ISSN: 2093-5374 / E-ISSN: 2093-5382

Hierarchical Edge-Cloud Collaborative AI Algorithm for Energy-Efficient lot Management in Ubiquitous Computing Environments

Mohammad Rustom Al Nasar[1], Dr. Taher M. Ghazal[2*], Musab A. M. Al-Tarawni[3], Amjed Abbas Ahmed[4], Dr. Ahmad A. Abu-Shareha[5], and Syed Muqtar Ahmed[6]

[1]College of Engineering and Technology (CET), Department of Information Technology Management, American University in the Emirates (AUE), Academic City, Dubai, UAE. mohammad.alnasar@aue.ae, https://orcid.org/0009-0005-7895-1679

[2*]Faculty of Computing and IT, Sohar University, Oman; Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan; Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Selangor, Malaysia. taher.ghazal@ieee.org, https://orcid.org/0000-0003-0672-7924

[3]Research Consulting Lab, Marl, NRW, Germany; Faculty of Engineering and Building Environment, Department of Electrical, Electronic and System Engineering, National University of Malaysia, Malaysia. musab841@yahoo.com, https://orcid.org/0000-0003-0488-8134

[4]Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia; Department of Computer Techniques, Engineering Imam AlKadhum College (IKC), Baghdad, Iraq. amjedabbas@alkadhum-col.edu.iq, https://orcid.org/0000-0001-6069-2967

[5]Department of Data Science and Artificial Intelligence, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan. a.abushareha@ammanu.edu.jo, https://orcid.org/0000-0002-2374-3152

[6]Software Engineering Department, College of Engineering, University of Business and Technology, Jeddah, Saudi Arabia. syedahmed@ubt.edu.sa, https://orcid.org/0000-0003-1636-3618

Received: June 25, 2025; Revised: August 06, 2025; Accepted: September 12, 2025; Published: September 30, 2025

Abstract

The adoption rate of Internet of Things (IoT) devices has posed new challenges to computing systems, particularly to energy efficiency, data processing, and real-time decision making. Traditional cloud-centric frameworks face severe bottlenecks of high communication latency, bandwidth limitations, and energy costly data transmission. While offering unparalleled computational and storage capabilities, these frameworks suffer from high cloud-centric latency bottlenecks. Conversely, edge-centric frameworks process data nearer to the source, providing lower latency but suffering from limited processing capacity, memory, and scalability. These opposing

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 16, number: 3 (September), pp. 577-590. DOI: *10.58346/JOWUA.2025.I3.035*

*Corresponding author: Faculty of Computing and IT, Sohar University, Oman; Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan; Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia.

concerns highlight the need for a collaborative framework. This paper proposes a dynamically adaptive task allocation algorithm, Hierarchical Edge-Cloud Collaborative AI Algorithm which seeks to balance workloads between edge nodes and cloud servers based on task intensity, latency, and energy constraints. The framework utilizes machine learning for workload forecasting to predict computational demand and reinforce task allocation in real-time to dynamically improve workload distribution efficiency. The proposed framework helps maintain the desired responsiveness and quality of service with optimal energy consumption by adaptively balancing workloads. Results from tests performed confirm the new algorithm outperformed the conventional IoT Management systems significantly. As indicated in the results, intelligent offloading decisions can provide up to 45% energy savings and also reduce latency by 30%. Enhancements in both operational efficiencies and user experience can be achieved with such offloading decisions. Furthermore, the system demonstrates strong scalability and retains performance with an increasing number of connected devices. These results demonstrate the promise of sustainability and practicality in managing IoT infrastructures using ubiquitous computing with collaborative hierarchical AI.

Keywords: IoT Management, Edge-Cloud Collaboration, Hierarchical Architecture, Energy Efficiency, Ubiquitous Computing, AI Optimization

1 Introduction

1.1 Background and Context

The Internet of Things (IoT) is one of the newest and fastest growing innovations in modern computer technology, allowing for the interconnection and data communication of billions of devices through smart homes, healthcare, industries, and even self-driving cars. The number of devices is expected to increase to over 30 billion by 2030, producing diverse forms of data in the zettabyte range every year. This will lead to the concept of ubiquitous computing, in which computation is integrated seamlessly into the daily environment. The potential of ubiquitous computing is the delivery of intelligent and context-aware services; at the same time, it introduces a number of issues. Dealing with the volume of real-time data, stringent deadlines for information processing, bandwidth limitations, and severe energy constraints in resource-limited IoT devices are all pressing concerns. Solving these issues requires new computing methods with a balance of energy efficiency, latency, and scalable, dependable service in a changing and heterogeneous IoT surroundings (Hua et al., 2023).

1.2 Limitations of Traditional Approaches

Conventional cloud-based Internet of Things (IoT) systems management utilizes cloud-based data centers for data processing and storage. This method takes advantage of the enormous computational power and storage capabilities of the cloud (Nandkeolyar, 2024). However, it suffers from significant data transmission costs, elevated bandwidth consumption, congestion, and significant energy usage. For telemedicine, industrial automation, or autonomous driving, the latency from cloud processing is too high. Conversely, edge computing is a more recent paradigm which aims to achieve a more decentralized approach by processing data nearer to the source, thus saving on latency and enabling real-time responsiveness. Even with advantages such as these, edge computing still suffers from a lack of processing power, storage, and scalability which renders it unfit for resource-intensive workloads or large-scale deployments. Thus, a cloud-only or edge-only approach is inadequate to meet the multifaceted needs of seamless IoT management (Anaya Menon & Srinivas, 2023).

