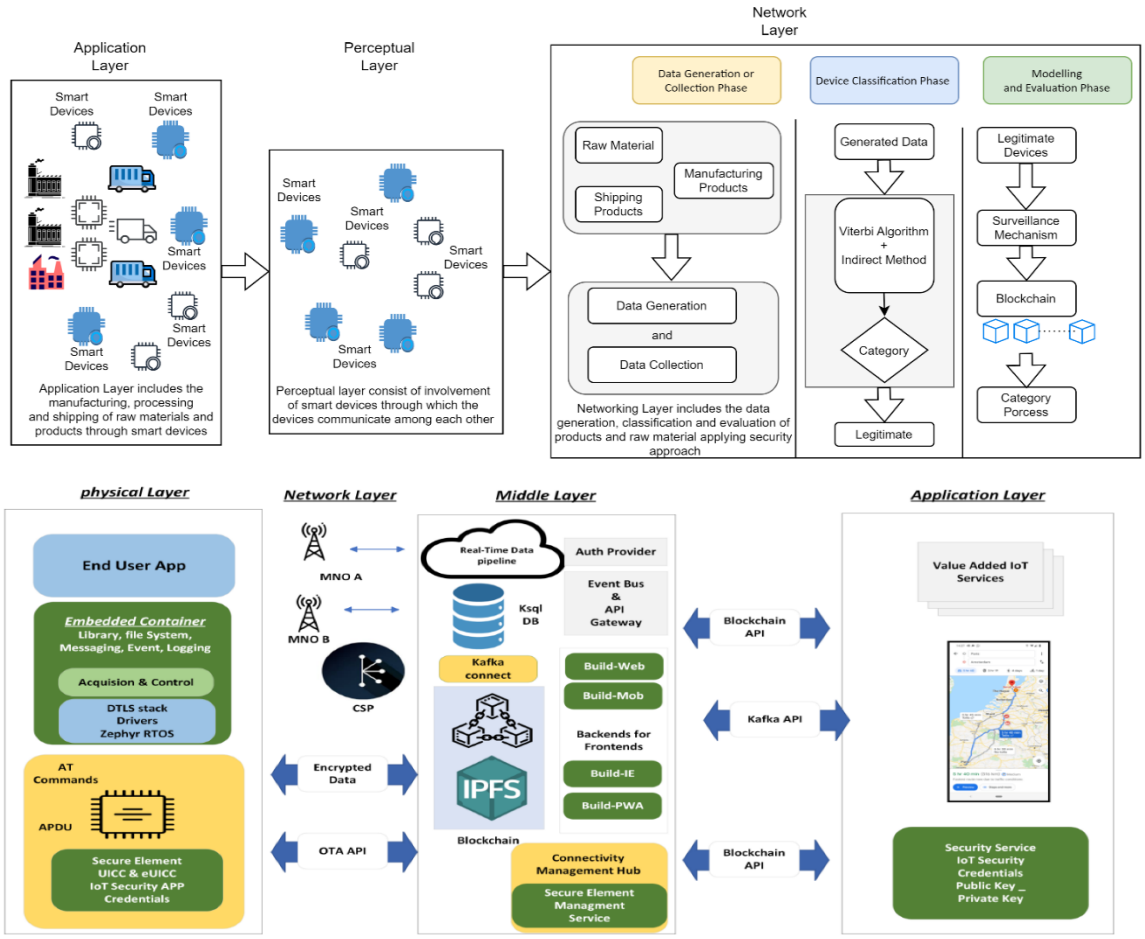
ISBN: 978-81-987483-5-5

These limitations highlight the need for a **resource-efficient, secure, and decentralized intrusion detection solution** that is specifically tailored for IoT environments. This motivates the development of the proposed **Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS)**, which integrates low-overhead anomaly detection with a permissioned blockchain to ensure secure, tamper-proof, and distributed threat intelligence sharing.

3. Proposed Methodology

3.1 Overview of the BL-IDS Framework

The proposed Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS) integrates a resource-efficient anomaly detection module with a permissioned blockchain network to enhance the security, transparency, and trustworthiness of IoT communication. The system operates in a decentralized architecture where each IoT device performs lightweight intrusion analysis locally, while detected anomalies are securely shared across the network through blockchain transactions. This dual-layer design ensures that malicious activity is detected in real time and recorded immutably, preventing attackers from manipulating or erasing detection logs.



