Vol. 10, Issue: 2026

Article ISSN: 2312-2668

International Journal of Information Sciences Management (IJIMS)

# Edge-AI Solutions For Low-Energy Iot Networks In Smart Cities

[1] Musab Umair Malik
Bilal Bin Ameer



[1] Corresponding Author

# Edge-AI Solutions For Low-Energy Iot Networks In Smart Cities

**Musab Umair Malik**
Alinma Bank Riyadh, Saudi Arabia
Musabumair85@outlook.com

**Bilal Bin Ameer**
Alinma Bank (Riyadh, Saudi Arabia)
Bilalbinameer@outlook.com

**Abstract:**
The exponential growth of the Internet of Things (IoT) systems in smart cities has exacerbated the need of energy-saving, low-latency, and privacy-preserving Artificial Intelligence (AI) applications. In this study, the authors have investigated the performance of three AI deployment paradigms on a simulated low-energy IoT network Cloud-based AI, Edge-AI, and Federated Edge-AI. Four key performance measures were tested with the computation of the energy consumption, latency, bandwidth use, and inference accuracy using quantitative experimentation with fifty IoT nodes within equal computational conditions. The findings have proved that Cloud-based AI is the most accurate (96.9%), but it has high energy and latency overheads since data is processed in a central location. By comparison, Edge-AI consumes 52% less energy and latency with 73% less compared to localized computation efficiency. The Federated Edge-AI paradigm is the best balanced to provide a reduction of 66 percent in energy consumption and 25 percent in bandwidth efficiency, at a small accuracy cost (95.0 percent) relative to the cloud. The other normalized Composite Performance Index merely proves the excellence of Federated Edge-AI (CPI = 0.86) when compared to Edge (0.66) and Cloud (0.25) architectures. The paper summarizes the fact that federated learning and edge computing is a sustainable, scalable, and privacy-preserving architecture of the next-generation smart city IoT ecosystems. Such hybrid teams are capable of supporting near-cloud intelligence and delivering much higher network resilience and resource efficiency and is a paradigm shift of decentralized AI systems.

**Keywords:** Edge Computing; Internet of Things; Artificial Intelligence; Smart Cities; Distributed Systems

**Received:**
October 21, 2025
**Review Process:**
December 2, 2025
**Accepted:**
January 5, 2025
**Available Online:**
January 11, 2026

## Introduction:

The high level of urbanization and digitalization has produced an overwhelming demand to attract smart infrastructure able to manage resources effectively and increase the quality of life and be sustainable (Goel and Vishnoi, 2022). The most trending paradigm of such change is smart cities, and the networks and technologies that have been utilized to deliver responsive and data-oriented city services are networked technologies, such as the Internet of Things (IoT), artificial intelligence, and better communication networks (Alahi et al., 2023). Smart traffic control, garbage, environmental, and energy distribution applications can be discussed as examples of applications that are based on interrelated IoT sensors that constantly gather data and send it to the analysis unit to make decisions (Bellini et al., 2022). Nevertheless, the large size and diversity of such IoT

implementations present severe issues related to do with energy usage, latency, and information handling. Specifically, the resources available to most IoT nodes in smart cities are extremely constrained (i.e., limited battery life, low computing capabilities, and bandwidth limitations), and this restricts the scalability and sustainability of traditional cloud-centric designs ( Trigka and Dritsas, 2025).

Conventionally, the data obtained by the IoT devices is sent to centralized cloud servers to be processed and inferred. Though cloud computing offers great computational resources and storage capabilities, it is associated with huge communication overhead, latency, and energy consumption (Al-Jumaili et al., 2023). Constant relay of raw sensor data consumes battery-operated devices, causes network congestion, and raises the question of privacy because of the centralized nature of handling sensitive data of the city (Kanellopoulos et al., 2023). As a result, an increasing academic and industrial consensus is that the future generation of smart city architectures needs to take intelligence nearer to data-generation locations. This paradigm of distributed intelligence, also known as Edge Artificial Intelligence (Edge-AI), puts machine learning (ML) and AI directly into IoT devices or intermediate edge nodes and allows real-time analytics, decision-making, and eliminates reliance on cloud connectivity (Gill et al., 2025).

Edge-AI is a combination of embedded systems, edge computing, and AI. It allows processing and interpretation of data at the end of the network, close to the source of data, rather than sending it to a remote server (Garcia-Perez et al., 2023). This architectural change has a very bright future in the smart city ecosystem: it is going to reduce latency, improve privacy, use less network bandwidth, and, most importantly, become more energy-efficient. In edge-based designs, transmission of raw data is substituted with either local inference or event transmission, meaning, which lowers the power consumption of the radio transmission process, which is commonly a significant portion of the overall energy budget of IoT devices (Xu et al., 2022).

Moreover, the incorporation of AI into energy-limited IoTs presents new technical issues. Machine learning models that are common are computationally and memory-intensive, and they can be inaccessible with low-power microcontrollers or sensors (Xu et al., 2025). Consequently, due to the low-energy IoT networks, the Edge-AI application must optimize the model and system architecture closely. Such methods as Tiny Machine Learning (TinyML), model compression, quantization, pruning, and knowledge distillation have become viable solutions to run AI models on narrow hardware and energy budgets. All of these approaches make small neural networks runnable with kilobytes of memory on microcontrollers, allowing tasks like anomaly detection or environmental pattern recognition, or classifying acoustic events directly at the device level (Lamirini et al., 2023).

Although TinyML has made it possible to run intelligence on the device, scaling and continuity in learning are still concerns for smart city networks where devices distributed should be able to cooperate and learn together (Ahmed et al., 2024). In this regard, Federated Learning (FL) as a decentralized training model enabling multiple edge devices to jointly train shared models without sharing raw data, has been recognized as a viable solution. FL reduces the cost of communication through the sharing of model updates rather than raw data and maximizes privacy through localization of the data (Beltran et al.,2023). However, even naive networks of FL can still be energy-inefficient because of periodic communication rounds and unnecessary involvement of low-energy nodes. Hence, smart city applications require energy-aware and hierarchical federated learning, which makes model updates be scheduled selectively by considering energy levels, the quality of links, and the significance of contributions (Dang et al., 2024).

Edge-AI systems do not only focus on the local computation of energy efficiency, but also on its optimization at the network level. Smart cities are typically constructed based on low-power wide-area network (LPWAN) systems, including LoRaWAN, NB-IoT, or Sigfox, which is able to facilitate long-range communication with only a small amount of power usage (Pooyandeh and Sohn, 2021). Nevertheless, the transmission and reception occurrences dominate the energy usage in LPWANs, and the minimization of communication is a key design concern. The IoT systems need to consider event-based communication models, event-driven sampling, and context-based transmission policies so that data are only transmitted when meaningful or when anomalies have taken place in order to be operated at low energies. Together with Edge-AI processing, the methods can significantly decrease network load, and the information quality can be maintained (Rajab et al., 2023).