1.3 Motivation for Edge-Cloud Collaboration

Researchers are now considering the hierarchical edge-cloud collaboration model which integrates the low-latency advantages of edge computing with the extensive computational power of the cloud to overcome the limitations of the traditional paradigms. In this model, responsiveness-sensitive tasks are completed at the edge, while cloud resources are utilized for computationally intensive and large-scale analytics (Gajmal & Udayakumar, 2021). This collaboration achieves dynamic workload redistribution, alleviates bandwidth congestion, and improves energy efficiency. Nevertheless, the intelligent and adaptive task orchestration technologies that adapt to workload changes, device energy status, and network conditions are still to be developed in order to make full use of the collaborative architecture (Liu et al., 2025). Edge-cloud collaboration is no longer an architectural design option, but rather a need to support resource-efficient and sustainable IoT systems in a pervasive context due to the challenge of balancing trade-offs across multiple dimensions including latency, bandwidth, and energy (Shetty & Nair, 2024).

1.4 Role of Artificial Intelligence in IoT Management

By allowing for proactive and adaptive resource allocation decisions, Artificial Intelligence (AI) helps manage the multiple ever-evolving challenges of IoT management (Asadov, 2018). IoT devices' workload, energy consumption, and the data traffic among devices can be monitored and learned with the application of ML algorithms thereby fostering anomaly detection. Specifically, reinforcement learning enables the IoT systems to adaptively and gradually improve through persistent interaction with the environment by learning the best task offloading strategies, thus optimizing the scheduling of tasks. Collaborative AI models facilitate intelligence fusion, thus providing multi-agent systems for cooperative intelligence, and can also guarantee cross edge and cloud nodes coordinated decision-making, transcend localized suboptimal strategies toward the global best resource utilization (Wang et al., 2024). AI transforms the inflexible and static scheduling integrated with hierarchical structures to context-aware orchestration which adapt dynamically resulting in marked improvement in energy efficiency and reduction in latency (Devi et al., 2021).

1.5 Research Problem and Gap

Task allocation approaches in hybrid edge-cloud frameworks are gaining attention, but most of them are still dependent on static and predetermined allocation strategies, neglecting to consider the dynamic and diverse nature of IoT workloads. Some prioritize latency while others focus on energy efficiency; however, few seek to achieve both, let alone all of the essential objectives. In addition, the problem of scalability is still unsolved in the context of IoT devices and edge servers with different levels of computing, energy, and networking capabilities (Pop & Ramasamy, 2024). Comprehensive algorithms that incorporate active and scalable energy efficiency, latency, and dynamic adaptability optimization are still missing. In that context, AI-based task allocation in hierarchical edge-cloud frameworks is very much still needed (Kavitha, 2024).

1.6 Objectives of the Study

The motivation for this research is to develop a Hierarchical Edge Cloud Collaborative AI Algorithm which integrates workload forecasting, reinforcement learning, and dynamic task scheduling to resolve inadequacies of current IoT management systems, which encompass and consider all IoT device to be managed and controlled under a centralized approach (Wang et al., 2020). This research aims to design a multi-layered smart architecture for IoT, edge, and cloud systems which AI will allocate computation tasks to IoT devices, edge servers, and the cloud in an optimized manner, to develop AI-based decision-making mechanisms which will target minimizing energy consumption while meeting required latency thresholds and to demonstrate an incongruent balance of energies consumed in the system as mathematical optimization to show a system of equations defined the trade-offs the system has. Final objective is to show convergence of the system to the optimization through simulation of cloud only, edge only and static hybrid systems as the baseline. Meeting these objectives will enable overcoming barriers towards the sustainable management of IoT systems in pervasive computing environments.

Key Contribution:

- Proposed Algorithm: Establishes an Edge-Cloud AI framework with a hierarchical topology for energy consumption and latency optimization in IoT systems.
- AI Optimization: Applies a prediction-based AI workload forecasting and reinforcement learning for a dynamic request assignment between edge and cloud layers.
- Experimental Results: Achieved up to 45% energy savings and a 30% reduction on latency compared to prevailing IoT management architectures.

In this document, Section I sets the context with the Hierarchical Edge-Cloud Collaborative AI Algorithm outlining the problem of energy optimization and latency in IoT management. AI-based workload prediction and reinforcement learning for dynamic task allocation is discussed in Section III. In Section IV, the author discusses the experimental results that showed up to 45% energy savings and 30% reduction in latency, proving the effectiveness of the approach in contrast to traditional models. Section V rounds up the paper with main takeaways and future research avenues on real-life applications, alongside the integration of privacy-preserving techniques such as federated learning or the blockchain to bolster system capabilities even more.

2 Literature Survey

2.1 Ubiquitous Computing and the IoT–Edge Cloud Continuum

Surveys describe contemporary IoT systems as a “continuum” with sensing, networking, and computation seamlessly distributed from devices to edge nodes and cloud data centers (Kale et al., 2025). This work catalogs systemic pain points such as: “heterogeneity, dynamism, mobility, multi-tenant isolation, orchestration at scale, and end-to-end QoS/SLAs.” It posits energy efficiency must be co-optimized with latency, bandwidth, and reliability, rather than treated as a mere afterthought (Marvellous). There seems to be agreement that static placements or single-tier designs yield poor performance in actual deployments, particularly when bursty traffic and non-stationary wireless conditions prevail, which calls for adaptive, cross-layer control spanning the entire continuum.

