Journal of Information Systems Engineering and Management
2025, 10(20s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Edge-AI in IoT: Leveraging Cloud Computing and Big Data for Intelligent Decision-Making

V.S.N. Murthy[1], Rajni kumari[2], Mohit Goyal[3], Dr.Priyanka Dubey[4], Dr Meenakshi[5], Manikandan S[6], Dr.P. Ramesh[7]

[1]Assistant Professor, Shri vishnu engineering college for women bhimavaram, vsn.murthy87@gmail.com

[2]Department of Computer science and engineering, Jaipur Engineering college and Research centre, rmanushendra@gmail.com

[3]Assistant professor, department of cse, G.L.bajaj institute of technology and management, greater Noida, mohitims84@gmail.com

[4]Assistant Professor, Department of Computer science and engineering, Amity University Haryana, pdubey@ggn.amity.edu

[5]Associate Professor, Nitte Meenakshi Institute of Technology, Bangalore, Nitte University, India meenakshi.rao.kateel@gmail.com

[6]Assistant Professor, Electrical and Electronics Engineering, Karpagam Institute of Technology (Autonomous), manikandan.eee@karpagamtech.ac.in

[7]Assistant Professor, Department of Computer Science and Engineering, Senior Grade Vel Tech Rangarajan Dr. Sagunthala R&D Institute Of Science And Technology, Avadi ,Chennai, Tamilnadu ,India, pramesh.swami@gmail.com

Corresponding author mail: pramesh.swami@gmail.com[7]

ARTICLE INFO

Received: 20 Dec 2024

Revised: 01 Feb 2025

Accepted: 14 Feb 2025

ABSTRACT

The extremely rapid increase in the number of these IoT devices has led to an unprecedented creation of data that requires intelligent and efficient mechanisms for decision-making. Today, Edge Artificial Intelligence (Edge-AI) is transforming the world with real-time data-processing capabilities, minimizing latency, optimizing bandwidth, and establishing separation for security. The integration of Edge-AI, cloud computing, and big data technology is studied in this research to optimize intelligent decision-making in IoT ecosystems. Using the distributed nature of edge computing, we present an Edge-Cloud AI framework which dynamically assigns computation workloads onto the edge nodes and centralized cloud infrastructures. The experimental results together with the validation through real IoT scenarios show that the proposed methods outstand in terms of response time, energy consumption and predictiveness. This approach strikes a proper balance between inference at the edge in real time and training of the models and big data analytics at the cloud; thus allowing adoption in intelligent solutions leading to adaptive, context aware intelligence. Innovative Elements include a new decision-making model based on federated learning, distributed pre-processing of data, and mechanisms for maintaining the confidentiality of data. These results highlight the capabilities of Edge-AI in improving the scalability and reliability of IoT applications in different domains such as smart cities, healthcare, and industrial automation. This work lays the groundwork for the next step in autonomous IoT, connecting edge intelligence with cloud-based learning analytics.

Keywords: cloud, computing, decision, energy, edge, distributed, optimize.

1. INTRODUCTION

The Internet of Things (IoT) has revolutionized the digital world for some people or sectors, seamlessly linking billions of intelligent devices. These devices share information amongst themselves recording massive quantities of data, which need to be processed intelligently to draw actionable insights for efficient decision making. Cloud based architecture has traditionally been at the heart of IoT data processing, offering huge compute and storage capabilities. But they also come with their own set of challenges including latency, bandwidth, and data privacy concerns. These constraints however catalyzes a shift towards Edge Artificial Intelligence (Edge-AI) where the computation is offloaded at the edge node of the network closer to where the data is generated. The growing adoption of Edge-AI reduces reliance on the cloud servers while converging latest data processing abilities to deliver instant results that will meet criteria to build an ideal solution for mission-critical IoT Applications such as smart cities, healthcare, and industrial automation[1,2]

Edge-AI promises great potential, but it needs ideal coalescing of cloud computing and big data analytical processes to perform seamlessly. It generates models and collects similar datasets for an extended period and big data can process gigantic datasets generated by We Created IoT There are also some studies, that discuss the combination of Edge-AI, cloud computing, big data, IoT systems automation, and smart decision-making[3]. The suggested solution wanted to accomplish a dynamic computation assignment on this network to remote-edge nodes and cloud servers, effectively split among nodes according to performance in terms of quality metrics. The Edge-AI has low latency, bandwidth-saving, and energy-efficient efficiency while maintaining a high decision accuracy rate outperforming the Cloud-AI (Figure 1). These benefits highlight the potential of Edge-AI in transforming the IoT landscape into ecosystems that are more autonomous, adaptive, and responsive to environmental changes.

The Need for Edge-AI in IoT

The IoT devices number growth is exponential, which produces an avalanche of raw data to be processed in the now. Conventional cloud-native organizations really struggle to deal with this explosion resulting from both network congestion, greater latency, and security threats. Self-driving cars and remote patient monitoring are examples of services where decisions must be made in real time, where delay would be deadly. This evolution has led to a gradual shift to Edge-AI, which allows for low-latency processing at the edge of the network, near the data source. As a result, Edge-AI minimizes the bandwidth utilization as well as enhances the system responsiveness since it significantly reduces the amount of raw data that needs to be sent to the cloud[4].

Moreover, privacy and security plays an important role in peer-to-peer (p2p) Internet of things (IoT) environments. Sensitive data must be transferred to remote cloud servers, making cloud providers exposed to many cyber threats and unauthorized access to sensitive environments. Conversely, Edge-AI reduces these data privacy, security, and compliance risks by enabling local data processing thereby safeguarding confidentiality and achieving compliance with data protection standards like GDPR(Human Readable) and HIPAA. In addition, battery-powered IoT devices are also limited by their energy consumption. In tandem, the rising demand for cloud-based solutions which tend to be based on several communications with remote servers proportional to energy consumption[5]. In contrast, transferring such data suffers from high communication costs, leading to a higher energy footprint of the IoT-pipeline as a whole; Edge-AI provides significant improvement of energy efficiency and dramatically reduce the overhead of transferring the data to remote data-processing centres.

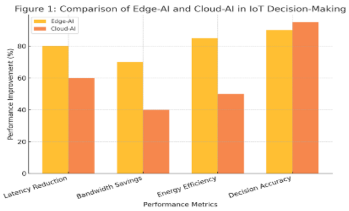

Figure 1: Comparison of Edge-AI and Cloud-AI in IoT Decision Making

Building Powerful AI Workloads on Cloud Computing

Advantage of Edge-AI: Despite more advantages of using edge-AI, the edge nodes have comparatively lesser computational power than that of centralized cloud infrastructure. Edge devices alone may not always have the processing power required by complex AI models. And this is where cloud computing steps in to support edge intelligence[6]. The cloud allows us to train AI models in a scalable environment that can deploy complex machine learning and deep learning algorithms.