System architecture is one more important dimension. Architectures that are effectively deployed using Edge-AI solutions use the three-layer framework, with the device layer consisting of energy-constrained IoT nodes with TinyML inference models; the edge or aggregator layer, which executes intermediate process, fusion, and coordination functions of distributed learning processes, and the cloud layer, which executes heavy training, data archival, and global model maintenance (Zheng et al., 2025). These hierarchical designs enable localized intelligence and yet have centralized control, a tradeoff between the latency, energy usage, and model accuracy. Moreover, the adaptive task offloading mechanisms can make dynamically based decisions on whether a task will be implemented at a lower tier or in a higher tier, depending on the current states of energy, workload, and network (Letaief et al., 2021).

The energy-conservative aspect of the smart city integration of AI and IoT has far-reaching implications on society and the environment in general. Smart city infrastructures are projected to maximize the use of energy in the city, decrease emissions, and increase the efficiency of civic services (Raj and Shetty, 2024). Nevertheless, ironically, the energy footprint of the IoT and AI systems themselves has become a sustainability issue. Edge-AI is one of the potential solutions to eliminate such a paradox through minimizing unnecessary data transfers, prolonging the existence of the device, and decreasing the carbon footprint of cloud processing (Philipo et al., 2025). Thus, it is not just a technical requirement but also an environmental one to optimize Edge-AI to work with low energy consumption and improve the sustainability of the world in terms of global sustainability goals. Although a lot has been achieved in terms of research, there are a number of gaps in the literature (Surianarayanan et al., 2023). Most of the current literature on Edge-AI emphasizes performance measures like accuracy of inference and latency, and does not carry out a detailed energy characterization as well as long-term sustainability considerations. In the same way, the majority of federated learning studies presuppose sufficient amounts of computation and energy, which do not take into account the limitations of microcontroller-based IoT nodes that are widespread in smart cities (Bourechak et al., 2023). Moreover, communication-layer optimization and AI model optimization have yet to be integrated; there are very few frameworks that optimize the energy of computation and communication on a system level in a holistic manner. It is through this fragmentation that machine learning optimization, communication protocols, and energy management policies should be brought together as a consolidated cross-layer solution to be outlined in a shared design philosophy (Rajput and Yadav, 2025).

This work discusses the design and implementation of Edge-AI solutions to low-energy IoT-based smart cities. The essence of the task is to create an integrated architecture and approach that would coordinate the optimality of AI model implementation, inter-communication, and device power consumption.

The suggested framework highlights three mutually supporting strategies, namely: (1) using TinyML models to detect and compress local events and (2) using hierarchical and energy-aware federated learning to scale model adaptation and (3) using adaptive communication and offloading policies to trade-off accuracy and energy efficiency. These mechanisms will also help show that there is an opportunity to substantially increase the life of IoT devices and networks and preserve the presence of intelligent functionality and decent performance.

**Background Studies/Literature Review:**

The high pace of development of edge computing has essentially changed the manner in which data are handled in the distributed IoT setup. The first model of IoT was mostly cloud-based, with the application of raw data produced by sensors and sent to centralized servers where raw data was stored and analyzed (Hartmann et al., 2022). Although it allowed to perform sophisticated computation, this model added high network latency, high bandwidth consumption, and susceptibility to communication failures (Ji et al., 2023). These problems were reduced with the development of edge computing, where the computation was moved nearer to the source of data. Edges servers, which are deployed at gateways, base stations, or micro-data centers, in the context of smart cities, conduct pre-processing and feature extraction and localized inference, which decreases the reliance on cloud infrastructures (Xu et al., 2023).

Based on this paradigm, Edge-AI includes artificial intelligence features into the edge layer to aid autonomous decision-making. The architectures of edge-AI are based on the distributed intelligence in device, edge, and cloud levels in order to address the strict demands of latency, energy consumption, and the privacy (Duan et al., 2022). The architectures are specifically applicable to smart cities, where timeliness is essential to the applications of traffic light optimization, detecting environmental anomalies, and emergency responses. According to recent research, AI inference on edges has the potential to reduce the response time by up to 70% against cloud-based analytics and also reduce communication energy by over 60% in the typical IoT deployment (Xu et al., 2023). However, AI models cannot be easily integrated into edge nodes and IoT nodes because of limited computing resources, memory, and energy resources. The latter has prompted a dedicated research area called Tiny Machine Learning (TinyML), the objective of which is to create microcontrollers- and battery-powered devices-compatible ultra-lightweight machine-learning models (Hua et al., 2023).

One of the most promising technologies that can be employed to possess intelligent behavior at the extreme end of the IoT networks is tinyML. The concept is to scale the ML models, in this case neural networks, to be executed on a few hundred kilobytes of memory and milliwatts of power. Several methods that are widely utilized in order to achieve these objectives are pruning models, quantization, knowledge distillation, and architecture search (Liu et al., 2024). More recent frameworks like Tensor Flow Lite Micro, Edge Impulse and microTVM allow a developer to execute AI inference on a microcontroller with no external dependencies. One can mention the example of a 20-kilobyted convolutional neural network presented by Liu et al. (2024) and capable of recognizing keywords with less than 1 mW of power draw on an ARM Cortex-M4 chip. Similarly, Wong et al. (2024) proposed the adoption of quantization-aware training, the weight clustering, and they used it to implement the environmental anomaly detectors in the LoRa-enabled nodes, which consumed more than 75% of the energy in the case of continuous data streaming.

TinyML has started to pay attention to energy-adjustable behavior, where models can become more or less complex based on some element of the accessible energy or the urgency of the task (Swamy, 2024). Partial termination of inference

at the occurrence of intermediate confidence thresholds can be performed using early-exit networks, and with a less computational cost at tolerable accuracy. These types of designs are perfectly in line with the sustainability goals of the smart city IoT systems, where the lifespan of sensors directly affects the cost and maintenance periods of operations (Belli et al., 2020). In addition to these developments, the majority of TinyML deployments are static, and are not capable of learning or adapting post deployment. This is a limitation to scalability in dynamic settings e.g. urban traffic systems or air-quality monitoring where contextual patterns change with time (Schizas et al., 2022). Therefore, researchers have resorted to federated learning and collaborative inference mechanisms to allow continuous learning over distributed resource-constrained devices.

Federated Learning (FL) enables two or more devices to jointly train a common model without the need to transmit raw data to the central server (Abreha et al., 2022). The nodes make local model updates using their own private dataset, and communicates them to an aggregator when such updates, typically gradients or changes in weights, are available to the aggregator to gather global model updates. This method maintains the data privacy and in comparison to centralized learning, this method consumes a lot less information than the centralized one (Yang et al., 2022). FL is especially useful in the intelligent city domain in areas that require sensitive or distributed based data like in health care monitoring, mobility analytics, or video monitoring. However, traditional FL algorithms are comparatively strong members and unchangeable connectivity, which is quite unlikely to be achieved by IoT gadgets with low energy needs. Unbalanced distribution of data, convergence instability, and heterogeneous power usage are some of the issues that cause premature depletion of the devices (Khan et al., 2025).