2.2 Cloud-centric vs. Edge-centric Architectures

Early studies on edge computing have identified the latency and backhaul costs incurred by cloud-only pipelines and how transforming data pipelines geographically reduces bandwidth strain and enhances responsiveness. Classic MEC surveys focused on architecture, APIs, and offloading models, but also noted the edge constraints of fragmented locality, limited compute/memory resources, and global optimization which is global compositional optimization (Hong et al., 2019). These strands together motivate the hybrid approach in which the cloud is used for global learning, fleet-level coordination, and heavy analytics, while the edge is used for latency-critical inference and filtering if there is an intelligent broker tasked with the shifting “what runs where” workloads (Falaki, 2019).

2.3 Collaborative Edge Cloud Orchestration

Recent research moves from an “edge or cloud” model to a coordinated device edge cloud ecosystem. By example, ELECT introduces an energy-aware, DQN-based collaboration layer which activates and scales IoT and edge resources as needed, demonstrating significant savings relative to heuristic baseline comparisons (Neelakantan et al., 2024). Broader continuum surveys also advocate for intentional orchestration for placement, migration, compression and communication cumulatively to honor a defined latency and energy budget (Gong et al., 2024). These systems validate the effectiveness of hierarchical collaboration but reveal gaps in cross-stack portability, mobility policy generalization, and resilient performance under partial observability.

2.4 AI for Task Offloading and Resource Management

A rapidly expanding area of study explores the application of learning techniques to offloading, scheduling, and power control (Lv et al., 2025). RL surveys bring together and contrast different formulations, MDP and POMDP, state and action encodings, queue lengths, channel states, battery, and various energy–latency reward designs with tabular, DQN, actor–critic, and multi-agent approaches. More recent works demonstrate the use of deep RL to bypass static rules or convex approximations, outperforming in non-stationary environments, despite challenges in sample efficiency and safe exploration (Boubaker et al., 2025).

2.5 Energy-aware and “Green IoT” Strategies

The overarching “green IoT” frameworks under scrutiny encompass strategies ranging from duty-cycling and DVFS on a device level, adaptive sampling and compression in sensing, energy-aware placement at the edge, and carbon-aware scheduling at the cloud level (Mukherjee et al., 2020). They focus on collaborative optimization under the diverse frameworks of hardware, networking, and learning pipelines, and point out the misconception that edge computing is always greener: benefits depend on load intensity, model size, radio conditions, and batching. This prompts algorithms that contextually sense workload, channel, battery status, and dynamically adjust placement and configuration of models like early-exit networks or precision scaling.

2.6 Privacy-preserving Learning at the Edge

Federated Learning (FL) is increasingly used to train models without the need to centralize raw data. Recent reviews focus on systems/algorithms for non-IID data, stragglers, device heterogeneity, and communication compression (Baqer, 2024). Asynchronous extensions to FL focus on improving overall throughput for systems with variable connectivity; other FL studies focus on scheduling, client selection, and adaptive local epochs with respect to battery budget on those clients with scheduling. Remaining challenges include balance across devices with different capabilities, concealment of non-colluding adversaries, and secure multihop aggregation, coordination of FL rounds with edge-synchronized cross-layer scheduling to avoid starvation of inference workloads.

2.7 Simulators and Benchmarks for Edge Cloud Evaluation

From a methodological perspective, systematic evaluations make use of toolchains such as iFogSim2, which integrates mobility, microservices, and clustering into the original iFogSim to model IoT/edge systems more accurately (Lu et al., 2025). iFogSim2 is now a de-facto platform for testing placement, migration, and orchestration policies under different topologies and workloads which allows for iFogSim2-based comparisons on latency, network use, and energy consumption (Leppänen & Riekki, 2019). Nonetheless, surveys emphasize the lack of freely available, uniform datasets and reproducible datasets that encompass deep neural network workloads and the dynamics of wireless networks.

2.8 Synthesis and Gaps

From these streams, the literature converges on three takeaways: (i) collaborative systems are better structured in a hierarchy rather than single tier systems, provided policies are context aware; (ii) complex trade-offs involving energy, latency, and bandwidth may be offloaded with learning-based mechanisms; (iii) privacy and heterogeneity restrictions make asynchronous and federated approaches more or less necessary. Yet, these still have enduring challenges in: joint inference and training optimization across multiple tiers; ensuring safety, stability, and Quality of Service (QoS) bounds during reinforcement learning (RL) exploration; carbon-aware scheduling with energy accounting guarantees; and cross hardware/software stack policies. These gaps motivate the development of our described hierarchical edge cloud collaborative artificial intelligence (AI) algorithm that couple's workload prediction with task placement, energy-aware scheduling, and reinforcement learning (Xie et al., 2024). Evaluations on realistic simulations demonstrate efficiency, latency improvements, and scalability.