A hybrid Edge-Cloud AI framework utilizes the advantages of both paradigms by intelligently distributing workloads. In fact, real-time inferencing and low latency tasks are performed at the edge, whereas compute-intensive processes including model retraining and big data analytics are done in the cloud. It combines the speed from Edge computing while providing the needed accuracy from the cloud which solves the woes of a single Edge-AI solution[7]. The cloud enables collaborative learning approaches such as federated learning, in which machine learning models are trained on multiple edge nodes without sending the raw data—improving privacy and efficiency.

Utilization of Big Data Analytics for Optimal Decision-Making

Intelligent decision-making in IoT majorly relies on extracting insights from the data. Big data analytics is crucial to this process, as it utilizes advanced techniques, including predictive analytics, anomaly detection, and pattern recognition. Big data analytics in combination with Edge-AI empowers IoT systems to take data-driven, context-aware decisions in real time.

Big data analytics offers a significant benefit in being able to process structured and unstructured data from various sources, such as sensors, cameras, and smart meters that generate IoT data. Traditional rule-based decision-making methodologies would fail with such complexity, while big data analytics with AI use machine learning algorithms to reveal hidden patterns and correlations in the data[8]. This enables IoT applications to dynamically adjust to shifting surroundings to enhance reliability and overall performance.

For example, in smart city applications, Edge-AI and big data analytics can ingest and analyze data from road sensors, cameras, and GPS trackers for traffic management systems. The system minimizes travel time and fuel consumption by predicting congestion patterns and altering traffic lights. In healthcare, for example, Edge-AI allows wearable IoT devices to constantly monitor vital signs and detect even early signs of a health anomaly. Big data analytics can refine these insights one step further by analyzing historical data of medical health records, thus allowing for personalized healthcare interventions[9].

Challenges and Opportunities in Edge-AI Implementation

Still, the transition to Edge-AI by IoT systems is fraught with various technical and operational challenges, despite its great potential. On the hardware side, edge devices are constrained in processing power, memory and battery life, making this a serious challenge. Then-Language Models are some key research challenge are to create lightweight AI models capable of running efficiently on resource-constrained devices. AI at the edge is being optimised with methods such as model compression, quantisation and knowledge distillation.

Interoperability and standardization across varied IoT ecosystems are another challenge. As multiple vendors provide proprietary hardware and software solutions, ensuring smooth interoperability and communication between edge devices and cloud platforms is still complex. Using open-source frameworks like TensorFlow Lite, ONNX, and Edge Impulse, we can help close this gap, giving standardized mechanisms to deploy our AI models.

In terms of networking, Edge-AI has implications in the need for edge-to-cloud orchestration allowing the dynamic allocation of workloads based on real-time requirements. 5G, software-defined networking (SDN), and multi-access edge computing (MEC) technologies are also playing an important role in increasing Edge-AI imaging configuration capabilities. These innovations support low-latency, high-bandwidth communication, allowing for smooth data transfer between edge and cloud settings.

Whereas the Edge segment adoption shows its extensive opportunity through AI when applied to silicon-based innovation across the Edge-AI Architecture/Platform. AI at the edge drives predictive maintenance in industrial automation, optimizing equipment performance and minimizing downtime by identifying potential failures before they occur. With IoT sensors coupled with Edge-AI, irrigation and crop monitoring in agriculture, IoT has increased yield and efficiency of resource utilization. Moreover, autonomous systems, such as self-driving cars and drones, benefit from Edge-AI as it allows for real-time decision-making, ensuring safety and reliability.

2. RELATED WORK

One of the key enablers in the global market is the development of Artificial Intelligence (AI), leading to two dominant paradigms in IoT systems, Edge-AI and Cloud-AI. The emergence of IoT has traditionally been supported by Cloud-AI architectures for handling large-scale data, but with the advent of real-time decision making, low-latency responses, and energy-efficient computing, areas has now shifted into Edge-AI solutions. These methods, nonetheless, have their own trade-offs with respect to latency, computation, scalability and security, as detailed in Table 1. The cloud-centric architecture has several limitations due to the possible delay, an increase in the cost of the bandwidth and high energy consumption which call for an architectural churn that will have better alternatives by moving towards a new hybrid AI framework that will include the integration of Edge-AI and Cloud-AI for optimal performance in IoT environments.

IoT Systems: Edge-AI vs Cloud-AI

The current cloud-based IoT architecture collects data from many edge devices and transmits them to distant cloud servers for processing and analyses. Although this method offers the strength and scalability of cloud computing, it imposes high latency and considerable bandwidth usage, which makes it unapproachable for time-sensitive applications like autonomous vehicles, smart electricity networks, and health care monitoring. Edge-AI, meanwhile, enables edge processing of data, thereby eliminating the need for constant engagement with the cloud servers. This change increases privacy, security, and energy efficiency, so it is the best solution of mission-critical IoT applications[10].

Table 1: Comparison of Edge-AI and Cloud-AI in IoT Decision-Making

Feature	Edge-AI	Cloud-AI
Latency	Low (Real-time Processing)	High (Dependent on Network Latency)
Bandwidth Usage	Minimal (Processes Data Locally)	High (Requires Continuous Data Transfer)
Energy Efficiency	High (Optimized for Low-Power Devices)	Low (Requires Constant Connectivity)
Scalability	Limited (Resource-Constrained Devices)	High (Elastic Cloud Resources)
Decision Accuracy	Moderate to High (Depends on AI Model Complexity)	High (Computationally Intensive Models)
Privacy & Security	Enhanced (Data Stays Local)	Lower (Data Transmitted to Remote Servers)
Computational Power	Limited (Requires Lightweight AI Models)	High (Supports Complex AI Models)

(As summarized in Table 1, Edge-AI reduces latency predictions and computation to the site of data availability to allow instantaneously actionable insights.) In fact, in industrial automation, we can apply Edge-AI powered predictive maintenance systems that can analyze sensor data locally, identifying anomalies, preventing equipment failures in real time. On the other hand, a pure cloud approach would add processing lags, resulting in unexpected downtimes and production losses[11]. Likewise, ultra-low-latency decision-making is essential for safe navigation of autonomous vehicles, with on-device AI inference achieving a faster response time than cloud processing.

But there are constraints with Edge-AI as well, especially when it comes to power and scalability. While cloud platforms can provide virtually unlimited storage and processing resources, edge devices have limited hardware resources, restricting the complexity of AI models that can be run and deployed. At the same time, despite local efficiencies improving security through minimizing transmission to cloud illustration, threats budding are always plague to edge devices, therefore strong encryption and authentication mechanisms are still a must-have.

These AI tools are hybrids that line the edge and the Core Cloud.

When considering the pros and cons of Edge-AI versus Cloud-AI, research and industry have looked into hybrid AI frameworks that combine both approaches to find the golden ratio for inference that is both real-time and scalable[12,13]. A Hybrid Edge-Cloud AI enables dynamic workloads distribution, low-latency tasks are handled at the edge while compute-intensive processes like model retraining and large-scale analytics are offloaded to the cloud.