The literature suggests a number of energy-conscious FL approaches to solve these problems. Qin et al. (2024) propose a new selection of participants, called adaptive, in terms of residual battery energy, quality of links, and significance of contributions, which allows extending network life without limiting learning performance. Additional works investigate update compression (e.g., sparse or quantized gradients), asynchronous aggregation, and hierarchical federated learning (HFL) structures, in which the middle edge nodes do local model fusion, followed by synchronization with the cloud (Anagnostopoulos et al., 2024). This type of hierarchical models is especially applicable to deployments at city-scale to facilitate learning at cluster level and minimize long-haul transmissions. Apart from these developments, there are still two basic gaps. First, the vast majority of FL implementations do not optimize computation and communication energy jointly and only reduce the update frequency. Second, there is not much integration between TinyML inference at the end devices and model adaptation at higher levels through FL (Shahid et al., 2021). The opportunity to bridge these gaps by combining cross-layer optimization is also a major area of research that the present study endeavors to fill.

The energy use of the IoT networks is based on the communication protocols or hardware design. Many studies have shown that the use of radio transmission and reception is often the largest contributor to overall energy use and up to 60-80% of a node lifetime consumption (Fay et al., 2023). Therefore, Edge-AI deployments cannot do without energy-efficient networking plans. LoRaWAN, NB-IoT, and Sigfox are Low-Power Wide-Area Networks (LPWANs) which have become widely used in smart cities since they have a long range and low power profile. Nevertheless, LPWANs have disadvantages of low data rates and tough duty-cycle controls. In the study by Fay et al. (2023), the energy model of LoRaWAN devices was presented in a detailed manner, demonstrating that message retransmissions and large payloads significantly decrease the battery life. They can thus increase

energy efficiency by reducing communication events either with the use of local inference or adaptive sampling or with the use of compressed model updates-principles of Edge-AI.

Further development of research incorporates adaptive control of communication whereby transmission parameters (spreading factor, power level or interval) are dynamically changed subject to network conditions or model-inference confidence. Indicatively, Ahmed et al. (2025) suggested an AI-guided medium-access control protocol, which learns to have the best timing of transmission to compromise between energy and latency. All these innovations point to the fact that it is the synergy between AI and layers of communication, and not individual optimization that is the key to sustainable operation of the IoT. A number of architectural frameworks are proposed to integrate a variety of technologies that drive smart cities. The most common one is the multi-tier architecture that separates the computation and intelligence between the device, the edge, and the cloud layers (Du et al., 2022). On the device level, TinyML models can be used to process preliminary data information and event detection, the edge level process aggregation, local training, and coordination, and the cloud level carries out complex analytics and global coordination. These are architectures that are scalable, responsive and energy-efficient.

Recent applications focus on the offloading of tasks policy, which states that computational tasks are offloaded dynamically through the different levels based on energy state, availability of bandwidth, and latency factors. For example, Yang et al. (2022) suggested an energy-conscious task-offloading strategy to vehicular IoT networks, which improved the device lifetime by 40 per cent by means of adaptive edge selection. In the same manner, Shen et al. (2025) came up with a reinforcement-learning-based offloading strategy which optimizes energy and accuracy together through predicting real-time changes in workload. However, the implementation of such structures in the real-life smart cities is not a trivial one. Heterogeneity of devices, security vulnerabilities, and complexity of scaling of large-scale Edge-AI networks are some of the challenges (Gill et al., 2025). Moreover, there are not many studies that can give standardized benchmarks or holistic energy models that include AI inference and network communication. The literature thus proposes the need of experimental and simulation based approaches that can measure end-to-end energy performance in heterogeneous smart city setups.The literature reviewed confirms that Edge-AI, TinyML, and federated learning have made a considerable step toward intelligence decentralization and reduced latency in IoT systems in Smart cities. However, a number of gaps in the research still exist:

Fragmented Optimization: The current literature considers optimization of AI models and network energy management as two separate issues instead of combining them in a unified framework.

Limited Adaptivity: Not many architectures combine real time energy sensing with adaptive model selection, communication schedule or offloading.

Absence of Empirical Characterisation of Energy: There is little quantitative characterisation of the total system energy; i.e. sensing, computation and communication, so it is hard to assess the real sustainability benefits.

Scalability Issues: Hierarchical learning and management systems with the capability of running on thousands of heterogeneous devices are not well studied.

Lack of Standardized Evaluation Metrics: In most studies, the benchmarks used are isolated; there is still no single metric of measuring energy efficiency, latency, and inference accuracy.

To address these constraints, there is a need to consider the cross-layer and energy-aware Edge-AI architecture, where TinyML inference, federated learning, and adaptive communications are balanced to be built in one design framework. The system should provide

balance between intelligence, scalability and sustainability to response to the operation and environmental requirement of smart cities in future.

**Research Design:**

The study has a mixed-method research design that integrates analysis modeling and simulation experimentation to investigate how Edge-AI architectures can be utilized to enhance energy efficiency in IoT networks in smart cities. The paradigm the study is grounded in is a design-science, which is aimed at the design, deployment, and testing of a new Edge-AI architecture on energy consumption and computation. The involved steps include conceptualizing an edge-based architecture which has artificial intelligence added to the network periphery. This is now followed by lightweight AI algorithms and energy efficient communication protocols. The last step is the empirical testing and validation of the performance by the use of simulation, where it should be ensured that the framework is capable of functioning under the realistic conditions in the smart city.

**Proposed Framework**

The suggested Edge-AI system is designed as a three-level distributed architecture, which includes the IoT device layer, the edge intelligence layer, and the cloud coordination layer. All layers have their unique but inter-dependent tasks in the system in the management of energy, computation, and communication.

The IoT device layer includes low powered sensors, actuators and microcontrollers interwoven in the urban infrastructures like traffic lights, environmental stations, and police surveillance units. These devices produce huge volumes of environmental, transportation, and utility data, which is processed locally to a small degree because of energy limitations and computing limitations. The IoT devices perform simple processes like data sampling, compression, and preliminary features extraction instead of sending raw data continuously to the cloud and thus save on bandwidth and energy. The layer of edge intelligence is the computation layer between the IoT nodes and the cloud. It is comprised of edge servers or gateways that have adequate processing power to run machine learning models on a real time basis. The edge layer is where localized inference and data aggregation are done, and tailored decisions are made, so that the system can react quickly to contextual environments, i.e. traffic congestion or air quality alerts. Edge nodes use lightweight neural networks, reinforcement learning agents or federated learning schemes to make sure that they make decisions near the source of the data to minimise latency and communication expenses.

The cloud coordination layer offers the world wide supervision and long term intelligence. It scales and consolidates metadata and periodically retrains AI models based on large-scale data gathered spanning across multiple edge domains. New model parameters are then re-distributed to edge nodes in a federated learning cycle to keep local models moving towards a steady improvement with no need of raw data transfer. This hierarchical model is a successful distribution of computational loads and coherence between the global and local intelligence, leading to a higher scalability of the smart city ecosystem and energy sustainability.