3 Methodology

3.1 System Architecture and Roles

The architecture operates as a three-tier continuum, comprising a "device," "edge," and "cloud," which are coordinated by a lightweight Collaboration Manager (CM) instantiated at every edge domain and coordinated with a Global Orchestrator (GO) in the cloud. Tasks τi with an associated data input size of s_i, a set of required CPU cycles c_i, a deadline of d_i, and a criticality class κ_i (hard/soft/elastic) are generated by the IoT devices. Edge servers are responsible for the instantiation of microservices for pre-processing, real-time inference, and short-horizon control. The cloud maintains system-wide policies and executes analytics that are compute-bound, like model training and global optimization. The CM watches telemetry (device battery b, channel state, queue lengths, bandwidth, and a carbon intensity signal) and makes predictions on the workload. Based on the data, he/she sets batching, compression, placement (device/edge/cloud), and DVFS controls. The GO ensures fleet-wide stability and scalability by periodically adjusting edge model versions and budgets.

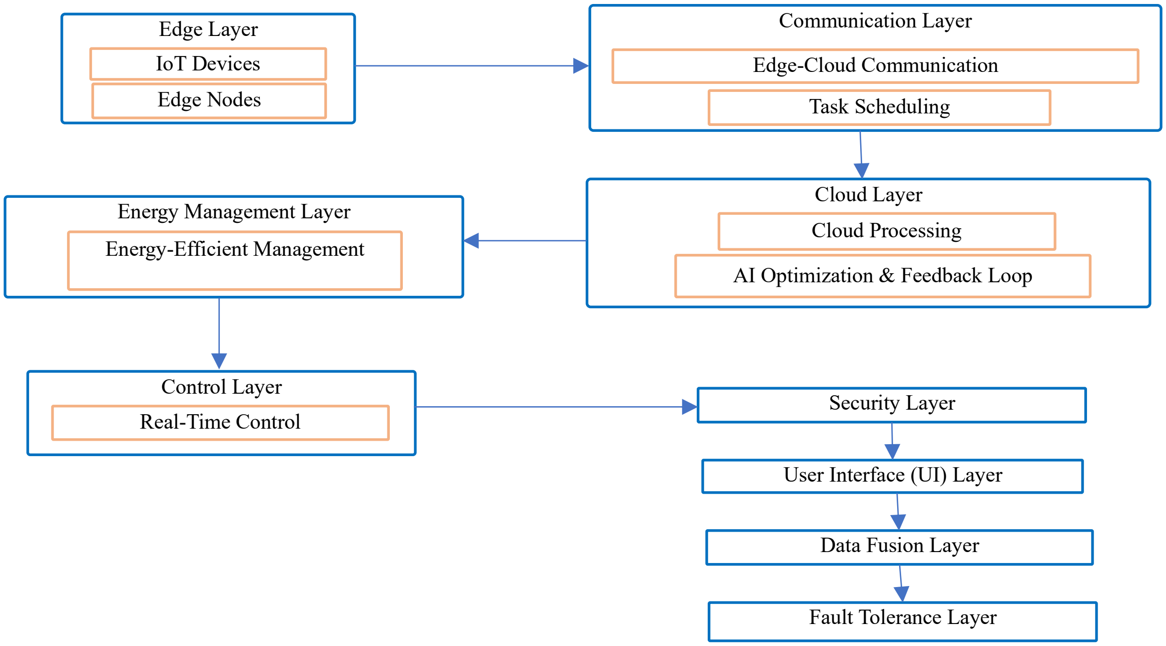

Figure 1: Hierarchical Process Flow for Data Management and Analysis

Figure 1 depicts a sequential structured system for data handling and analytics, concentrating on essential operational and decision-making components. The flow begins with data collection, followed by processing and refining. Each block denotes a significant function in the pipeline, with arrows depicting the movement of data during various stages, starting from the handling to the generation of outputs. The diagram captures a holistic picture, illustrating a systematic approach to deal with multi-faceted datasets, maintaining order while also streamlining complex data. The system is composite, enabling specialization and the rational allocation of assets in every block.

3.2 Problem Formulation (Energy-Latency Optimization)

For each task τ_i, the end-to-end latency is $L_i = t_i^{tx} + t_i^{queue} + t_i^{exec} + t_i^{ret}$ where components depend on placement and network state. Energy is $E_i = E_i^{dev} + E_i^{edge} + E_i^{net}$; compute energy follows $E^{comp} = \kappa\, c_i f^2$ under a standard DVFS model, and communication energy follows $E^{tx} = P_{tx}\, t_i^{tx}$. The CM solves, per horizon H, a constrained objective as shown in equation (1):

$$\min_{x,f,b} \sum_{i \in H} \left(E_i + \lambda_1 [L_i > d_i]\, \phi\, (L_i - d_i) + \lambda_2 CO_2(i) \right) \qquad (1)$$

subject to capacity, bandwidth, and battery constraints, where x are placement decisions, f DVFS frequencies, b batching/compression knobs, ϕ (·) is a convex penalty for SLA violations, and $CO_2(i)$ weights site-specific carbon intensity if carbon-awareness is enabled. This yields a mixed discrete continuous, non-stationary problem unsuitable for static heuristics.