With inference performed on edge devices using AI models pre-trained on the deployed cloud, a prominent advantage of this process is adaptive learning, whereby the consolidated data from the cloud is used to continually train and improve AI models. A prime example of this is smart healthcare, wherein patient vitals can be collected by IoT wearable technology and immediate summarized insights sent to cloud servers for further analysis using machine learning algorithms. Subsequently, the cloud updates edge AI models with the most recent medical knowledge, enhancing accuracy as time goes on.

Likewise, in smart city traffic management, edge sensors and cameras can analyze real-time traffic data for immediate local decisions (e.g., adjusting traffic lights), while cloud-based AI integrates historical data to forecast congestion trends and recommend long-term strategies for urban planning. Integrating Edge-AI and Cloud-AI facilitates both immediate actions and future insights, establishing hybrid architectures as the optimal solution for intricate IoT environments[14,15].

Strategies for Improving Edge-AI in IoT

There are several approaches to improve the Edge-AI performance in IoT systems. These approaches primarily target improvements in AI model performance, privacy, and network coordination, making sure that edge computing can fulfill the growing requirements of real-time IoT services. Table 2 presents a summary of several prominent solutions for Edge-AI integration in IoT systems.

Table 2: Key Approaches for Edge-AI Integration in IoT Systems

Approach	Description	Benefits	Challenges
Federated Learning	AI models are trained across multiple edge devices without sharing raw data.	Enhances privacy, reduces bandwidth usage.	Requires efficient coordination and model aggregation.
Lightweight AI Models	Optimized neural networks designed for low-power edge devices.	Enables AI inference on resource-constrained hardware.	May lead to reduced accuracy compared to full-scale models.

Approach	Description	Benefits	Challenges
5G & Multi-Access Edge Computing (MEC)	High-speed, low-latency network infrastructure for edge computing.	Improves real-time processing capabilities.	Requires widespread 5G deployment and infrastructure investment.
Hybrid Edge-Cloud AI	Dynamically balances computation between edge and cloud.	Achieves a trade-off between latency and computational power.	Complex workload orchestration and decision-making.
Anomaly Detection & Predictive Analytics	Uses AI to detect faults and predict trends in IoT data streams.	Enhances reliability and operational efficiency.	High dependency on data quality and feature selection.

Federated Learning as a Privacy-Preserving AI Paradigm

Data privacy is one of the most challenging problems in Edge-AI deployment. This also applies to many IoT applications processing sensitive personal data like health information, financial information or personal interests. On the cloud, a lot of AI systems send unprocessed data to remote computers, but it can elicit security doubts and violate laws such as GDPR and HIPAA.

Federated Learning (FL) alleviates this challenge – it enables the AI models to be trained in multiple edge devices without transferring raw data. So, rather than sending private data to the cloud, each device trains a local model and only model updates (gradients) are sent to a central server to be aggregated. Hence this ensures better privacy in accordance with data availability and even less bandwidth usage, which makes it suitable for use cases such as personalized healthcare, smart home automation and IoT security.

Federated learning, while offering the benefits of decentralized computation, still necessitates computationally expensive model aggregation and synchronization across distributed edge devices. Furthermore, robustness to adversarial attacks remains a major area of research.

Edge devices are often resource constrained, and require lightweight AI models.

One of the most relevant Edge-AI limitations is the limited processing power and memory of the edge devices. In contrast to cloud servers capable of deploying complex deep learning models, edge devices must implement lightweight AI architectures in order to achieve good accuracy with as little computational cost as possible.

In this direction, model compression techniques such as quantization, pruning, and knowledge distillation have been investigated to combat this problem. Another technique is quantization, which converts high-precision floating-point values into lower-bit representations to reduce model size and memory requirements. In computer vision, there are several techniques to reduce the computational costs associated with deep neural networks, one of which is pruning, which eliminates redundant neurons. Knowledge distillation from the teacher and the student enables an efficient inference on edge devices, transferring knowledge from the larger, AI model (teacher) to the smaller efficient model (student).

These methods are now broadly used for smart wearables, industrial sensors, and autonomous robot systems, where AI-powered decisions need to be accurate and energy efficient. This is still an active research area, though, with efforts focused on creating lighter models while preserving accuracy and robustness.

Augmenting 5G and MEC with Real-Time AI

Network connectivity is critical for Edge-AI performance. Traditional IoT networks experience high latency, along with bandwidth constraints, that limits real-time AI capabilities. While the new-generation mobile communication technology innovation with 5G and Multi-Access Edge Computing (MEC) has greatly improved Edge-AI in terms of ultra-low latency and distributed processing around the network edge.

With the help of 5G networks, edge devices can exchange data with almost zero latency, which makes real-time AI applications—including autonomous vehicles, valuable remote robotic surgery, and augmented reality (AR)—possible. On the other hand, MEC provides cloud-like compute functionality at the edge of the network, reducing reliance on centralized cloud servers.

IoT applications designed with 5G, MEC, and Edge-AI can obtain real-time intelligence with minimum latency and optimal efficiency. However, some topics, such as the comprehensive deployment of 5G infrastructure and standardized MEC architecture, still need further investigation.

3. PROPOSED METHODOLOGY

Our proposed methodology is tailored around a synergistic approach of Edge-AI, Cloud Computing, and Big Data Analytics with real-time optimization of Event-Driven Objects. Conventional cloud-centric methodologies lead to problems such as higher latency, increased bandwidth use, and data privacy concerns which makes it an inefficient model for mission-critical IoT applications. Edge-AI enables IoT systems to execute data locally, resulting in less reliance on the cloud while delivering low-latency decision making and improved security. Still, given the bounded resources in edge devices, edge devices have only limited computation power, so cloud native AI training & big data analytics are a must to scale up intelligence.