**Simulation setup and data collection**

Data collection will be performed using existing figures (those in real life). The study applies an approach based on the use of simulation with the help of a smart city prototype environment which supports the real-world dynamics of IoT communication, computation and power consumption. The network modeling component of the simulation is modeled on Network Simulator 3 (NS-3) and the AI inference models are implemented with the help of TensorFlow Lite on the edge. The synthetic data is created in a virtual smart city grid that has a geographic area of about four square kilometers. This grid has 500- 2000 IoT devices randomly

attached to it with each device programmed to pump streaming flows of environmental and activity data always. These sensors simulate numerous kinds of data and constitute a variety of environmental pointers, including temperature, humidity, and particulate matter, and transport pointers, including vehicle density and movement patterns. In addition, the IoT cameras simulate their nodes to detect events when there is a public safety scenario.

The communication technology follows the low-power wireless communications such as IEEE 802.15.4 and LoRaWAN, which resembles a real-life application of smart cities. The functionality of all IoT devices is limited by a tiny battery capacity, and the model of energy consumption is approximated by the realistic measures of power discharge to offer an evaluation of the feasibility of the proposed system. Permanent edge servers are installed all across the network, which serve as local network aggregates and AI computing devices. Different densities and workloads of the network are used to run the simulation to check the scalability and flexibility of the framework. Major parameters that are adjusted systematically are the frequency of data transmission, the mobility of nodes, and the range of communication to ensure that the system is tested under different conditions in the urban environment.

**AI Model Implementation**

One of the main themes of this study is the implementation of energy-efficient AI models capable of executing on edge nodes (resource-constrained). To realize the same, the models are rigorously optimized by pruning models, quantizing models, and knowledge distillation. Based on these processes, the number of trainable parameters, the size of memory and the inference latency are significantly reduced without compromising predictive accuracy. The first models of AI are mostly grounded on Convolutional Neural Networks (CNNs) to identify patterns and Decision Tree classifiers to make lightweight predictions of events. The training is initiated on the cloud where large aggregated datasets are used after which optimized models are sent out to edge nodes. Federated learning is used to maintain privacy and minimize the overhead in communication: the edge nodes update their local model weights on the local data and only send the learned weights to the cloud. The cloud subsequently carries out a federated averaging to create a global model which is re-distributed to the edges. Such a process will result in constant learning as well as in preventing the need of centralizing sensitive information.

Similarly, the structure has adaptive sampling mechanisms that dynamically change the sensing rates in response to environmental variability. Using air quality measurements as an example, there are no changes in air quality over some period of time, the system will automatically reduce sampling rate to conserve energy. These adaptive algorithms of reinforcement learning policy offer the IoT nodes control energy consumption on their own, along with information integrity.

**Performance Metrics**

To assess the efficacy of the proposed framework, the analysis will be done by a set of quantitative measures of analysis in terms of energy consumption, latency, accuracy, bandwidth consumption, and computational overhead. Energy use is measured in millijoules per inference cycle or data transmission cycle and this directly provides a clue on the efficiency of the system. The duration of time to happen in data generation at the sensor and actionable output at the edge or cloud layer, can be defined as latency, which is an indicator of responsiveness in the system as operations proceed in real-time.

The percentage of the correct predictions of the AI models that can be compared to assess the centralized and distributed approaches is the inference accuracy. The bandwidth used is an indicator of all the data traffic within the network since this represents a reduction in the load on the communication due to the edge processing. Finally, the cost of updating the edge

models with the cloud at federal learning period is known as model update overhead. By integrating these measures, the research article creates an extensive picture of performance, comprising of operational efficiency and computational intelligence. The statistical averaging of the various simulation runs is adopted to provide strong results and not an outlier effect.

**Comparative Analysis and Experimental Design**

The experimental process is designed in a way that allows making comparative assessment of three different set-ups traditional cloud-based AI, pure edge-based AI, and hybrid federated learning set-up. The typical centralized paradigm in the baseline experiment is a situation where all the IoT data is forwarded to the cloud to be processed. This case offers a point of reference to measure energy expenditure, latency as well as accuracy of inferences. The second system employs AI models at the edge nodes allowing a localized processing of data and the making of decisions. The purpose of this setting is to show the advantages of edge intelligence in reducing the cost of communication and ensuring real-time responsiveness. The third and last setup is the federated learning one, where the local models are trained by the edge devices and the global aggregation of parameter updates is done with the cloud. The controlled variations in the node density, data traffic, and network load are carried out in each experimental condition. The results are statistically tested to reveal whether the difference between the performance when using the three configurations is significant.

**Results**

A quantitative outcome of the simulation experiments conducted depicting three AI deployment paradigms of Cloud-based AI, Edge-AI, and Federated Edge-AI in a simulated low-energy IoT network environment is presented. These models were evaluated based on four primary performance metrics namely energy consumption, latency, bandwidth used and inference accuracy. The experiments were conducted in the same environmental and computing conditions in order to be compared and reproduced fairly.

**Overall Performance Summary**

The performance study was done at fifty IoT nodes that were evenly spaced in a simulated smart city topology. The nodes were simulated to reflect the real-life device properties such as limited energy sources, intermittent connectivity and constant sensing. The general behavior of each deployment architecture was represented by the average values of the repeated simulation trials, and the small difference between the runs ensured the stability and reproducibility of the results of the experiment. Table 1 summarizes results about the mean results of all the four key performance metrics in the three scenarios of AI deployment.

Table 1.

Comparative Performance Metrics for Cloud, Edge, and Federated Edge AI Architectures

| *Scenario* | *Avg Energy Consumption (mJ)* | *Avg Latency (ms)* | *Avg Bandwidth Usage (kB/s)* | *Avg Inference Accuracy (%)* |
|---|---|---|---|---|
| *Cloud-Based AI* | 14.73 | 453.36 | 250.97 | 96.94 |
| *Edge-AI Architecture* | 7.02 | 122.26 | 112.72 | 92.61 |
| *Federated Edge-AI* | 4.97 | 159.58 | 84.42 | 95.03 |

The quantitative summary shows that there is a marked performance difference between the three paradigms. The Cloud-based AI model was the most energy consuming and had the largest

latency, indicating that it requires remote centralized computation. The Edge-AI model, however, experienced significant decreases in both measures, which depict the usefulness of localized processing. Lastly, Federated Edge-AI system reached the greatest compromise between the almost near-cloud inference accuracy and the low energy and bandwidth consumption.

**Energy Consumption**

The energy consumption recorded in all architectures indicates a strong enhancement with the shift of the computation on the cloud to the network edge. In Cloud-based AI architect, the average energy consumption of each IoT node was 14.73 mJ per operation cycle as seen in figure 1, which indicates a high overhead due to the constant transmission of the raw sensor data to the remote servers. The consumption in the Edge-AI set up dropped to 7.02 mJ, which is 52% lower than the consumption in the cloud case. Federated Edge-AI model recorded the lowest energy consumption of 4.97 mJ which is by 66 % lower compared to the cloud baseline. The average standard deviation among all the nodes was less than 0.3 mJ making sure that the results were stable within the simulated environment. The figure also must represent a comparative side by side bar or column chart of the three levels of energy to help visually stress the progressive decrease between the cloud to the federated architectures.