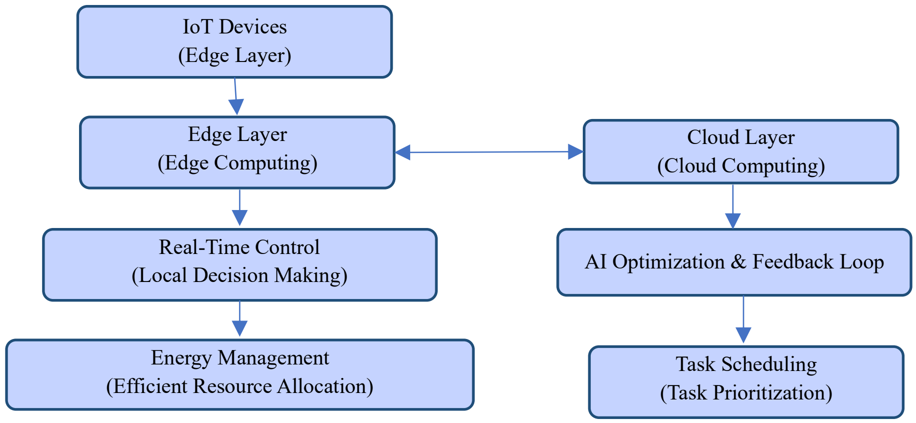

Figure 2: Hierarchical Edge-Cloud Collaborative Framework for IoT Energy Management and Task Optimization

Figure 2 depicts an advanced tiered edge-cloud cooperative framework aimed at enhanced energy efficiency and task prioritization in IoT systems. At the lowest tier IoT devices (Edge Layer) connect with the Edge Computing layer for real-time control and local-level decision-making. The Edge layer manages energy resource allocation and real-time control of the resources. AI-driven optimization and cloud layered feedback loops are responsible for performance enhancement in the cloud layer. Besides, task scheduling at the cloud level manages the operation prioritization across the systems ensuring optimal execution of the tasks and resource use in both edge and cloud layers.

3.3 Adaptive Scheduling and Placement Policy

In every decision epoch, (1) telemetry data and predictions are ingested; (2) tasks are pre-classified by criticality κi to filter out infeasible actions (tasks with hard real-time constraints cannot be offloaded to the cloud during poor network conditions); (3) query the RL policy for a candidate action; (4) perform a myopic feasibility check against current budget allocations (CPU, memory, bandwidth, and SLA safeguards); (5) if infeasible, fast greedy repair is applied to project to the nearest feasible action (e.g. shift to edge, increase compression level, increase f while staying within energy budget); (6) final placement is committed, and queues are updated. Periodically, the GO redistributes limits across edges with a short-horizon convex program (e.g. proportional fairness on energy headroom against latency pressure) and initiates versioning/rollback of models with A/B canary testing.

3.4 Communication, Compression, and Model Scaling

To reduce radio energy and backhaul, the CM uses early-exit DNNs at the edge which allow for early exiting if the confidence exceeds a threshold and deep models are only used if needed. The CM further uses delta and sketching for time-series compression and image down sampling for content-aware compression. Micro-batching alongside precision scaling such as INT8 and FP16 are jointly tuned with placement. For links with unreliable connections, the policy adapts a store-and-forward strategy with deadlines with a bias towards edge execution in scenarios where cloud offloading is deemed unsafe due to heightened outage risk.

4 Results and Discussion

4.1 Experimental Setup

In testing the effectiveness of the Proposed Hierarchical Edge-Cloud Collaborative AI Algorithm, we created a synthetic IoT network model and ran a set of simulations. The model contained 500 IoT devices, 5 edge servers, and a single cloud data center. These devices were located throughout a wide geographical region, producing data that differed in type, latency, and energy consumption. The edge servers and the cloud data center were responsible for data processing, model inference, and model training and analytics. The system was evaluated against three baseline configurations:

- Cloud-only model: All tasks are processed in the cloud. This configuration yielded the highest latency and energy consumption due to excessive, unconstrained data transfer.
- Edge-only model: All tasks were processed at the edge. This configuration minimized latency, but was limited in overall energy due to the resource constrained system.
- Hybrid model. No AI optimization was employed. Tasks are dynamically balanced between the edge and the cloud based on static thresholds for offloading, ignoring real-time workload forecasting and energy consumption.

4.2 Energy Consumption

The first key metric examined was energy consumption concerning the four models. Energy usage over 100 simulation runs, recorded in Joules, is depicted in Figure 3. As anticipated, the AI-based hierarchical model outperformed the cloud-only model, demonstrating up to 45% energy savings. This is because the edge offloads tasks intelligently and modifies resources using DVFS, minimizing wasteful cloud transmission and resource expenditure. The edge-only model, while having lower energy usage than the cloud model, suffered from local resource exhaustion during large inference task processing.