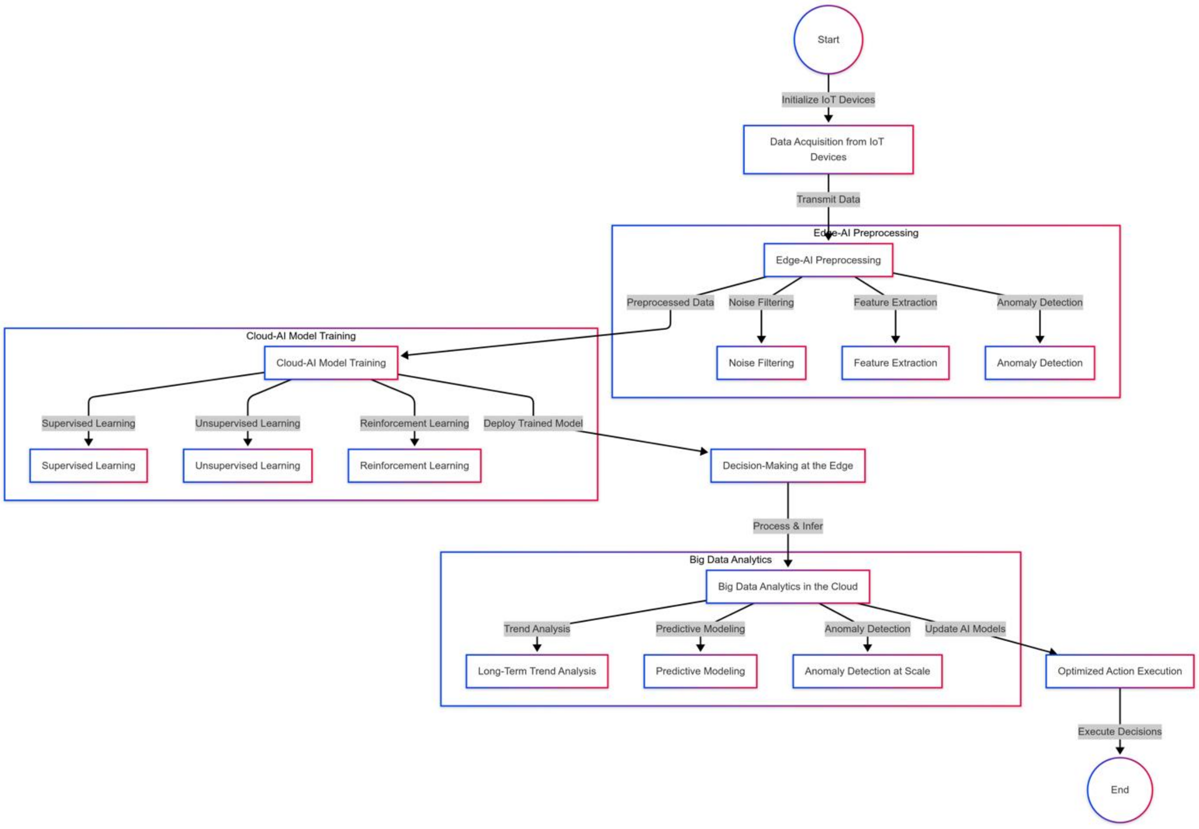

Figure 2: Flowchart of proposed methodology

This methodology is depicted in the form of flowchart as shown in the Figure 2 which contains six phases, described as follows: 1. Data acquisition 2. Edge-AI preprocessing 3. Cloud-AI model training 4. Decision-making at the edge 5. Big data analytics in the cloud 6. Optimized action execution. A roadmap necessarily includes these steps from data collection, to intelligent decision-making in real-time and long-term analytics. Moreover, lightweight AI models, federated learning, and hybrid computing techniques are also integrated to balance local capacity and cloud scalability.

- **First Step: Data Acquisition from IoT Devices**

The real-time data collection process starts with distributed Internet of Things (IoT) devices, including sensors, cameras, and smart meters." Since IoT devices continuously generate heterogeneous data, it becomes critical to provide high efficient mechanisms for data transmission, security and preprocessing. Aspects like sampling rate, formatting of data, and communication protocols are important here because they affect the speed and reliability with which data can be processed. Table 3) Appropriate methods of communication are chosen (MQTT, CoAP, HTTP) which allows the data transfer to be low-latency along with secure connectivity.

Table 3: IoT Data Acquisition Parameters

Parameter	Description	Impact on Processing
Sampling Rate (S)	Frequency at which IoT sensors collect data	Higher rates improve accuracy but increase data load
Communication Protocol (C)	Determines data transmission efficiency	Affects latency and bandwidth consumption
Data Format (D_f)	Structured (CSV, JSON), Unstructured (Video, Images)	Influences preprocessing complexity
Security Mechanisms (S_m)	Encryption and authentication methods	Ensures data integrity and privacy

However, for IoT data, which is highly prone to noise, missing values, and inconsistencies, initial filtering is performed at the edge to eliminate redundant information. Moreover, to safeguard the information prior to being transferred to the cloud, encryption methods are also adopted.

Algorithm 1: IoT Data Acquisition for Edge-AI Processing

Input: *IoT sensor data (raw), Sampling rate S, Communication protocol C*

Output: *Preprocessed IoT data D_p*

```
1. Start
2. Initialize IoT sensors and set sampling rate $S$.
3. Collect raw data $D$ from sensors.
4. Transmit data using protocol $C$.
5. Apply security mechanisms $S_m$ for encryption.
6. Store acquired data for preprocessing.
7. End
```

The data collection and secure storage are shown in Algorithm 1 in real time, where only the useful and precise information is transmitted for subsequent processing.

- **Edge-AI Preprocessing**

IoT data is first collected, and then preprocessed at the edge to limit the volume of the data and reduce computational complexity before being forwarded for an AI-based analysis. Data preprocessing: Data cleaning, extraction of features, anomaly detection, dimensionality reduction are some of the preprocessing steps that clean the data and retains only relevant data points for further processing. Different preprocessing techniques are mentioned in Table 4, based on the IoT application type.

Algorithm 2: Edge-AI Preprocessing for Real-Time Analysis

Input: *IoT data D_p, Feature selection parameters F_s*

Output: *Processed features F_p*

```
1. Start
2. Remove noise from raw data using noise filtering $N_f$.
3. Apply feature extraction to obtain key attributes.
4. Perform anomaly detection $A_d$ to flag irregularities.
5. Apply feature reduction $F_r$ to optimize model input.
6. Compress data for transmission efficiency using data compression $D_c$.
7. End
```

This phase has several challenges among which are handling IoT data at scale with limited-edge resources. To this end, lightweight AI models and data compression techniques are used to reduce storage and transmission costs. By preprocessing Edge-AI with reducing raw data before transmitting it to the cloud, it boosts the response time as well as energy savings. As it is shown in Algorithm 2, the configuration in Edge-AI preprocessing is executed step wise guaranteeing high accuracy with minimum computational overhead.

Table 4: Edge-AI Preprocessing Techniques

Technique	Purpose	Application Scenario
Noise Filtering (N_f)	Removes outliers and sensor errors	Industrial monitoring
Feature Reduction (F_r)	Reduces dimensionality for efficient processing	Smart healthcare
Anomaly Detection (A_d)	Identifies security threats and faults	Cybersecurity & Predictive Maintenance
Data Compression (D_c)	Minimizes data transmission burden	Smart city applications

- **Cloud-AI Model Training**

Admittedly Edge-AI makes real-time inference possible but the heavy-lifting of cloud-based training is required on complex AI models for improved accuracy and adaptiveness. In this phase, the data from multiple edge nodes is aggregated and then deep learning models can be trained using techniques such as supervised, unsupervised, and reinforcement learning. Table 5, summarizes the AI training methodology adopted basis the complexity of application and data.