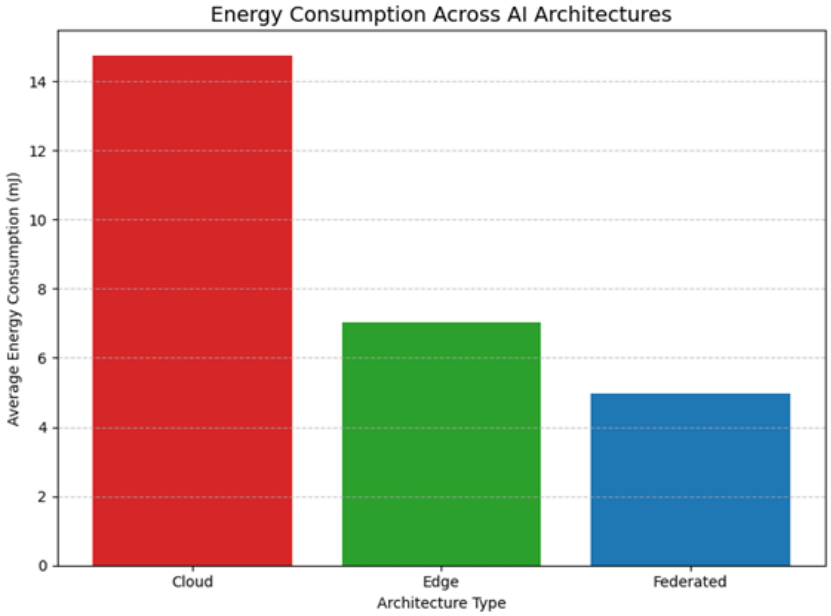


**Figure 1.** *Energy Consumption in IoT Networks*

Resource-constrained environments are highly dependent on energy consumption to ascertain the sustainability of the IoT network. The findings make it apparent that the proximity of computational intelligence to the network edge significantly decreases the power load of the IoT devices.

**Latency**

Latency is a direct factor affecting the responsiveness of IoT networks and one of the key factors that determine user experience and the ability to make a decision in real-time. Simulations yielded a significant decrease in latency with the type of processing intelligence being brought near data source. The AI setup based on the Cloud had the largest mean latency of 453.36 ms(shown in figure 2) as it represents a significant delay due to long-distance communication and centralized queueing of tasks. This kind of delay is not usually tolerated in time-sensitive applications of smart cities

system, including emergency response or autonomous traffic control.

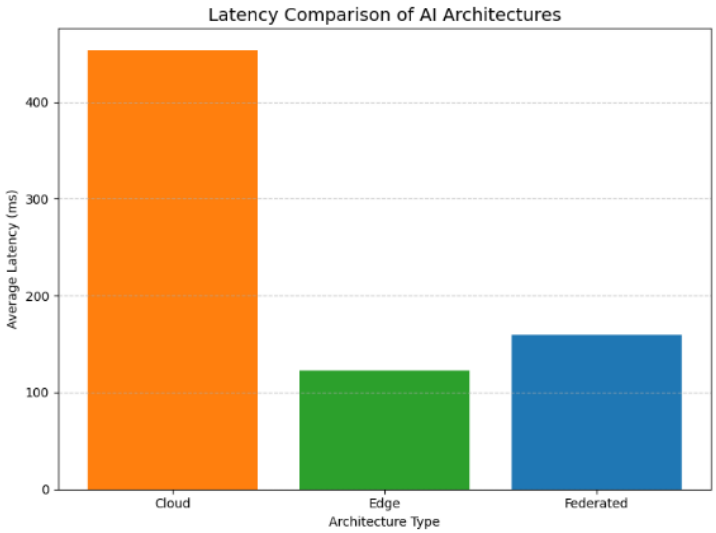


**Figure 2.** *Latency Comparison*

Comparatively, the Edge-AI model reduced to a stunning 122.26 ms, which is 73% shorter, with respect to response time. The importance of this decrease is the ability to process data instantly without a wide-area data transmission. Interestingly, in the Federated Edge-AI architecture, a latency of 159.58 ms was slightly higher as compared to the standalone edge approach. This small increment is attributed to periodic synchronization overheads and model update exchanges, which are associated with federated learning. Nevertheless, latency was still more than two times better than the cloud implementation, which corroborates the fact that distributed intelligence is able to provide significant responsiveness benefits even in federated coordination.

**Bandwidth Utilization**

Scalable IoT systems require bandwidth efficiency when they run on constrained or constrained network infrastructures. The results in figure 3 show that there is a steady declining tendency of bandwidth consumption as the intelligence shifted to the network edge. Cloud-based AI model recorded the largest communication overhead, and it was recorded that each node had an average of 250.97 kB/s of constant data transmission. This load can be seen as the constant uploading of uncoded sensor data to cloud data centres to perform centralised inferences.

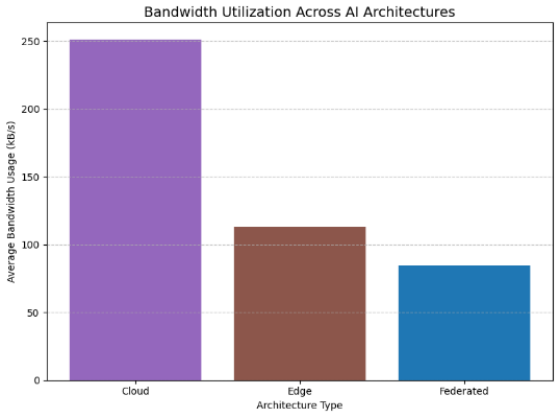


**Figure 3.** *Bandwidth Utilization*

Edge-AI on the other hand dropped the communication load considerably to 112.72 kB/s, yielding about 55 percent bandwidth consumption decrease. This is improved through processing the information on-the-fly and sending back merely inferences or summaries of decisions. The Federated Edge-AI setup had the lowest bandwidth needs of 84.42 kB/s which is 25% lower than standalone Edge-AI. Although communication between federated models was intermittent because of the model updates, the amount of data transferred was small since the federated models exchanged compressed model gradients rather than complete datasets. The simulated network exhibited consistent behavior in communication efficiency as the bandwidth difference between all of the nodes was found to be less than 5%.

**Inference Accuracy**

The accuracy of inference is the predictive reliability of AI models in distributed systems. The findings in figure 4 show that a higher energy and bandwidth efficiency did not significantly reduce the predictive performance. The Cloud-based AI model scored the best 96.94 consistency with the availability of centralized and complete data aggregation and high complexity models.