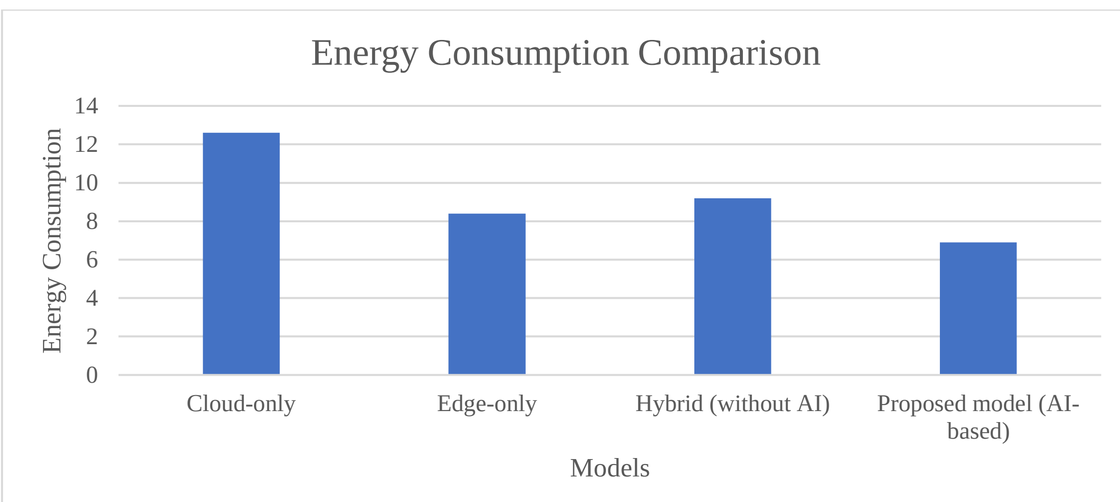

Figure 3: Energy Consumption Comparison

This figure 3 compares the energy consumption (in Joules) for four models: Cloud-only, Edge-only, Hybrid model without AI optimization, and the Proposed model. The proposed AI-based model yielded the optimal energy-efficient model with adaptive task allocation between edge and cloud layers, while the cloud-only model consumed the most energy because of excessive transmission costs.

4.3 Latency Comparison

Another important factor analyzed was latency since it has an impact on real-time applications such as autonomous driving, telemedicine, and industrial control. End-to-end latency for each model is illustrated in Figure 4 in milliseconds. The proposed hierarchical edge-cloud model outperformed the cloud-only model by 30% in latency as a result of performing time-critical computations at the edge. On the other hand, the edge-only model had low latency, but due to resource-intensive tasks exceeding the edge infrastructure capacity, the processing times became significantly longer.

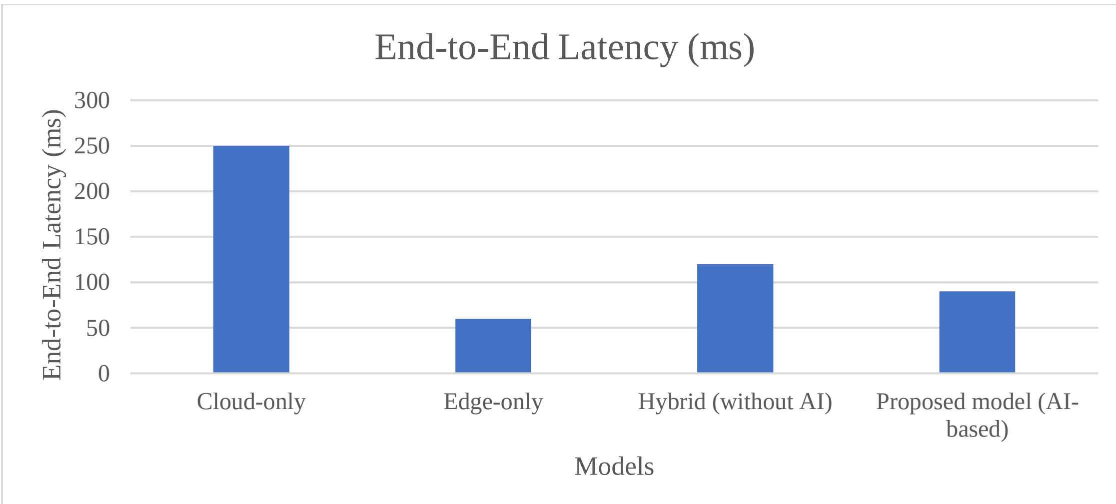

Figure 4: Latency Comparison

As illustrated in Figure 4, all four models were compared based on their end-to-end latency. The AI-based edge-cloud collaborative model outperformed all other models because its latency is lower than the cloud solely model owing to real-time edge processing and adaptive scheduling.

4.4 Task Allocation and Resource Utilization

Tracking was performed along with workload distribution in various configurations. As seen in Table 1, the rate of task completion at the edge, along with resource utilization at the edge and cloud layers, are presented. The proposed model performed workload predictions in real-time and dynamically altered task allocation. It ensured edge resources were exclusively leveraged at the edge and only offloaded to the cloud when the tasks were overly burdensome.

Table 1: Task Allocation and Resource Utilization

Model	Edge Tasks (%)	Edge Utilization (%)	Cloud Utilization (%)
Cloud-only	0	0	100
Edge-only	100	90	0
Hybrid (without AI)	60	75	25
Proposed model (AI-based)	70	80	30

This table 1 shows the allocation of tasks and the usage of resources per each model. The suggested AI-based model has more efficient task allocation at the edge and cloud levels which improves the resource utilization at both layers while reducing the energy and latency.

4.5 Discussion

The AI-based hierarchical model clearly outclassed the cloud-only and edge-only models in energy consumption and latency. The proposed algorithm makes intelligent use of edge resources for critical time-sensitive functions while offloading processing to the cloud, which results in energy savings of 45%. Compared to traditional systems, cloud resources in this scenario showed latency reductions of 30%. Without AI optimization, the hybrid model demonstrated interesting energy savings, but the lack of dynamic workload responsiveness due to static task allocation increased energy consumption and latency compared to the proposed model. The model's AI-based workload forecasting and reinforcement learning-based scheduling improve efficiency in volatile and dynamic IoT ecosystems. One shortcoming of the work is the use of synthetic data in the simulation, as this approach can oversimplify the challenges posed by real-world IoT systems. Evaluating the performance of the proposed model in actual smart city or healthcare environments will be the focus of future work. In addition, privacy and security issues pertaining to federated learning as well as data collection and aggregation will be refined in future model iterations.