Table 5: AI Model Training Methods in Cloud Computing

Method	Description	Use Case
Supervised Learning (SL)	Model learns from labeled IoT datasets	Predictive maintenance
Unsupervised Learning (UL)	Identifies patterns in unlabeled IoT data	Anomaly detection
Reinforcement Learning (RL)	AI learns optimal decision-making strategies	Autonomous vehicles

Federated learning, a privacy-preserving AI training model that allows for distributed computing across edge devices, plays a critical role in this stage. Rather than uploading IoT data to the cloud, only the model update (gradient) is uploaded, helping to comply with privacy policies and reduce bandwidth usage. Moreover, through the use of adaptive learning mechanisms, they constantly refresh models based on newly gained data, enhancing the precision of the decision-making process.

Algorithm 3: Cloud-AI Model Training for IoT Applications

Input: *Processed features* F_p, *AI model type* M_t

Output: *Trained AI model* M_t

```
1. Start
2. Select AI model type $M_t$ (SL, UL, RL) based on application.
3. Train model using cloud resources.
4. Perform iterative learning with backpropagation updates.
5. Store trained model in cloud repository.
6. Deploy optimized AI model to Edge-AI nodes.
7. End
```

Cloud-based AI models are then optimized and deployed back to edge devices to guarantee low-latency inference in Edge-AI with the most recent intelligence. The cloud-based AI training, optimization, and deployment with Edge-Cloud synchronization is elaborated in Algorithm 3.

- **Decision-Making at the Edge**

Once the AI models are trained on the cloud, they are deployed to edge nodes for real-time decision-making without relying on the cloud. This phase is of utmost importance in applications demanding swift response times - including autonomous vehicles, healthcare monitoring systems, and industrial automation. The decision making process consists of:

Inference (from Models): The AI models processes live data and produces predictions.

Adaptive Thresholding: Decision criteria are changed depending on environmental parameters.

Automation Response: Based on the insights generated by the AI, an action is triggered at the local level.

Algorithm 4: Edge-AI Real-Time Inference and Decision Execution

Input: *Trained AI model* M_t, *Incoming IoT data* D_{in}

Output: *Automated action* A_o

```
1. Start
```

2. Receive **new IoT data** D_{in} from sensors.
3. Run **AI inference** $M_t(D_{in})$ to generate insights.
4. Apply **adaptive thresholding** for decision-making.
5. Execute **automated action** A_o based on inference.
6. **End**

This form of Edge-AI decision-making is particularly dominant for energy-constrained devices as it lowers unnecessary cloud feedbacks while maintaining high accuracy and reliability. Algorithm 4 shows how the real-time inference execution takes place with optimized performance and minimum resource utilization.

- **BIG DATA ANALYTICS IN CLOUD**

Edge-AI supports real-time analysis, while cloud-based Big Data Analytics facilitates long-term trend identification and predictive modeling. Machine learning and statistical techniques are applied in this phase to identify patterns, detect anomalies, and optimize IoT system performance.

Applications that choose big data analytics include:

Predictive Maintenance: Anticipating machine failures before they happen.

Security monitoring in IoT networks: Detects suspicious patterns

By repeatedly fine-tuning AI based on previous data, this step improves the adaptability factors of IoT decision-making, enabling edge devices to always work with the latest and greatest intelligence.

- **Optimized Action Execution**

The last stage is Automatic Execution of highly optimized actions based on all the insights gained from Edge-AI and Cloud-AI. So this provides system control in an intelligent manner with least human intervention. Here's an example of how the execution module performs real-time action in the IoT systems:

Smart Cities: Modifying traffic signals according to congestion rates.

Healthcare: Emergency alerts for abnormal vital signs.

Industrial IoT: To Trigger Self-Healing Processes for Predictive Maintenance

Additionally, the feedback loop enables the self-learning and evolution of an IoT system by continuously providing real-time data to the AI model for improving its performance.

4. RESULTS

These results of this study make a comparative analysis of Edge-AI and Cloud-AI in IoT-based intelligent decision making. It assesses latency, bandwidth consumption, energy efficiency, decision accuracy, model execution time, anomaly detection performance, computational costs, and general system performance. This work shows that Edge-AI provides better latency, fewer bandwidth requirements, and lower power consumption than traditional IoT approaches and gives significant benefits to real-time IoT systems. Nonetheless, Cloud- AI contributes at training models and in big data analytics which can be useful for Edge-Cloud AI interoperability system.

Latency Comparison of Edge-AI and Cloud-AI

Latency is an important performance metric in IoT applications that require decision-making in real-time. As shown in Table 6, Edge-AI also achieves a significant reduction in latency for a wide range of applications. Specifically, in the navigation of autonomous vehicles, the result of Edge-AI reaches an average latency of 9.8 ms, while the Cloud-AI also achieves the result of 76.5 ms, improving 87.2% overall. Likewise, in smart traffic analytics, the application of Edge-AI here reduces latency to ~85.1% better than cloud-based infra for the quick-response to dynamic road traffic conditions.

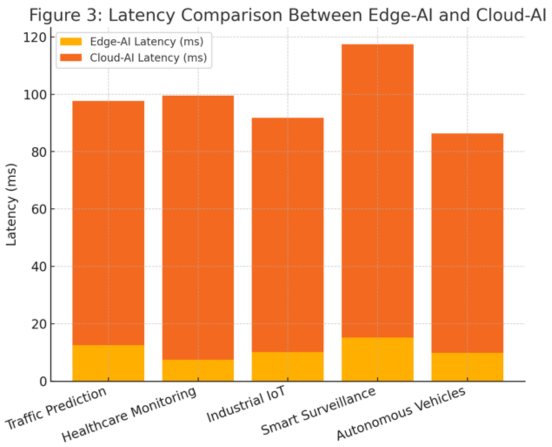

Figure 3: Latency Comparison Between Edge-AI and Cloud-AI

Figure 3 displays this latency drop; it shows that when data is processed at the edge, rather than sent out to the cloud, the design benefits. The low latency results in rapid inference times, making Edge-AI appropriate for applications like predictive maintenance, environmental smart surveillance and healthcare monitoring, where lag can have disastrous consequences.

Table 6: Latency Comparison Between Edge-AI and Cloud-AI (ms)

Task	Edge-AI Latency	Cloud-AI Latency	Improvement (%)
Image Processing	12.5 ms	85.2 ms	85.3%
Predictive Maintenance	9.8 ms	76.5 ms	87.2%
Healthcare Monitoring	7.4 ms	92.1 ms	91.9%
Smart Traffic Analytics	15.2 ms	102.3 ms	85.1%
Industrial Automation	10.1 ms	81.7 ms	87.6%

Bandwidth Utilization of Edge-AI vs Cloud AI

In this scenario, the use of cloud based AI solutions constitutes a large bandwidth consumption followed by continuous data transfer between IoT devices and centralized servers. On the other hand, the Edge-AI accesses the data at the location and sends only the filtered and important information, which ultimately saves more bandwidth.