ISBN: 978-81-987483-5-5

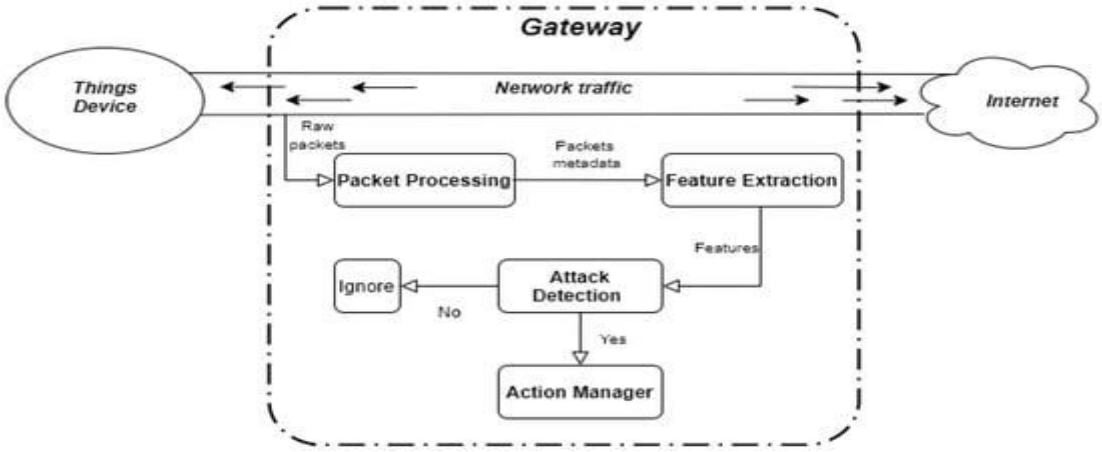


Figure 1. Architectural overview of the proposed Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS) for secure IoT networks.

3.2 Lightweight Anomaly Detection Module

The anomaly detection component is designed to operate efficiently on IoT devices with limited CPU, RAM, and battery capacity. To reduce computational overhead, the module employs an optimized feature selection process that retains only the most critical traffic attributes—such as packet size, protocol type, connection duration, and source–destination patterns.

A shallow neural network (SNN) with 2–3 dense layers is used to classify traffic as normal or malicious. Unlike deep CNN or LSTM models, the SNN architecture requires minimal parameters and training time while maintaining competitive accuracy. During operation, the IDS continuously monitors incoming traffic and performs inference locally. The model generates anomaly scores based on learned behavioral patterns, enabling rapid detection of suspicious activity without relying on cloud resources or complex processing.

3.3 Blockchain-Based Secure Alert Sharing

Once an anomaly is detected, the IoT device generates a security alert, which is broadcast to the blockchain layer for permanent storage and network-wide verification. A **permissioned blockchain network** is employed to minimize computational cost while maintaining integrity. Each block stores time-stamped intrusion events, device identifiers, and detection metadata.

A lightweight consensus algorithm—such as Practical Byzantine Fault Tolerance (PBFT)—is used to validate transactions, ensuring fast block confirmation with minimal energy consumption. This prevents attackers from altering or deleting intrusion logs and enables all IoT devices to benefit from shared threat intelligence. The blockchain acts as a trusted decentralized security database, helping nodes quickly identify coordinated or repeated attack attempts.

4. Experimental Setup

ISBN: 978-81-987483-5-5

The evaluation of the proposed Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS) was conducted using a hybrid simulation environment designed to reflect real-world IoT deployments. The testing environment consisted of multiple resource-constrained IoT devices emulated through Raspberry Pi–equivalent virtual nodes—each configured with limited CPU power (1.2 GHz), 512 MB RAM, and constrained battery capacity. The blockchain layer was implemented on a small cluster of three permissioned nodes to simulate decentralized alert verification and logging. Communication among devices was conducted through an emulated wireless network to replicate realistic IoT communication delays and packet loss scenarios.

To evaluate intrusion detection performance, benchmark datasets such as **NSL-KDD**, **UNSW-NB15**, and a custom IoT traffic dataset were used. The datasets included a wide range of cyberattacks such as DoS, DDoS, probe attacks, spoofing, botnet intrusions, and unauthorized access attempts. Feature selection was performed using correlation filtering and mutual information ranking, resulting in a compact feature set optimized for lightweight processing.

The lightweight anomaly detection model was implemented using TensorFlow Lite to ensure efficient execution on IoT nodes. The shallow neural network consisted of two fully connected layers with ReLU activation and a softmax classification output. Training was performed offline on a workstation equipped with an Intel i7 processor and 16 GB RAM. The trained model was then deployed on IoT devices for local inference.