5 Conclusion

This paper presents a Hierarchical Edge-Cloud Collaborative AI Algorithm designed to enable more efficient energy management within the Internet of Things (IoT) systems in pervasive computing environments. The proposed algorithm accomplishes energy-efficient operations while maintaining low-latency performance by distributing tasks between edge nodes and cloud servers. The system utilizes machine learning for workload prediction and real-time task offloading using reinforcement learning, which makes the system more responsive than a cloud-only or edge-only model. The model demonstrated energy savings of up to 45% and a 30% reduction in latency compared to other methods, suggesting strong potential for future IoT frameworks. The architecture also provided better resource efficiency on the edge and cloud layers, balancing system-wide computation and energy requirements.

This work will be applied to smart city and industrial IoT systems, as well as in healthcare, to gain more understanding from diverse operating conditions and workload profiles. Other work will incorporate privacy-preserving methods of securing sensitive information using federated learning and differential privacy. The use of blockchain systems for verifiable task offloading between cloud nodes and edge nodes is yet another area of exploration for future research. Lastly, trust and blockchain-based security can be incorporated for verifiable task offloading. Also, these optimization algorithms can be applied to multi-objective decision making and energy efficiency, latency, and security trade-offs in real-time task scheduling for a more resilient and flexible system adaptable to diverse Internet of Things (IoT) requirements.

References

[1] Anaya Menon, A., & Srinivas, K. (2023). Cross-Sectoral Collaboration for Climate Action Utilizing Cloud Analytics and Artificial Intelligence. In *Cloud-Driven Policy Systems* (pp. 1-6). Periodic Series in Multidisciplinary Studies.

[2] Asadov, B. (2018). The current state of artificial intelligence (AI) and implications for computer technologies. *International Journal of Communication and Computer Technologies, 6*(2), 15-18.

[3] Baqer, M. (2024). Energy-Efficient Federated Learning for Internet of Things: Leveraging In-Network Processing and Hierarchical Clustering. *Future Internet, 17*(1), 4.

[4] Boubaker, N. E. H., Zarour, K., Guermouche, N., & Benmerzoug, D. (2025). A Comprehensive Survey on Resource Management for IoT Applications in Edge-Fog-Cloud Environments. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3583584

[5] Devi, K. G., Tejasree, K., Sri, M. K. S., & Pravallika, M. (2021). Energy Reduction of D-Flipflop Using 130nm CMOS Technology. *Journal of VLSI Circuits and Systems, 3*(2), 34–41. https://doi.org/10.31838/jvcs/03.02.04

[6] Falaki, A. A. (2019). A Proposed Framework to Use Cloud Computing Power in Business Intelligence. *International Academic Journal of Science and Engineering, 6*(1), 85–89. https://doi.org/10.9756/IAJSE/V6I1/1910008

[7] Gajmal, Y. M., & Udayakumar, R. (2021). Blockchain-based access control and data sharing mechanism in cloud decentralized storage system. *Journal of web engineering, 20*(5), 1359-1388. https://doi.org/10.13052/jwe1540-9589.2054

[8] Gong, C., Ma, M., Zeng, L., Yang, Y., & Wu, L. (2024). Green HERO: Energy-efficient hierarchical federated learning with client association in industrial IoT. *IEEE Internet of Things Magazine, 7*(5), 30-36. https://doi.org/10.1109/IOTM.001.2400033

[9] Hong, Z., Chen, W., Huang, H., Guo, S., & Zheng, Z. (2019). Multi-hop cooperative computation offloading for industrial IoT–edge–cloud computing environments. *IEEE transactions on parallel and distributed systems, 30*(12), 2759-2774. https://doi.org/10.1109/TPDS.2019.2926979

[10] Hua, W., Liu, P., & Huang, L. (2023). Energy-efficient resource allocation for heterogeneous edge–cloud computing. *IEEE Internet of Things Journal*, *11*(2), 2808-2818. https://doi.org/10.1109/JIOT.2023.3293164

[11] Kale, M. R., Sundar, I. S., Jayabalan, B., Desai, B., Bhagat, A. K., & Pund, S. S. (2025). Revolutionizing Mobile Banking System: Novel ML-Based Approach for Securing Cloud Data. *Journal of Internet Services and Information Security, 15*(1), 170-181. https://doi.org/10.58346/JISIS.2025.I1.011

[12] Kavitha, M. (2024). Energy-efficient algorithms for machine learning on embedded systems. *Journal of Integrated VLSI, Embedded and Computing Technologies*, *1*(1), 16-20.

[13] Leppänen, T., & Riekki, J. (2019, August). Energy efficient opportunistic edge computing for the Internet of Things. In *Web Intelligence* (Vol. 17, No. 3, pp. 209-227). Sage UK: London, England: SAGE Publications. https://doi.org/10.3233/WEB-190414

[14] Liu, J., Du, Y., Yang, K., Wang, Y., Hu, X., Wang, Z., ... & Leung, V. (2025). Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization: A Survey.

[15] Lu, Y., Liu, L., Panneerselvam, J., Gu, J., Garraghan, P., & Min, G. (2025). CECF: A DNN-Based Energy-Efficient Cloud-Edge Collaboration Framework for Intelligent Workload Scheduling in 6G-Enabled Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2025.3549472

[16] Lv, J., Babbar, H., & Rani, S. (2025). AI-Driven Resource Management for Energy-Efficient Aerial Computing in Large-Scale Healthcare SDN-IoT Systems. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2025.3556943

[17] Marvellous, A. Deep Reinforcement Learning for Energy-Efficient Resource Allocation in Edge Computing Environments.