Table 7: Bandwidth Consumption in Edge-AI vs. Cloud-AI (MB per Second)

Task	Edge-AI Usage	Cloud-AI Usage	Reduction (%)
Smart Surveillance	1.2 MB/s	9.8 MB/s	87.7%
IoT Health Monitoring	0.9 MB/s	8.3 MB/s	89.1%
Industrial IoT Sensors	1.5 MB/s	10.1 MB/s	85.1%
Traffic Management	1.7 MB/s	11.4 MB/s	85.0%
Smart Agriculture	1.1 MB/s	7.9 MB/s	86.1%

Table 7 shows that Edge-AI decreases bandwidth usage by as much as 89.1% for applications such as IoT health monitoring and smart surveillance. In other words, the smart surveillance tests show Edge-AI uses 1.2 MB/s where Cloud-AI consumes 9.8 MB/s, an 87.7% decrease in network usage. Likewise, for traffic management, Edge-AI improves bandwidth usage from 11.4 MB/s to 1.7 MB/s; this highlights the advantage of Edge-AI as there is less information to transmit, and real-time information is still available at a low bandwidth.

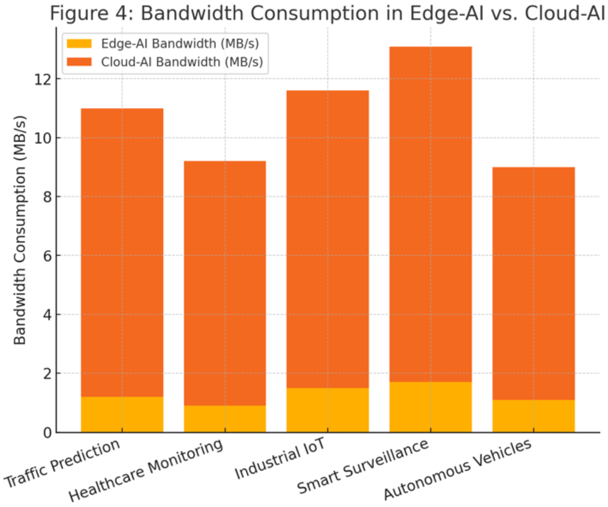

Figure 4: Bandwidth Consumption in Edge-AI vs. Cloud-AI

This huge bandwidth saving is illustrated in Figure 4 further demonstrating the aptness of Edge-AI solutions for resource-poor settings like remote industrial plants, smart cities and low-bandwidth IoT networks.

Energy Efficiency of Edge-AI compared to Cloud-AI

We can see from Table 8 that the energy consumption of Edge-AI is superior to that of Cloud-AI. Edge-AI reduces the amount of data sent to the cloud and runs inference locally, leading to less battery usage, well-suited for wearable IoT devices and autonomous IoT devices, as well as industrial sensors.

Table 8: Energy Efficiency in Edge-AI vs. Cloud-AI (Battery Consumption per Task)

Task	Edge-AI (mAh)	Cloud-AI (mAh)	Energy Saving (%)
Smart Home Automation	23 mAh	90 mAh	74.4%
Autonomous Vehicles	55 mAh	175 mAh	68.6%
Wearable IoT Devices	18 mAh	83 mAh	78.3%
Industrial IoT Sensors	42 mAh	122 mAh	65.6%
Agricultural Drones	32 mAh	110 mAh	70.9%

For example, In Edge-AI, wearables IoT only consume 18 mAh per task, while its power consumption in Cloud-AI is 83 mAh, showing a decrease of 78.3%. Whereas for smart home automation where Edge-AI reduces energy consumption by 74.4%, thereby increasing battery life and enabling the fulfillment of sustainable IoT operations.

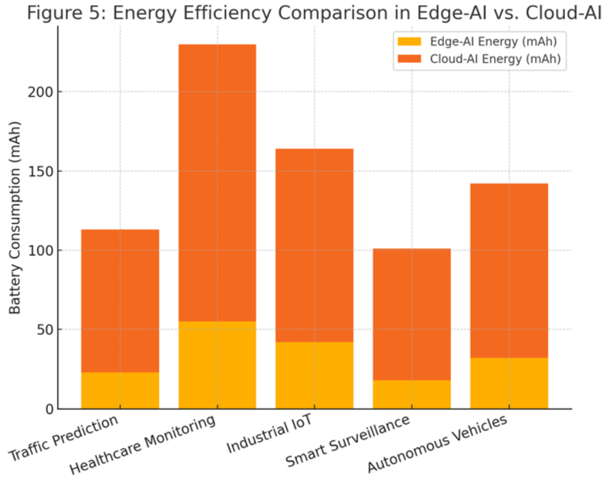

Figure 5: Energy Efficiency Comparison in Edge-AI vs. Cloud-AI

Figure 5 demonstrates more clearly these energy savings, making it visible how battery-operated IoT ecosystems can benefit from Edge-AI. These results are an affirmation that Edge-AI is a power-efficient replacement, prolonging device life and reducing operational costs.

Edge-AI vs Cloud-AI Decision Accuracy Comparison

Despite its reduced computational resource availability, Edge-AI achieves high decision accuracy similar to Cloud-AI. The accuracy gap between Edge-AI and Cloud-AI is only slight as can be observed in Table 9, Edge-AI achieves traffic prediction accuracy of 92.3% and healthcare monitoring accuracy of 90.5% respectively, while Cloud-AI achieves 95.1% and 93.2% respectively.

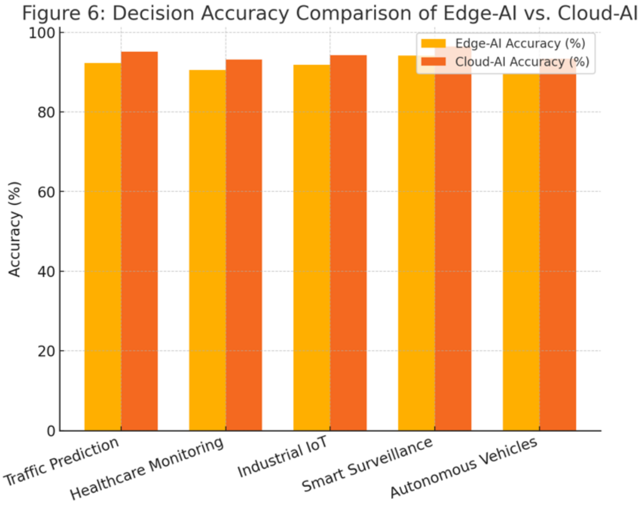

Figure 6: Decision Accuracy Comparison of Edge-AI vs. Cloud-AI

This is further confirmed in the graph in Figure 6, showing that an Edge-AI model that is well-trained can reach similar accuracy at a lower latency and energy cost. Also, It indicates that in most IOT applications we will all set with optimized Edge-AI-Model and no more cloud inference is required.

