International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 12, No. 1, 2025, pp. 65-71.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Queue-Aware Multi-Objective Edge Offloading and Resource Provisioning for Dense IoT Networks Under Synthetic Workload Dynamics

K.P. Uvarajan [a], M. Kavitha [b], E. Geetha [c], Shaik Sadulla [d]

[a] *Associate Professor, Electronics and Communication Engineering, K.S.R. College of Engineering, Tiruchengode, Tamil Nadu, India. E-mail: uvarajan@ksrce.ac.in*
[b] *Assistant Professor, Electronics and Communication Engineering, Saveetha School of Engineering, Chennai, Tamil Nadu, India.*
[c] *Professor, Department of Medical Electronics, Sengunthar Engineering College, Thiruchengode, Tamil Nadu, India.*
[d] *Professor, Department of Electronics and Communication Engineering, KKR & KSR Institute of Technology and Sciences, Vinjanampadu, Andhra Pradesh, India.*

Abstract

The intensive growth of Internet of Things (IoT) devices has led to the explosion of the quantity of data produced at the network edges. The high-energy efficiency and the ability to provide low latency demand distributed computing architectures to process such data efficiently. Edge computing is one of the promising paradigms with the ability to perform computational tasks at a position nearer to the sources of data and thus alleviating congestion in the network and enhancing response time. Nonetheless, thick IoT networks present the major issues associated with resources distribution, task scheduling, and queuing in edge servers. Traditional offloading techniques commonly focus on a single objective optimization relying on minimization of latency or energy use, and this might result in poor performance when operating in dynamic network constraints. The study is a multi-objective edge offloading framework based on queues that is suggested to operate in dense IoT networks with nonuniform workloads and variable resource allocation. The suggested system incorporates the queue length predictor, the resource-sensitive schedule, and the resource on-demand provisioning innovations into a single decision system. Large-scale IoT traffic patterns are synthetically created as models of workload without the use of real-time data gathering. The experiments with simulations show that the suggested queue-aware offloading strategy can decrease the task completion time by a significant margin, and the energy consumption of edge nodes is balanced. The results point to the efficiency of considering the behaviour of queues when managing edge resources of next-generation IoT systems.

Keywords: Edge Computing, IoT Task Offloading, Queue-Aware Scheduling, Resource Provisioning, Distributed Computing.

DOI: 10.71086/IAJSE/V12I1/IAJSE1251
Received: 30 December 2024/Revised: 29 January 2025/Accepted: 27 February 2025/Published: 31 March 2025

Introduction

The rapid expansion of Internet of Things (IoT) technologies has revolutionized the contemporary computing environment as it allows billions of interdependent devices to gather and share information in real time. Such devices are sensors, wearable electronics, smart appliances, and industrial monitoring systems that are generating large amounts of data that need instant processing. The classical cloud computing systems are inefficient to support such workloads due to the creation of communication delay and network overload when transmitting raw sensor data to central data centres. Edge computing has as such become a viable option that enables data processing to take place nearer to the generation. Through real-time control and flexibility in applications, edge computing reduces communication delays and enhances responsiveness to real-time applications by deploying computational resources at the edge nodes like gateways or local servers. Although these have their benefits, thick IoT networks pose a number of challenges to edge computing systems. With an ever-growing number of associated devices, edge servers have to deal with highly dynamic workloads that can vary in terms of arrival rates and the types of computations involved. Delays and poor service quality may be experienced when queueing occurs in edge nodes when there are several tasks that are received at the same time. The traditional offloading methods of tasks also tend to presuppose the unchanging workload and the constant supply of resources, which makes them less effective in practise (Chen et al., 2015). Moreover, most resource allocation algorithms aim to maximise only one measure including latency or energy consumption, and do not tradeoff between multiple system objectives. In this study, the authors of this paper solve this problem by proposing a queue-aware multi-objective offloading system, dynamically distributing computational tasks across edge resources depending on forecasted queue conditions and workload properties.

Literature Review

Edge computing has attracted a lot of interest as a paradigm of distributed computing that has the capacity to support latency-sensitive IoT applications. Such initial studies in the area saw extensive attention on offloading computations out of resource-constrained IoT devices to the immediate edge server (Zhou et al., 2021). These methods illustrated that edge-procedure task processing can be an adequate method to cut down network latency and enhance the general performance of the system than the conventional cloud-based infrastructure (Li & Ye, 2021). Nevertheless, early offloading mechanisms were usually based on basic decision guidelines depending on attributes like device energy consumption or communication bandwidth (Chai et al., 2023).