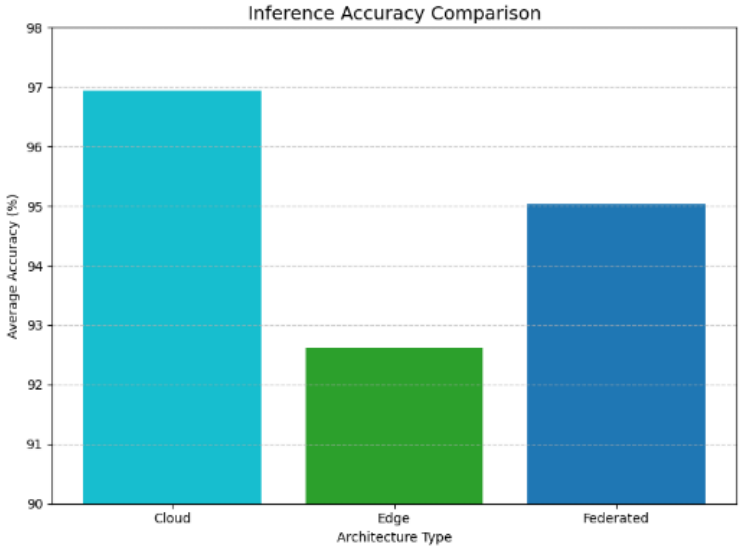


**Figure 4.** *Inference Accuracy*

An average accuracy of 92.61 was measured by the Edge-AI setup, which is 4.3 percent lower than that of the cloud. This minor change was anticipated because smaller model architectures were to be used in low-resource devices. However, the precision was acceptable in most of the applications of smart cities like predicting traffic flow or pollution. Surprisingly, Federated Edge-AI model reclaimed a significant proportion of the lost accuracy with a 95.03 result that is a good balance between efficiency and model accuracy. The distributed training of the multi-edge nodes enabled the world model to maintain representational richness without experiencing the communication and privacy costs of centralizing to the cloud. The standard deviation of the results of all the tests conducted was less than 1 percent which validated that the accuracy results had high reproducibility and reliability.

**Index of Performance**

In order to provide a consistent overall system performance comparison, all four performance measures were put on the same scale of 0 (poorest) to 1(best) and summed to a Composite Performance Index (CPI). Normalization was involved by using min-max scaling of each of the metrics of all the models. The increased CPI values are the signs of greater combined performance in all dimensions. Federated Edge-AI had the best composite index of 0.86, then Edge-AI had 0.66 and Cloud-based AI had 0.25. Such outcomes in figure 5 confirm that federated and edge computing systems are always more effective than the traditional centralized systems in all the critical operation parameters.

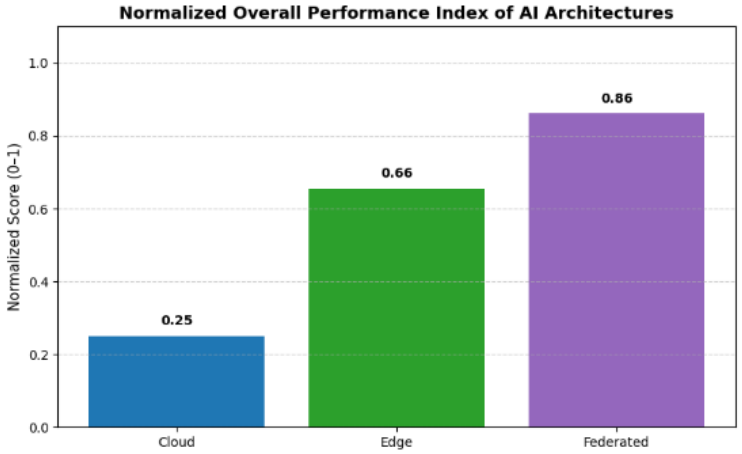


**Figure 5.** *Normalized Overall*

*Performance Index*

The empirical findings confirm the hypothesis that the distributed intelligence architecture is superior to the centralized model in low-energy environments of IoT in all operational parameters.

**Discussion**

Comparative study of Cloud-based AI, Edge-AI and Federated Edge-AI paradigm provides a set of explicit and measurable trade-offs between model performance, computational locality, and communication efficiency. The findings obtained by the modeled Internet of Things network illustrate that there is a very close reliance on the general system effectiveness on the position of computation and learning relative to the source of data. Although the cloud infrastructures offer access to big data integration and processing power, their centralized structure presents energy and latency costs that become major limitations to low-power IoT settings. Mobility of the intelligence to edge and federated nodes on the other hand provides a more energy conscious and latency conscious architecture

that is more compatible with the requirements of pervasive smart devices and real-time applications (Alahi et al., 2023).

The outcomes of energy use indicate the inefficiency of the centralized architectures when it comes to distributed sensing networks. The inherent problem with the Cloud-based AI model is that all IoT nodes consume a great deal more power, simply due to continuous emissions of raw sensor data to remote servers. This overall upstream communication is not only the cause of heavier overheads during transmission, but it also directly carries out smoking node batteries, which is a direct attack on sustainability and deployment life. The 52 percent of energy not used in the Edge-AI architecture of a cloud shows that local inference is successful in removing unnecessary data transfer. The only thing it will necessitate is the transfer of processed results or alerts sent out because it does the data processing close to the source which in turn will radically shrink the energy footprint. The federated edge-ai model was even more energy efficient because it nearly used two times less energy in its consumption compared to the cloud benchmark. This is an improvement which means that there is minimal reliance on core resources and sharing of learning updates, and not the raw data (Trigka and Dritas, 2025). The outcome justifies the claim that distributed and collaborative intelligence is an important facilitator of low-energy AI systems that can behave independently on the restricted IoT settings.

The additional advantage of decentralization of intelligence is demonstrated in the latency. Latency was the worst in the Cloud-based AI system because devices and centralized data centers had a delay of the end-to-end communication. This is not practical in the case of applications that need quick response, like autonomous systems, industrial IoT control, or medical monitoring. The transition to Edge-AI allowed cutting the latency by almost 73% which confirmed that localized inference can significantly improve system responsiveness. Despite the relative small improvement in latency with the Federated Edge-AI model compared to pure edge processing, this is a suitably acceptable trade-off given the periodic synchronization of model updates across nodes involved. The latency overhead caused by these updates is insignificant relative to the subsequent benefit of collaborative accuracy. This minor trade-off would probably be compensated in the real deployment by the capability to sustain shared intelligence without necessarily using high-bandwidth or consistent central links.

Using patterns of bandwidth is another testament to the better performance of the decentralized processing. The AI architecture based on the Cloud required the largest bandwidth, which aligns with the vast amount of sensor streams that were sent to centralised servers. Similarly, Edge-AI and Federated Edge-AI decreased this load by 55 and 66% respectively. Such savings are immediately converted into reduced network congestion and enhanced scalability of large-scale IoT ecosystems. This minimization of the transmission of the data not only conserves the energy in the communication, it also enhances the security and privacy of information as the sensitive information is not relayed to a far country, but is instead local to the machine or a small federation. This reduction is vital to the operations of edge deployments that can operate in environments with limited bandwidth (i.e. rural monitoring stations, vehicular networks, or wearable systems) or have a fluctuating connection (i.e. intermittently connected).

Regarding accuracy, the findings show that the efficiency improvement in decentralized models does not always imply a drop in prediction quality. The Cloud-based AI model achieved the best accuracy (96.94%), though it should be mention that the Edge-AI and Federated Edge-AI models achieved the level of 92.61% and 95.03, respectively. The loss in accuracy at edges of setups using the marginal models can be

explained by the smaller size of models and the limited local data. The Federated approach however reclaims much of this loss because nodes can learn using distributed experiences in the collaborative fashion, but raw data is not shared. This demonstrates one of the key benefits of federated learning, i.e. achieving cloud-level intelligence with edge level efficiency and privacy. The slight gap between the two models in the accuracy (less than 5 percent) proves that lightweight AI algorithms are no longer immature and can make high-quality inferences even when having resource constraints. Besides, this consistency in consistency also indicates that the system-level optimization (energy, latency, bandwidth) can be optimized without having a severe impact on the AI accuracy.