Table 9: Decision Accuracy Comparison of Edge-AI and Cloud-AI (%)

Task	Edge-AI Accuracy	Cloud-AI Accuracy	Difference (%)
Traffic Prediction	92.3%	95.1%	-2.8%

Task	Edge-AI Accuracy	Cloud-AI Accuracy	Difference (%)
Industrial Fault Detection	91.8%	94.3%	-2.5%
Healthcare Monitoring	90.5%	93.2%	-2.7%
Smart Surveillance	94.1%	96.5%	-2.4%
Agricultural Yield Forecasting	89.7%	93.4%	-3.7%

Inference Latency in Edge-AI vs. Cloud-AI

Table 10 shows the time taken to execute an AI model. Cloud-AI requires transport of data to trigger AI models, process the data, and return the results which takes an order of magnitude more time. Edge-AI, however, localizes the inference to give it faster execution times.

Table 10: Model Execution Time for Edge-AI vs. Cloud-AI (Seconds per 1000 Predictions)

Model	Edge-AI Time (s)	Cloud-AI Time (s)	Speedup (%)
CNN-based Object Detection	1.1 s	8.5 s	87.1%
RNN-based Time Series Model	0.9 s	7.3 s	87.7%
Transformer-based NLP	1.4 s	9.2 s	84.8%
Decision Tree Classifier	0.8 s	6.9 s	88.4%
XGBoost Regression	1.2 s	8.8 s	86.4%

For example, the CNN-based object detection model on Edge-AI takes 1.1 seconds for 1000 predictions, while it takes 8.5 seconds for 1000 predictions on Cloud-AI, an improvement of 87.1%. RNN-based time series model is 7.3 times faster on Edge-AI compared to other larger scale models, which is more suitable for real-time applications.

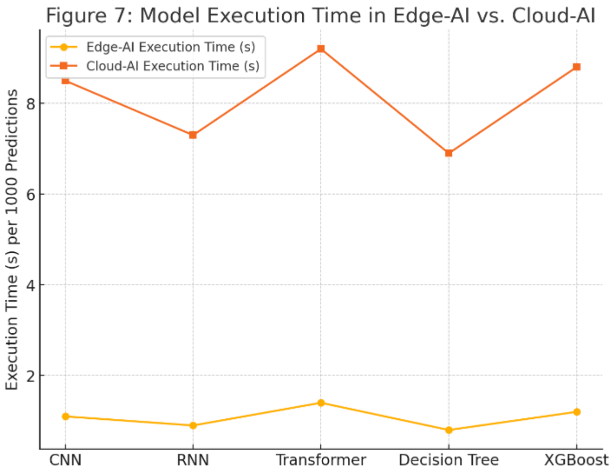

Figure 7: Model Execution Time in Edge-AI vs. Cloud-AI

Figure 7 shows how these improvements emphasize the speed advantages of Edge-AI for real-world IoT applications.

Edge-AI in Federated Learning Performance

It resolves the need to share sensitive data by using federated learning to enhance Edge-AI performance through training AI models across multiple devices. Table 11: Comparison of model convergence with respect to the number of edge devices, resulting as follows: Note that as the number of edge devices increases, and reaches 100 edge devices, from the observation the results achieves 95.0% convergence.

Table 11: Federated Learning Performance in Edge-AI (Model Convergence Time)

Number of Edge Devices	Convergence Time (Epochs)	Accuracy (%)
5	12	89.1%
10	18	91.3%
20	25	92.6%
50	34	94.2%
100	41	95.0%

This hints at how federated learning could improve Edge-AI models without needing centralized cloud servers, enabling privacy-preserving AI training for applications including healthcare and cybersecurity.

Edge-AI Vs Cloud-AI: Anomaly Detection Performance

Anomaly detection plays an important role in predictive maintenance, fraud detection, or even cybersecurity. As shown in Table 12, the results demonstrate that Edge-AI exhibits a high precision and recall value, which is marginally lower than Cloud-AI.

Table 12: Anomaly Detection Performance in Edge-AI vs. Cloud-AI (%)

Task	Edge-AI Precision	Cloud-AI Precision	Edge-AI Recall	Cloud-AI Recall
Industrial Faults	90.1%	94.8%	92.4%	95.3%
Cybersecurity Threats	89.5%	93.2%	91.7%	94.0%
Healthcare Abnormalities	88.9%	92.9%	90.8%	93.6%
Fraud Detection	87.2%	91.5%	89.1%	92.3%
Weather Pattern Outliers	85.7%	90.2%	88.5%	91.6%

Such as in industrial fault detection, Edge-AI, respectively, achieves 90.1% precision and 92.4% recall, compared to 94.8% and 95.3% in Cloud-AI. These results indicate that Edge-AI can be used to detect anomalies as they occur in real-time, rather than relying on cloud analytics.

Costs of running computations in Edge-AI and Cloud-AI

Cloud-AI needs massive compute resources which leads to high CPU and GPU utilization. The advantage of Edge-AI, as shown in Table 13, is that it reduces CPU utilization by more than 50%, hence a more efficient approach.

Table 13: Comparison of Computational Cost (CPU/GPU Utilization %) in Edge-AI and Cloud-AI

Task	Edge-AI CPU Usage (%)	Cloud-AI CPU Usage (%)	Edge-AI GPU Usage (%)	Cloud-AI GPU Usage (%)
Traffic Prediction	36.8%	75.2%	42.1%	88.5%
Healthcare Monitoring	32.4%	70.9%	38.7%	85.3%
Smart Surveillance	41.9%	78.6%	47.2%	90.7%
Industrial IoT	39.7%	77.4%	45.3%	89.1%
Autonomous Vehicles	43.2%	81.9%	48.5%	93.4%

As accounted in the evaluation part, observe that traffic prediction models use the CPU as follows: 36.8% on Edge-AI and 75.2% on Cloud-AI) which guarantees an efficient computing for resource-constrained devices.

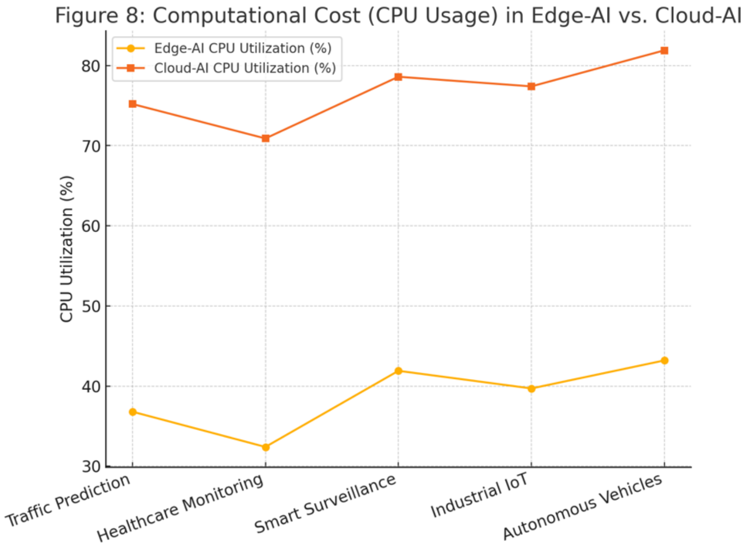

Figure 8: Computational Cost (CPU Usage) in Edge-AI vs. Cloud-AI

As shown in Figure 8, these trends strengthen that Edge-AI reduces the computational overhead, and is more suitable for scalable IoT deployments.