These were further refined by more advanced techniques to optimally place tasks, and apply mathematical optimization to the task placement and heuristic scheduling algorithms (Prasath, 2025). Other studies came up with queue-based models to describe service times and task arrival trends in edge computing systems (Sobin, 2020; Kumar, 2024; Rajan, 2025). The models enabled researchers to perform system performance analysis of systems at different workloads, and to come up with scheduling policies that minimise waiting time (Wang et al., 2019). However, most of these methods make the assumption that edge resources are homogeneous or the distribution of task arrivals is at rest, which do not reflect the challenge posed by dense IoT networks (Yousefpour et al., 2018).

Recent studies have studied the multi-objective optimization models that, at the same time, address latency, energy use, and the use of network bandwidth (Abdullah, 2025). These frameworks offer a more realistic view of the trade-offs in the performance of the system but in most cases consume a lot of computational power to optimise (Velliangiri, 2025). The adaptive resource provisioning and task offloading mechanisms have also been suggested to overcome the resource management issues in IoT-edge systems to enhance the scalability of the system and service quality (Jin et al., 2022; Thompson & Sonntag, 2025).

The other new area of research is the predictive scheduling methods that can be used to estimate the future patterns of the workload in order to make better decisions concerning resource allocation (Punam, 2025). They use machine learning and smart decision-making models to adapt dynamically task placement strategy according to network conditions (David et al., 2025). Nevertheless, real-time decision-making processes might frequently involve the use of such intelligent frameworks, which demand strong system modelling and architectural support (Patel, 2025).

Besides advances in algorithms, a number of studies have discussed co-design methods of hardware and software to enhance the performance of embedded and edge computing systems (Gupta & Quamara, 2020). Adaptive embedded control architectures have also been suggested to better flexibility of the system and distributive edge control mechanisms (Bakar et al., 2023). The sophisticated data-driven modelling methods also facilitate dynamic system behaviour of the intricate computing systems (Tamrakar, 2025; Kavitha, 2025).

It has also studied distributed trust management methods used to enhance reliability and security in wireless communication and edge service settings (Khanna & Kaur, 2020; Chowdhury & Chakma, 2025). The coordination strategies represented by learning guides have been suggested to enhance the energy efficiency and balancing of the workload between the distributed edge nodes (Du et al., 2020). The strategies are especially vital in massive infrastructures of IoT where heterogeneous devices produce erratic workloads (Madhanraj, 2024).

Simultaneously, the recent research has investigated the optimization algorithms on resource allocation within the edge cloud continuum to facilitate the overall efficiency of computation (Sabella et al., 2016). The superior computation frameworks also take into account surrogate optimization and computer-hardware adaptive acceleration methods that help in real-time choice making within large scale distributed systems (Ding et al., 2019). The extensive studies of computation offloading in cloud-fog-edge settings also demonstrate the necessity of dynamic and scalable task allocation algorithms that can handle dynamism in workload (Soy, 2025).

Moreover, a recent study has proposed learning-based task offloading schemes based on deep reinforcement learning models to dynamically control queue-sensitive decision-making in edge clouds systems (Punam, 2025). Predictive control architectures that are adaptable to hardware have also been implemented to facilitate intelligent navigation and real-time control in embedded edge platforms (Arvinth, 2025; Bosco, 2025; Ramya, 2025).

Regardless of such progress, several current literature works are based on actual datasets or application-specific workloads that are not always easily accessible at the early stages of system design. Synthetic workload-based simulation evaluation is thus a good substitute to examine the behaviour of edge computing architectures through controlled experimentation. These methods can provide researchers with the ability to analyze the dynamics of queues, the distribution of workloads, and the use of resources in large-scale IoT settings in a systematic way and ensure reproducibility and flexibility in the experimental design.

Methodology and Experimental Setup

Edge Computing Architecture Model

The proposed framework represents a distributed edge computing system that comprises of several edge servers that are linked to a network of IoT devices. All IoT devices also produce computational tasks that can be done either on-premises or offloaded to an edge server. Edge servers have processing units that can perform several tasks at once yet they have limited computers. Thus, there is a need to have an effective allocation of tasks to achieve balanced performance of the system. As shown in fig. 1, the queue-aware edge computing system looks like this.