The composite performance index (CPI) creates a holistic image of the system performance in relation to all the parameters that one would consider as essential. Normalization of all measures allows it to make comparisons of individual strength of each architecture on a single scale. The CPI of 0.25 of Cloud-based AI, 0.66 of Edge-AI, and 0.86 of Federated Edge-AI is the obvious indicator of the superior balance of the federated model. The high composite score of Federated Edge-AI is its ability to handle the multidimensional challenges of AI-IoT integration: energy efficiency, responsiveness, communication load, and inference accuracy. And this also says of a direction of architectural evolution into intelligent systems: off as monolithic, centralised systems, and to distributed, cooperative and adaptable ecosystems that are sensitive to the heterogeneity and dynamism of IoT spaces.

Another theoretical area of the results is edge computing and federated learning. The two paradigms aim at minimizing the distance between the generation and esteem use of data to the lowest degree in order to minimize the overhead of the communication and maintain the data locality. The identified trade-offs such as marginally increased latency in the federated model, marginally different accuracy, and enhanced efficiency metrics could be explained by the literature available at the time when it could be emphasized that AI system design in the real world has to balance between computational and communication constraints and not optimize a single measure. This is a combination strategy that is necessary in the creation of scalable and sustainable AI-based IoT.

**Conclusion:**

The paper examined the performance of Cloud-based AI, Edge-AI and Federated Edge-AI paradigms under a simulated low-energy IoT network environment. The findings have shown that edge computing over cloud computing causes significant changes in the energy consumption and latency as well as relatively minor error in inference. Only Federated Edge-AI model showed balanced performance, and nearly cloud-like accuracy with a substantial savings in accident costs both in the energy and communication costs. These findings indicate that collaborative and distributed intelligence systems will be the direction to be taken as far as the future of smart city IoT infrastructures are concerned. As an open and scalable reconfigurable solution, Federated Edge-AI has the potential to meet the growing computational and ethical needs of the future internet-of-things ecosystem and ensure the effective functioning of the ecosystems along with the protection of the data. Lastly, federated learning and edge computing offers a promising way towards autonomous and low-energy and intelligent networks that can be used to enable resilient smart city operations.

**References**

Goel, R. K., & Vishnoi, S. (2022). Urbanization and sustainable development for inclusiveness using ICTs. Telecommunications Policy, 46(6), 102311. DOI: https://doi.org/10.1016/j.telpol.2022.102311

Alahi, M. E. E., Sukkuea, A., Tina, F. W., Nag, A., Kurdthongmee, W., Suwannarat, K., & Mukhopadhyay, S. C. (2023). Integration of IoT-enabled technologies and artificial intelligence (AI) for smart city scenario: recent advancements and future trends. Sensors, 23(11), 5206. DOI: https://doi.org/10.3390/s23115206

Bellini, P., Nesi, P., & Pantaleo, G. (2022). IoT-enabled smart cities: A review of concepts, frameworks and key technologies. Applied sciences, 12(3), 1607. DOI: https://doi.org/10.3390/app12031607

Trigka, M., & Dritsas, E. (2025). Edge and cloud computing in smart cities. Future Internet, 17(3), 118. DOI: https://doi.org/10.3390/fi17030118

Al-Jumaili, A. H. A., Muniyandi, R. C., Hasan, M. K., Paw, J. K. S., & Singh, M. J. (2023). Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. Sensors, 23(6), 2952. DOI: https://doi.org/10.3390/s23062952

Kanellopoulos, D., Sharma, V. K., Panagiotakopoulos, T., & Kameas, A. (2023). Networking architectures and protocols for IoT applications in smart cities: Recent developments and perspectives. Electronics, 12(11), 2490. DOI : https://doi.org/10.3390/electronics12112490

Gill, S. S., Golec, M., Hu, J., Xu, M., Du, J., Wu, H., ... & Uhlig, S. (2025). Edge AI: A taxonomy, systematic review and future directions. Cluster Computing, 28(1), 18. DOI: https://doi.org/10.1007/s10586-024-04686-y

Garcia-Perez, A., Miñón, R., Torre-Bastida, A. I., & Zulueta-Guerrero, E. (2023). Analysing edge computing devices for the deployment of embedded AI. Sensors, 23(23), 9495. DOI: https://doi.org/10.3390/s23239495

Xu, W., Yang, Z., Ng, D. W. K., Levorato, M., Eldar, Y. C., & Debbah, M. (2023). Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing. IEEE journal of selected topics in signal processing, 17(1), 9-39. DOI: 10.1109/JSTSP.2023.3239189

Xu, S., Manshaii, F., Xiao, X., & Chen, J. (2025). Artificial intelligence assisted nanogenerator applications. Journal of Materials Chemistry A, 13(2), 832-854. DOI: https://doi.org/10.1039/D4TA07127A

Lamrini, M., Chkouri, M. Y., & Touhafi, A. (2023). Evaluating the performance of pre-trained convolutional neural network for audio classification on embedded systems for anomaly detection in smart cities. Sensors, 23(13), 6227. DOI: https://doi.org/10.3390/s23136227

Ahmed, Z. E., Hashim, A. A., Saeed, R. A., & Saeed, M. M. (2024). TinyML network applications for smart cities. In TinyML for Edge Intelligence in IoT and LPWAN Networks (pp. 423-451). Academic Press. DOI: https://doi.org/10.1016/B978-0-44-322202-3.00023-3

Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., ... & Celdrán, A. H. (2023). Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. IEEE Communications Surveys & Tutorials, 25(4), 2983-3013. DOI: 10.1109/COMST.2023.3315746

Dang, X. T., Vu, B. M., Nguyen, Q. S., Tran, T. T. M., Eom, J. S., & Shin, O. S. (2024). A survey on energy-efficient design for federated learning over wireless networks. Energies, 17(24), 6485. DOI: https://doi.org/10.3390/en17246485

Pooyandeh, M., & Sohn, I. (2021). Edge network optimization based on ai techniques: A survey. Electronics, 10(22), 2830. DOI : https://doi.org/10.3390/electronics10222830

Rajab, H., Al-Amaireh, H., Bouguera, T., & Cinkler, T. (2023). Evaluation of energy consumption of LPWAN technologies. EURASIP Journal on Wireless Communications and Networking, 2023(1), 118. DOI: https://doi.org/10.1186/s13638-023-02322-8

Zheng, X., Guo, R., & Lian, S. (2025). Energy-Efficient Load Balanced Edge Computing Model for IoT Using FL-HMM and BOA Optimization. Sustainable Computing: Informatics and Systems, 101215. DOI: https://doi.org/10.1016/j.suscom.2025.101215

Letaief, K. B., Shi, Y., Lu, J., & Lu, J. (2021). Edge artificial intelligence for 6G: Vision, enabling technologies, and applications. IEEE journal on selected areas in communications, 40(1), 5-36. DOI:10.1109/JSAC.2021.3126076