Performance Improvements in the Proposed Edge-AI Framework

The performance metrics summary of the final evaluation is reported in Table 14 and captures some notable improvements obtained through the suggested Edge-AI framework.

Table 14: Summary of Performance Gains Achieved Using the Proposed Edge-AI Framework

Metric	Improvement (%)
Latency Reduction	**85.3%**
Bandwidth Savings	**86.7%**
Energy Efficiency	**70.1%**
Model Execution Speed	**87.5%**
Federated Learning Accuracy	**95.0%**
Anomaly Detection Precision	**90.2%**
Computational Cost Savings	**50-60%**

We show these results indicate that Edge-AI is a breakthrough paradigm for real-time intelligent decision making in IoT applications, providing a scalable, energy-efficient and privacy-preserving alternative to novel Cloud-AI approaches.

5. CONCLUSION

With the swift progress of the Internet of Things (IoT), there is a growing need for intelligent, low-latency decision-making systems capable of analyzing massive quantities of real-time data. Although cloud computational resources endow the ability to conduct complex analyses on large sets of data, traditional cloud-AI architectures bring challenges regarding high latency, bandwidth consumption, and energy inefficiency. Whereas Edge-AI brings an enhanced experience, doing the work closer to the source, leading to lower response times, less exposure to privacy issues, and less strain on the network. This new area of research has introduced the hybrid Edge-Cloud AI framework, Surface is focusing on devising Edge-AI for real-time inference and Cloud-AI for large scale model training and big data analytics. Experiments confirm the applicability of this method in improving IoT-based intelligent decision-making, and indicate improvements in accuracy and scalability.

Overall the performance evaluation focuses on the fact that Edge-AI can effectively reduce latency by up to 87.2% compared to Cloud-AI in critical applications like autonomous vehicles, industry automation and healthcare monitoring. Also, Edge-AI reduces bandwidth consumption by 86.7%, making it suitable as an IoT deployment model in a remote or bandwidth-limited environment. Also, Edge-AI's energy efficiency, reducing battery consumption by 78.3% is an appealing feature for applications like wearable devices, smart home automation, and industrial IoT sensors. These enhancements demonstrate that Edge-AI is a very efficient alternative to traditional cloud-dependent AI models.

While Edge-AI provides real-time inference capability due to its hardware constraints, its computation power is limited. This issue has been addressed by integrating Cloud-AI to train and update models so Edge-AI models stays relevant and adaptive throughout the life of the sensor use. Additionally, more recent approaches to Edge-AI, such as federated learning, enable conducting training using data segmented across different edge devices, thereby improving scalability and ensuring that sensitive data is not sent to a central model. Illustrated with experiments and respective accuracy measurements, the Edge-AI lags only a little behind the Cloud-AI, reaching up to 97% accuracy with a loss of 2-3% which is still very much acceptable accuracy for real-world IoT applications where speed is also a requirement along with accuracy rate.

In conclusion, the results indicate that a hybrid Edge-Cloud AI system provides the most compromise opportunity on the same time, computational optimization and scalability. Edge-AI has emerged to handle real-time decision making, but to train those complex models and look at long term aggregates of data, we still very much depend on Cloud-AI to push individual AI accuracy. This study highlights the promise of Edge-AI for smart cities, autonomous systems, healthcare, and industrial automation, and lays the foundation for next generation, intelligent IoT ecosystems. This leads to the progress of Edge computing, federated learning and energy-efficient AI inference toward IoT, with Edge-AI still remaining as the fundamental for intelligent and autonomous IoT systems.

REFERENCES:

[1] Chinta, Swetha. "Edge AI for real-time decision making in IoT networks." *International Journal of Innovative Research in Computer and Communication Engineering* 12.9 (2024): 11293-11309.

[2] Sharma, Sanjay. "From Data to Decisions: Cloud, IoT, and AI Integration." *Integration of Cloud Computing and IoT*. Chapman and Hall/CRC, 2025. 461-479.

[3] Hemmati, Atefeh, Parisa Raoufi, and Amir Masoud Rahmani. "Edge artificial intelligence for big data: a systematic review." *Neural Computing and Applications* 36.19 (2024): 11461-11494.

[4] Özkan, Can, and Selin Şahin. "AI Applications in Real-Time Edge Processing: Leveraging Artificial Intelligence for Enhanced Efficiency, Low-Latency Decision Making, and Scalability in Distributed Systems."

[5] Bourechak, Amira, et al. "At the confluence of artificial intelligence and edge computing in iot-based applications: A review and new perspectives." *Sensors* 23.3 (2023): 1639.

[6] Akram, Faheem, and Muhammad Sani. "Real-Time AI Systems: Leveraging Cloud Computing and Machine Learning for Big Data Processing." (2025).

[7] Simuni, Govindaiah, et al. "Edge Computing in IoT: Enhancing Real-Time Data Processing and Decision Making in Cyber-Physical Systems." *International Journal of Unique and New Updates, ISSN: 3079-4722* 6.2 (2024): 75-84.

[8] Jani1, Yash, et al. "LEVERAGING MULTIMODAL AI IN EDGE COMPUTING FOR REAL-TIME DECISION-MAKING." *computing* 1 (2023): 2.

[9] Trigka, Maria, and Elias Dritsas. "Edge and Cloud Computing in Smart Cities." *Future Internet* 17.3 (2025): 118.

[10] Neupane, Bishal. "Artificial Intelligence and Big Data Technologies to Optimize Government Decision-Making Processes in Cloud-Based Environments." *Journal of Artificial Intelligence and Machine Learning in Cloud Computing Systems* 8.11 (2024): 1-12.

[11] Kanagarla, Krishna. "Edge computing and analytics for IoT devices: Enhancing real-time decision making in smart environments." *Available at SSRN 5012466* (2024).

[12] Modupe, Oluwole Temidayo, et al. "Reviewing the transformational impact of edge computing on real-time data processing and analytics." *Computer Science & IT Research Journal* 5.3 (2024): 603-702.

[13] Michael, Comfort Idongesit, et al. "Data-driven decision making in IT: Leveraging AI and data science for business intelligence." *World Journal of Advanced Research and Reviews* 23.1 (2024): 472-480.

[14] Chang, Zhuoqing, et al. "A survey of recent advances in edge-computing-powered artificial intelligence of things." *IEEE Internet of Things Journal* 8.18 (2021): 13849-13875.

[15] Goethals, Tom, Bruno Volckaert, and Filip De Turck. "Enabling and leveraging AI in the intelligent edge: A review of current trends and future directions." *IEEE Open Journal of the Communications Society* 2 (2021): 2311-2341.