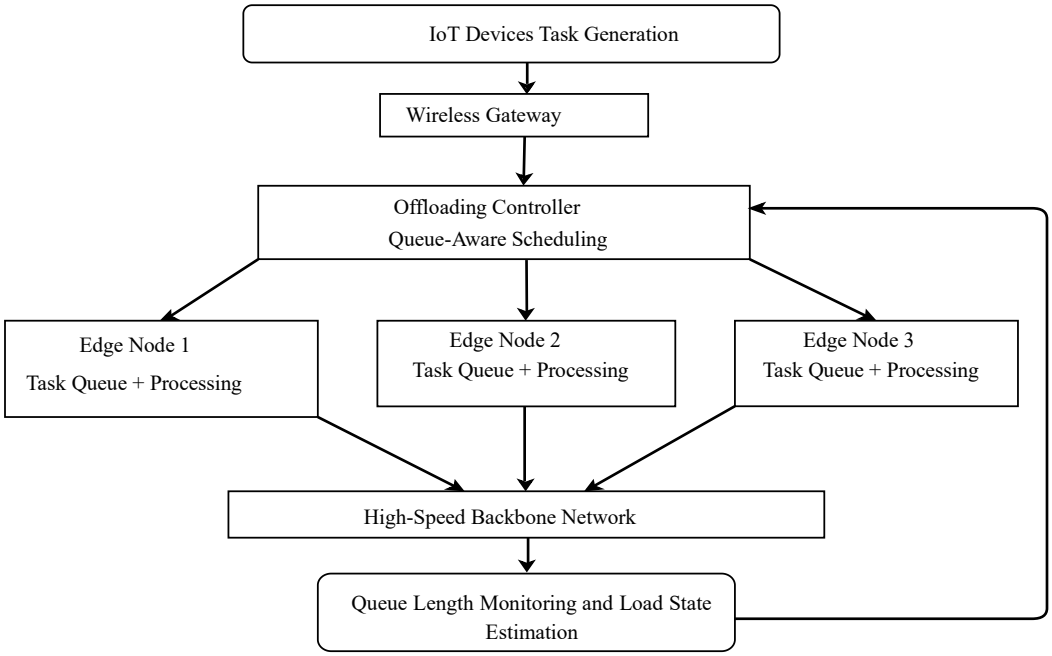


Fig. 1: Queue-Aware Edge Computing Architecture for IoT Networks

Under this architecture, the IoT devices are sending request of tasks to the closest edge server using a wireless gateway. Each edge server has a task queue that is objected to pending computational jobs that are awaiting execution. The time that this line spends indicates the immediate service of the server and is important to decide whether new tasks can be handled locally or be redirected to a different edge node. To model large-scale IoT systems, the system has a variety of edge computers across a high-speed backplane network. In case a node becomes congested, the offloading controller can bypass it and distribute tasks to the neighboring nodes with shorter queues. This distributed architecture guarantees that resources of computations are used efficiently in the network. Table 1 presents the main parameters of configuration to be used to model the edge computing system.

Table 1: Edge System Configuration Parameters

Parameter	Value
Edge nodes	10
IoT devices	500
Average task size	2 MB
Edge CPU frequency	2.5 GHz
Network bandwidth	100 Mbps

These parameters reflect typical large-scale IoT deployments.

Queue-Aware Offloading Optimization Model

The offloading decision is a multi-objective optimization problem, where the system latency, as well as the energy consumption, are taken into consideration. These metrics are combined by the objective function in the following way:

$$F = \lambda L + (1 - \lambda)E$$

Where L and E are the task and energy consumption respectively. The weighting factor λ is used to associate the relative importance of latency and energy efficiency. The optimization algorithm will compare several edge nodes and choose the node at which the objective function is expected to be the lowest. The decision process is also fitted with queue length prediction on the basis of estimating the waiting time before the execution of a task.

Algorithm 1 Queue-Aware Offloading Decision

1 Receive task request from IoT device
2 Estimate queue length at each edge node
3 Predict execution delay and energy cost
4 Compute optimization objective value
5 Select node with minimum objective cost
6 Dispatch task to selected edge node

This algorithm makes sure that an assignment of tasks to the nodes is made to ones that will provide a balanced performance.