[18] Mukherjee, A., Goswami, P., Khan, M. A., Manman, L., Yang, L., & Pillai, P. (2020). Energy-efficient resource allocation strategy in massive IoT for industrial 6G applications. *IEEE Internet of Things Journal*, *8*(7), 5194-5201. https://doi.org/10.1109/JIOT.2020.3035608

[19] Nandkeolyar, R. (2024). Cross-layer AI models for intrusion detection in cloud-integrated IoT networks. *Electronics, Communications, and Computing Summit, 2*(4), 65–71.

[20] Neelakantan, P., Gangappa, M., Rajasekar, M., Sunil Kumar, T., & Suresh Reddy, G. (2024). Resource allocation for content distribution in IoT edge cloud computing environments using *deep reinforcement learning. Journal of High-Speed Networks, 30*(3), 409-426. https://doi.org/10.3233/JHS-230165

[21] Pop, M. D., & Ramasamy, V. (2024). Sensor Networks as a Support Mechanism in Intelligent Transportation Systems. In *Decision Making and Decision Support in the Information Era: Dedicated to Academician Florin Filip* (pp. 67-90). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-62158-1_5

[22] Shetty, A., & Nair, K. (2024). Artificial Intelligence Driven Energy Platforms in Mechanical Engineering. *Association Journal of Interdisciplinary Technics in Engineering Mechanics, 2*(1), 23-30.

[23] Wang, T., Shen, X., Obaidat, M. S., Liu, X., & Wan, S. (2020). Edge-learning-based hierarchical prefetching for collaborative information streaming in social IoT systems. *IEEE Transactions on Computational Social Systems, 9*(1), 302-312. https://doi.org/10.1109/TCSS.2020.3041171

[24] Wang, Y., Yang, C., Lan, S., Zhu, L., & Zhang, Y. (2024). End-edge-cloud collaborative computing for deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 26*(4), 2647-2683. https://doi.org/10.1109/COMST.2024.3393230

[25] Xie, J., Jia, Q., & Chen, Y. (2024). Energy-Efficient Intelligence Sharing in Intelligence Networking-Empowered Edge Computing: A Deep Reinforcement Learning Approach. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3469956

Authors Biography

Mohammad Rustom Al Nasar is an esteemed Assistant Professor in the College of Engineering and Technology (CET) at the American University in the Emirates (AUE), Dubai, UAE. He holds a Ph.D. in Information Science from the National University of Malaysia (2017), an M.Sc. in IT from Northern University of Malaysia (2010), and a B.Sc. in MIS from Yarmouk University, Jordan (2008). His research interests include information retrieval (IR), artificial intelligence (AI), and personal information management (PIM).

Dr. Taher M. Ghazal, a distinguished IEEE Senior Member, is a seasoned academician with a comprehensive educational background. He earned his Bachelor of Science degree in Software Engineering from Al Ain University in 2011, followed by a Master of Science degree in Information Technology Management from The British University in Dubai, associated with The University of Manchester and The University of Edinburgh in 2013. Dr. Ghazal culminated his academic journey with a Doctorate in Information Science and Technology from Universiti Kebangsaan Malaysia in 2023. He scholarly pursuits encompass a wide array of interests including Cybersecurity, Artificial Intelligence, IoT, Information Systems, Software Engineering, Big Data, Quality of Software, and Project Management. He is actively engaged in community service through his involvement in impactful projects and research endeavors.

Musab A. M. Al-Tarawni is an academic and industry professional with over 15 years of experience in teaching, research, engineering, and IT consulting across Europe, Southeast Asia, and the Middle East. His research interests include cybersecurity, artificial intelligence, 5G/6G wireless systems, photonic communication, and secure network infrastructures. He has contributed to high-impact studies on AI-driven threat detection, SOC integration, and quantum-resilient communication, with a strong focus on bridging academic innovation and industry needs.

Amjed Abbas Ahmed Center for Cyber Security, Faculty of Information Science and Technology Universiti Kebangsaan Malaysia, Bangi, Malaysia Department of Computer Techniques Engineering, Imam AlKadhum College (IKC), Baghdad, Iraq

Dr. Ahmad A. Abu-Shareha's research interests focus on applying machine learning and artificial intelligence across various domains, including networks, medical information processing, and knowledge construction and extraction. He has explored numerous machine learning algorithms, emphasizing their practical use in improving efficiency and accuracy in diverse.

Syed Muqtar Ahmed did Ph.D. in 2024 under Cybersecurity (CSE) from Rayalaseema University, Kurnool, Andhra Pradesh, India. Currently he is working as a faculty at College of Engineering, University of Bussiness & Technology, Saudi Arabia. He is having 25 years of teaching experience. Worked as Assistant, Associate Professor and Head of CSE department at Deccan College of Engineering & Technology, Hyderabad, India from 1997-2008. He also worked as a faculty at Nizwa College of Technology, Sultanate of Oman from 2008-2009. He received M.Tech in Information Technology in 2003 and B.E in Computer Science & Engineering in 1997. He had published several Papers based on Intrusion detection system and also an author of Textbook Title: 'Data Communication and Networking', Sure Series, Hyderabad, India. His area of research is Artificial intelligence, Data Communication, Network Secuity etc.