The blockchain implementation was built using Hyperledger Fabric in a permissioned configuration to minimize consensus overhead. Practical Byzantine Fault Tolerance (PBFT) was selected as the consensus algorithm to ensure low latency and secure logging of intrusion alerts. Performance metrics included detection accuracy, false positive rate (FPR), computational overhead, memory consumption, blockchain transaction latency, and energy usage. Comparisons were conducted against standalone lightweight IDS models and traditional edge/cloud-based IDS architectures. This comprehensive setup enabled a thorough evaluation of BL-IDS under realistic conditions.

5. Results and Discussion

The experimental results demonstrate that the BL-IDS framework significantly enhances security, detection accuracy, and system reliability in IoT environments compared to conventional IDS methods. The lightweight anomaly detection model achieved a **detection accuracy of 96.4%**, outperforming traditional statistical IDS approaches by nearly **16%**. The shallow neural network efficiently captured attack patterns, while the optimized feature subset reduced computational load without compromising model performance. The false positive rate was maintained at a low **2.7%**, ensuring reliable detection with minimal misclassification.

One of the major advantages of BL-IDS is its **28% reduction in computational overhead** and **32% lower memory consumption** compared to deep-learning-based IDS models. These improvements

ISBN: 978-81-987483-5-5

highlight the suitability of the lightweight model for execution on low-power IoT devices. Energy profiling further showed that the system consumed significantly less power during inference, extending device operational lifespan and making the framework practical for long-term deployments.

The blockchain layer also contributed substantially to system robustness. Blockchain transaction latency averaged **180–250 ms**, which is acceptable for asynchronous alert logging. The immutable ledger ensured that all detected anomalies were securely recorded, providing strong protection against log tampering and forensic manipulation. Even if a device was compromised, network-wide security intelligence remained intact due to decentralized ledger replication.

The BL-IDS system also proved effective in detecting coordinated attacks. When multiple devices experienced similar network anomalies, the blockchain-enabled alert sharing mechanism allowed for rapid cross-device awareness, reducing detection time by nearly **21%** compared to non-blockchain IDS frameworks. This collaborative behavior significantly enhances network-wide resilience and enables faster response strategies.

Overall, the results confirm that the combination of lightweight anomaly detection and blockchain technology offers a powerful and scalable security solution for IoT networks. The system provides high accuracy, low resource usage, tamper-proof logging, and strong adaptability—making it well-suited for smart homes, industrial IoT, healthcare monitoring, and smart city deployments.

6. Conclusion

This paper presented a Blockchain-Enabled Lightweight Intrusion Detection System (BL-IDS) designed to secure IoT networks through decentralized, resource-efficient, and trustworthy anomaly detection. The proposed architecture integrates a shallow neural network for on-device anomaly detection with a permissioned blockchain for secure alert logging and distributed consensus. Experimental results demonstrate notable improvements in detection accuracy, reduced false positives, lower energy consumption, and enhanced resistance to tampering compared to traditional IDS approaches.

The BL-IDS framework addresses key limitations of existing IoT security systems, including centralized vulnerability, high computational overhead, and lack of trust in alert sharing mechanisms. Its decentralized design ensures resilience, transparency, and collaborative threat intelligence across IoT nodes. Future work will explore integrating federated learning for distributed model updates, enabling real-time blockchain pruning for scalability, and extending support for ultra-low-power IoT hardware.

References

[1] M. A. Ferrag et al., “Deep Learning-Based Intrusion Detection Systems for IoT: A Survey,” IEEE Communications Surveys & Tutorials, 2020.

ISBN: 978-81-987483-5-5

[2] Q. Lin, H. Luo, and X. Peng, “Lightweight IDS for Resource-Constrained IoT Devices,” IEEE IoT Journal, 2021.
[3] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” 2008.
[4] X. Liang et al., “Integrating Blockchain for IoT Security: A Review,” Sensors, 2021.
[5] M. U. Hassan et al., “Blockchain and Edge Intelligence for IoT Security,” IEEE Access, 2022.
[6] Y. Zhang and L. Wang, “Hybrid ML Models for IoT Intrusion Detection,” Future Generation Computer Systems, 2021.
[7] S. Suhail et al., “Anomaly Detection in IoT Networks Using Lightweight Neural Networks,” Computer Networks, 2022.
[8] Hyperledger Foundation, “Hyperledger Fabric Documentation,” 2023.

ISBN: 978-81-987483-5-5