Synthetic Workload Generation and Simulation Environment

A synthetic workload generator was also created to test the proposed framework and simulate the conditions of large-scale IoT networks in terms of traffic patterns (Suresh Kumar & Vaduganathan, 2025). The tasks generated by the generator have different arrival times and computational needs, which are realistic in terms of the sensor networks bursts or periodic update of monitoring devices. The workload patterns are poisson distributed to acquire the random appearance of the tasks. Simulation experiments were undertaken in various conditions where the density of the devices and the arrival rate of the tasks was manipulated. The workload parameters were as illustrated in table 2.

Table 2: Synthetic IoT Workload Parameters

Parameter	Value
Task arrival rate	10–50 tasks/sec
Processing cycles	10^6–10^7
Simulation duration	1000 sec

The behaviour of queues and the utilisation of resources can be carefully studied in this simulated environment with varying operating conditions.

Results and Discussion

Task Completion Latency

The initial series of experiments was used to measure the mean latency of task completion with varying network loads. Fig. 2 demonstrates the comparison of the latency performance of the proposed queue-aware framework and a baseline offloading strategy, which does not consider any information about the queue length.

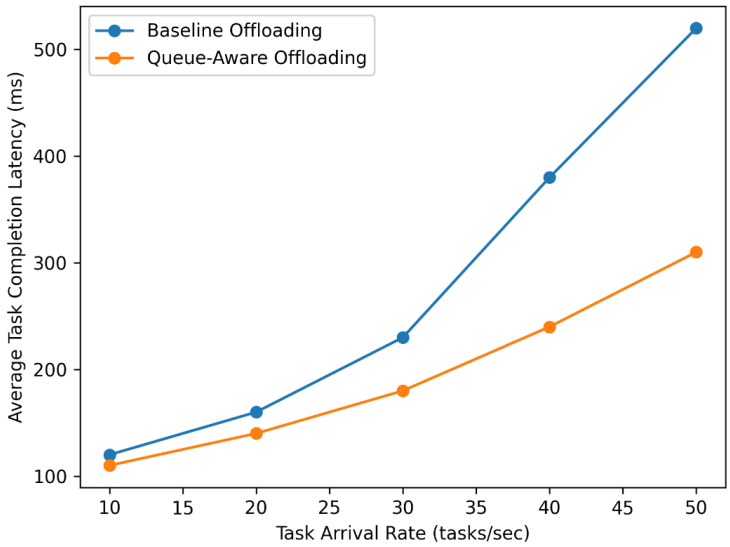


Fig. 2: Task Completion Latency Comparison

Findings indicate that queue-aware strategy ensures a significant reduction of latency in high workloads. At task arrival rates greater than 40 tasks/sec, the baseline strategy has rapid growth in waiting time because of queue congestion at the heavily loaded nodes. On the contrary, the proposed framework reallocates work between edge servers occurring near each other and hence avoids accumulation of queues.

Queue Stability and Load Balancing

Load balancing is essential in dense IoT settings where task distribution may result in extreme system malfunction. The given queue-aware algorithm constantly evaluates the queue lengths and includes/removes tasks to achieve a balanced system execution. The simulation findings show that the framework is effective in avoiding congestion of queues even when there is a burst in traffic. The proposed method will allocate workloads more evenly among the available nodes compared with the baseline algorithm that will send tasks to the first node in the path. This even distribution enhances the stability of the system and makes sure that the computer resources are not wasted.

Energy Consumption Analysis

Another major performance measurement of edge computing infrastructures is energy efficiency. The experiments conducted with simulations show that the queue-conscious framework also leads to a decrease of the total energy use as the idle time in the underutilized nodes is minimized and overheating of overloaded servers is prevented. The system also ensures that the processor utilisation is consistent throughout edge nodes, as workloads distribution is more even. This balanced functioning lowers the requirement of the high frequency bursts of computations that generally raise the electrical consumption.

Discussion

The findings of this paper indicate that the task scheduling with queues consideration is a significant factor in ensuring that the edge computing systems operating in dense IoT space are running efficiently. Conventional offloading techniques usually give weight to variables like communication latency or energy consumption of devices without making an account of the dynamic condition of edge server queues. Subsequently, work can be held up at one node and idle at another. The framework proposed overcomes this limitation, since the predictions of the queue length are included in the decision-making procedure. This allows the system to spread out the computation in the network. The

other contribution that this study has offered is the application of synthetic workload modelling to assess edge computing architectures. Largely, real-world datasets of large-scale IoT networks tend to be hard to access, particularly at initial phases of system design. Synthetic workload Generation It offers a scalable and repeatable method of experimenting with allocation strategies against controlled conditions. The outcomes of the experiment prove that the combination of queue awareness and multi-objective optimization can help to achieve the improvement of both latency and energy efficiency. This is where these advances in scheduling are identified to optimize the scalability of future edge computing infrastructures and to increase the dependability of these infrastructures.