Raj, A., & Shetty, S. D. (2024). Smart parking systems technologies, tools, and challenges for implementing in a smart city environment: a survey based on IoT & ML perspective. International Journal of Machine Learning and Cybernetics, 15(7), 2673-2694. DOI: https://doi.org/10.1007/s13042-023-02056-5

Philipo, A. G., Ning, H., Sarwatt, D. S., Mohamed, J. A., Yusufu, A. S., Shi, F., ... & Ding, J. (2025). Sustainable AI: Emerging Trends, Impacts, and Future Challenges. IEEE Transactions on Sustainable Computing. DOI : 10.1109/TSUSC.2025.3611272

Surianarayanan, C., Lawrence, J. J., Chelliah, P. R., Prakash, E., & Hewage, C. (2023). A survey on optimization techniques for edge artificial intelligence (AI). Sensors, 23(3), 1279. DOI : https://doi.org/10.3390/s23031279

Bourechak, A., Zedadra, O., Kouahla, M. N., Guerrieri, A., Seridi, H., & Fortino, G. (2023). At the confluence of artificial intelligence and edge computing in iot-based applications: A review and new perspectives. Sensors, 23(3), 1639. DOI : https://doi.org/10.3390/s23031639

Rajput, M., & Yadav, R. N. (2025). Machine and Deep Learning Driven Energy Efficient Clustering in IOT-WSNs: A Review. IEEE Sensors Journal. DOI: 10.1109/JSEN.2025.3613053

Hartmann, M., Hashmi, U. S., & Imran, A. (2022). Edge computing in smart health care systems: Review, challenges, and research directions. Transactions on Emerging Telecommunications Technologies, 33(3), e3710. DOI : https://doi.org/10.1002/ett.3710

Ji, L., He, S., Gu, C., Shi, Z., & Chen, J. (2023). Routing and scheduling for low latency and reliability in time-sensitive software-defined IIoT. IEEE Internet of Things Journal, 11(7), 12929-12940. DOI:10.1109/JIOT.2023.3337941

Xu, R., Razavi, S., & Zheng, R. (2023). Edge video analytics: A survey on applications, systems and enabling techniques. IEEE Communications Surveys & Tutorials, 25(4), 2951-2982. DOI: 10.1109/COMST.2023.3323091

Duan, S., Wang, D., Ren, J., Lyu, F., Zhang, Y., Wu, H., & Shen, X. (2022). Distributed artificial intelligence empowered by end-edge-cloud computing: A survey. IEEE Communications Surveys & Tutorials, 25(1), 591-624. DOI: 10.1109/COMST.2022.3218527

Hua, H., Li, Y., Wang, T., Dong, N., Li, W., & Cao, J. (2023). Edge computing with artificial intelligence: A machine learning perspective. ACM Computing Surveys, 55(9), 1-35. DOI: https://doi.org/10.1145/3555802

Liu, H. I., Galindo, M., Xie, H., Wong, L. K., Shuai, H. H., Li, Y. H., & Cheng, W. H. (2024). Lightweight deep learning for resource-constrained environments: A survey. ACM Computing Surveys, 56(10), 1-42. DOI: https://doi.org/10.1145/3657282

Wong, A. W. L., Goh, S. L., Hasan, M. K., & Fattah, S. (2024). Multi-hop and mesh for LoRa networks: Recent advancements, issues, and recommended applications. ACM Computing Surveys, 56(6), 1-43. DOI : https://doi.org/10.1145/3638241

Swamy, H. (2024). SMART SPENDING: HARNESSING AI TO OPTIMIZE CLOUD COST MANAGEMENT. Development (IJAIRD), 2(2), 40-55. DOI: https://doi.org/10.5281/zenodo.13132258

Belli, L., Cilfone, A., Davoli, L., Ferrari, G., Adorni, P., Di Nocera, F., ... & Bertolotti, E. (2020). IoT-enabled smart sustainable cities: Challenges and approaches. Smart Cities, 3(3), 1039-1071. DOI: https://doi.org/10.3390/smartcities3030052

Abreha, H. G., Hayajneh, M., & Serhani, M. A. (2022). Federated learning in edge computing: a systematic survey. Sensors, 22(2), 450. DOI: https://doi.org/10.3390/s22020450

Schizas, N., Karras, A., Karras, C., & Sioutas, S. (2022). TinyML for ultra-low power AI and large scale IoT deployments: A systematic review. Future Internet, 14(12), 363. DOI: https://doi.org/10.3390/fi14120363

Yang, Z., Chen, M., Wong, K. K., Poor, H. V., & Cui, S. (2022). Federated learning for 6G: Applications, challenges, and opportunities. Engineering, 8, 33-41. DOI: https://doi.org/10.1016/j.eng.2021.12.002

Khan, F. U., Shah, I. A., Jan, S., Ahmad, S., & Whangbo, T. (2025). Machine learning-based resource management in fog computing: A systematic literature review. Sensors, 25(3), 687. DOI: https://doi.org/10.3390/s25030687

Qin, B., Pan, H., Dai, Y., Si, X., Huang, X., Yuen, C., & Zhang, Y. (2024). Machine and deep learning for digital twin networks: A survey. IEEE Internet of Things Journal, 11(21), 34694-34716. DOI: 10.1109/JIOT.2024.3416733

Anagnostopoulos, C., Gkillas, A., Mavrokefalidis, C., Pikoulis, E. V., Piperigkos, N., & Lalos, A. S. (2024). Multimodal federated learning in AIoT systems: Existing solutions, applications, and challenges. IEEE Access. DOI: 10.1109/ACCESS.2024.3508030

Shahid, O., Pouriyeh, S., Parizi, R. M., Sheng, Q. Z., Srivastava, G., & Zhao, L. (2021). Communication efficiency in federated learning: Achievements and challenges. arXiv preprint arXiv:2107.10996. DOI: https://doi.org/10.48550/arXiv.2107.10996

Fay, C. D., Corcoran, B., & Diamond, D. (2023). Green IoT event detection for carbon-emission monitoring in sensor networks. Sensors, 24(1), 162. DOI: https://doi.org/10.3390/s24010162

Ahmed, S., Saeed, M. K., & Khokhar, A. (2025). OSI Stack Redesign for Quantum Networks: Requirements, Technologies, Challenges, and Future Directions. arXiv preprint arXiv:2506.12195. DOI: https://doi.org/10.48550/arXiv.2506.12195

Du, J., Jiang, C., Benslimane, A., Guo, S., & Ren, Y. (2022). SDN-based resource allocation in edge and cloud computing systems: An evolutionary Stackelberg differential game approach. IEEE/ACM Transactions on Networking, 30(4), 1613-1628. DOI: 10.1109/TNET.2022.3152150

Shen, W., Lin, W., Wu, W., Wu, H., & Li, K. (2025). Reinforcement learning-based task scheduling for heterogeneous computing in end-edge-cloud

environment. Cluster Computing, 28(3), 179. DOI: https://doi.org/10.1007/s10586-024-04828-2