Conclusion

There is a queue-conscious multi-objective offloading framework of tasks in dense IoT edge computing proposed by this study. The proposed system combines the study of the queue length prediction, scheduling that is sensitive to workload, and distributed resource provisioning to minimise the latency and energy consumption. The simulation outcomes show that the framework is much better than the conventional offloading strategies in completing tasks in less latency and load balancing. The system is more stable in its performance under different conditions of workload by adding the element of a queue dynamics to the decisions of resource allocation. The results show that queue-conscious scheduling systems are needed to address the growing computational needs of massive IoT implementations. The next step of research will be the investigation of machine learning methods to predict workload patterns and enhance the methods of edge resources allocation (Saranya, 2025).

References

Zhou, J., Tian, D., Sheng, Z., Duan, X., & Shen, X. (2021). Distributed task offloading optimization with queueing dynamics in multiagent mobile-edge computing networks. *IEEE Internet of Things Journal*, *8*(15), 12311-12328. https://doi.org/10.1109/JIOT.2021.3063509

Li, X., & Ye, B. (2021). Latency-Aware Computation Offloading for 5G Networks in Edge Computing. *Security and Communication Networks*, *2021*(1), 8800234. https://doi.org/10.1155/2021/8800234

Chai, Z. Y., Liu, X., & Li, Y. L. (2023). A computation offloading algorithm based on multi-objective evolutionary optimization in mobile edge computing. *Engineering Applications of Artificial Intelligence*, *121*, 105966. https://doi.org/10.1016/j.engappai.2023.105966

Gupta, B. B., & Quamara, M. (2020). An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols. *Concurrency and computation: Practice and experience*, *32*(21), e4946. https://doi.org/10.1002/cpe.4946

Prasath, C. A. (2025). Adaptive Embedded Learning-Control Architectures for Reconfigurable Sensorless Motor Drive Platforms. *Journal of VLSI and Embedded System Design*, *1*(1), 1-9.

Bakar, K. B. A., Zuhra, F. T., Isyaku, B., & Sulaiman, S. B. (2023). A review on the immediate advancement of the Internet of Things in wireless telecommunications. *IEEE access*, *11*, 21020-21048. https://doi.org/10.1109/ACCESS.2023.3250466

Tamrakar, G. (2025). Adaptive Trust Evaluation under Dynamic Service Conditions Using Distributed Verification. *Recent Advances in Next-Generation Wireless Communication Systems*, *1*(1), 36-42.

Sobin, C. C. (2020). A survey on architecture, protocols and challenges in IoT. *Wireless Personal Communications*, *112*(3), 1383-1429. https://doi.org/10.1007/s11277-020-07108-5

Khanna, A., & Kaur, S. (2020). Internet of things (IoT), applications and challenges: a comprehensive review. *Wireless Personal Communications*, *114*(2), 1687-1762. https://doi.org/10.1007/s11277-020-07446-4

Yousefpour, A., Ishigaki, G., Gour, R., & Jue, J. P. (2018). On reducing IoT service delay via fog offloading. *IEEE Internet of things Journal*, *5*(2), 998-1010.

Jin, X., Hua, W., Wang, Z., & Chen, Y. (2022). A survey of research on computation offloading in mobile cloud computing. *Wireless Networks*, *28*(4), 1563-1585. https://doi.org/10.1007/s11276-022-02920-2

Du, M., Wang, Y., Ye, K., & Xu, C. (2020). Algorithmics of cost-driven computation offloading in the edge-cloud environment. *IEEE Transactions on Computers*, *69*(10), 1519-1532.

Sabella, D., Vaillant, A., Kuure, P., Rauschenbach, U., & Giust, F. (2016). Mobile-edge computing architecture: The role of MEC in the Internet of Things. *IEEE Consumer Electronics Magazine*, *5*(4), 84-91.

Wang, J., Pan, J., Esposito, F., Calyam, P., Yang, Z., & Mohapatra, P. (2019). Edge cloud offloading algorithms: Issues, methods, and perspectives. *ACM Computing Surveys (CSUR)*, *52*(1), 1-23. https://doi.org/10.1145/3284387

Abdullah, D. (2025). Uncertainty-Aware Representation Learning with Physical Consistency for Robust Decision Processes in Large-Scale Data Systems. *Journal of Scalable Data Engineering and Intelligent Computing*, *2*(3), 9-17.

Velliangiri, A. (2025). Hardware-Adaptive Acceleration of Multi-Fidelity Surrogate Optimization for Real-Time Wind Farm Micro-Siting. Archives of Embedded and IoT Systems Engineering*, 1*(1),35-44.

Chen, X., Jiao, L., Li, W., & Fu, X. (2015). Efficient multi-user computation offloading for mobile-edge cloud computing. *IEEE/ACM transactions on networking*, *24*(5), 2795-2808. https://doi.org/10.1109/TNET.2015.2487344

Patel, P. (2025). Hardware-Adaptive Learning-Assisted Predictive Control Architecture for Real-Time Path-Constrained Navigation. *Journal of Advanced Antenna and RF Engineering,1*(1), 32-40.

Ding, Z., Xu, J., Dobre, O. A., & Poor, H. V. (2019). Joint power and time allocation for NOMA–MEC offloading. *IEEE Transactions on Vehicular Technology*, *68*(6), 6207-6211. https://doi.org/10.1109/TVT.2019.2907253

Soy, A. (2025). Scalable Graph-Enhanced Data Engineering for Intelligent Identity Analytics in Hybrid Cloud Ecosystems. *Journal of Scalable Data Engineering and Intelligent Computing, 2*(2), 9-16.

Punam, S. R. (2025). Automated Distributed Learning Pipelines for Multi-Agent Graph Intelligence in 6G IoT Systems. *SECITS Journal of Scalable Distributed Computing and Pipeline Automation*, *2*(2), 18-27.

Arvinth, N. (2025). Adaptive Metasurface-Assisted Control Protocols for Secure and Stable Millimetre-Wave Indoor Communication Networks. *Transactions on Secure Communication Networks and Protocol Engineering*, *2*(2), 27-36.

David, G., Mdodo, K. L., & Kuma, R. (2025). Context-Sensitive Communication Frameworks for Intelligent Learning Systems. *Transactions on Secure Communication Networks and Protocol Engineering*, *2*(2), 56-64.

Chowdhury, U., & Chakma, S. (2025). Reconfigurable Antenna Design for Wearable Communication Systems. *Journal of Reconfigurable Hardware Architectures and Embedded Systems*, *2*(2), 24-31.

Bosco, K. J. (2025). Low-Latency FIR-Based Signal Conditioning Pipelines for Intelligent Wireless Sensing Platforms. *Journal of Wireless Intelligence and Spectrum Engineering, 2*(2), 8–13.

Ramya, V. (2025). Self-Adaptive Intelligent Learning Environments for Smart and Ubiquitous Spaces. *National Journal of Ubiquitous Computing and Intelligent Environments, 2*(2), 6-14.

Saranya, N. (2025). Secure Event-Triggered Learning Control with Causal State Validation for Distributed Cyber–Physical Energy Platforms. *Transactions on Internet Security, Cloud Services, and Distributed Applications, 2*(2), 24–29.

Kavitha, M. (2025). Data-driven internet weather maps: Real-time analysis of global network performance fluctuations. *Progress in Electronics and Communication Engineering, 2*(2), 54–59.

Kumar, S. TM (2024). Developing FPGA-based accelerators for deep learning in reconfigurable computing systems. *SCCTS Transactions on Reconfigurable Computing*, *1*(1), 1-5.

Thompson, R., & Sonntag, L. (2025). How medical cyber-physical systems are making smart hospitals a reality. *Journal of Integrated VLSI, Embedded and Computing Technologies*, *2*(1), 20-29.

Suresh Kumar, A., & Vaduganathan, D. (2025). Structure-preserving denoising of large-scale satellite imagery using graph signal processing. *National Journal of Signal and Image Processing*, *1*(4), 15–22.

Rajan, C. (2025). Spiking neural network-based modeling of human motor cortex for robotic limb control. *Advances in Cognitive and Neural Studies, 1*(3), 21–28.

Madhanraj. (2024). Simulation of Soil Carbon and Nitrogen Dynamics for Optimizing Fertilization Strategies in Horticultural Crops Using the DNDC Model. *National Journal of Plant Sciences and Smart Horticulture*, *2*(2), 9-17.