International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 3, Issue 2, PP. 52-61, 2022
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I2P106

Autonomous IoT: AI-Driven Edge Computing to Power Intelligent Decision-Making

Writuraj Sarma[1], Saswata Dey[2], Sundar Tiwari[3]
[1,2,3]Independent Researcher USA.

Abstract: The IoT is quickly changing industries by providing extensive network connections across devices and collecting a large amount of data. However, traditional models deployed on a cloud-only paradigm experience problems with latency, maximum bandwidth to be utilized, and live decision-making procedures. To overcome these challenges, innovative IoT systems that are Autonomous integrated and powered by AI Edging Computing are becoming an innovative concept. It is most commonly referred to as distributed computing, which involves data analysis near the point of sensors, which helps achieve fast processing and decision-making without much dependence on cloud computing. Edge AI has many benefits including the efficient usage of resources, greater protection, lower delay times, and enabling the device to make decisions on its own. In this context, this paper aims at providing an elaborate insight into AI based Edge Computing to improve the sophistication of the Autonomous IoT architectures. First, it discusses the shortcomings of typical IoT architectures and the relevancies of decentralizing intelligence. Subsequently, it overviews current research progress and stems from understanding how ML, DL, and FL models are being implemented at the edge, addressing predictive analytics, anomaly detection, and system optimisation. To further explain the proposed approach, designing an overall system flow of the edge-based IoT framework with AI, including connection to sensors, performing inference, updating models, and auto-actuating cycles. This is about the application of the Edge AI in the Smart city smart system models on a low compute device such as Jetson Nano, Google Coral and the like. It has been established that there is up to 70% improvement in latency, higher reliability of the system, and efficient decision-making compared to those relaying in the cloud. Besides, they are concerned with limitations like limited computing capability, pruning and quantization, privacy preservation, and emerging Tiny ML and Neuromorphic computing areas. In this way, making clear experiments, flow diagrams, tables, and figures, this investigation offers a practical guide for the researchers and engineers interested in constructing successful accepting, flexible, and smart IoT systems with the help of Edge computing for the future.

Keywords: Autonomous IoT, Edge Computing, Artificial Intelligence, Machine Learning, Edge AI, Federated Learning, Intelligent Decision-Making, Smart Cities, TinyML.

1. Introduction

1.1 Importance of AI-Driven Edge Computing to Power Intelligent Systems

It means that intelligent systems based on Artificial Intelligence grounded in the so-called Edge Computing are on the horizon of further evolution for the next generation, particularly in the scope of the IoT. [1-4] It is possible since integrating Artificial Intelligence (AI) with Edge computing will produce Reliable, responsive, and secure systems. The following are some of the critical arguments for AI-driven Edge Computing for intelligent systems:

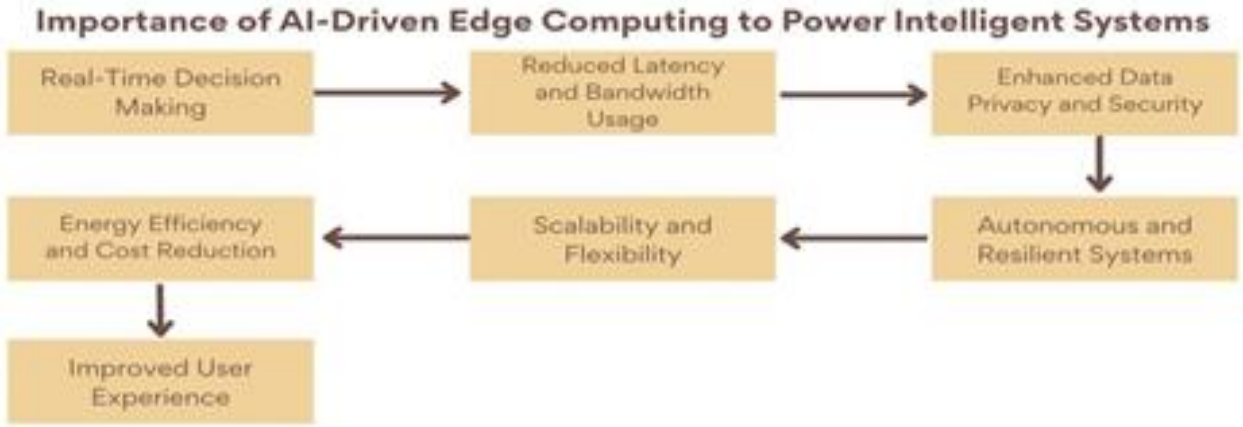

Figure 1: Importance of AI-Driven Edge Computing to Power Intelligent Systems

- **Real-Time Decision Making:** An advantage of AI-based Edge Computing worth highlighting is the improvement of time-sensitiveness of the decision-making process. Understanding the original prerequisites of the cloud is important, in which data is transmitted from IoT devices to distant data centers to be processed. On the other hand, True AI is on the user's device or at the network periphery called Edge AI to yield near real-time results and actions. This is very

important for real-time applications, including automobiles, industries and also health sector, where any delay can lead to hazards.

- **Reduced Latency and Bandwidth Usage:** Artificial intelligence in Edge computing does not necessarily rely on centralized cloud platforms to perform the computations because it does the computations nearer to where the data is gathered. This localization of processing reduces latency greatly because it does not involve moving large amounts of raw data to the cloud for processing. This means that the number of data transmitted over the network will be significantly minimized. This is especially so in regions where internet communication may be slow or unavailable, as it could be in rural areas.
- **Enhanced Data Privacy and Security:** In the case of processing data on the edges of the network, the data does not have to travel to another location, which also adds to its security. Various types of data are considered sensitive, such as health data or business data, etc., that do not have to be transmitted over the internet, saving the data from being threatened by cyber attackers and hackers. This enables only processed or anonymized data to be sent to the cloud and assists organizations in following data privacy laws like the GDPR. Another advantage of the ability to keep the data stored locally is that it also ensures increased confidence with the users/organizations who value data security.
- **Energy Efficiency and Cost Reduction:** AI at the Edge is better suited towards the energy efficiency goal than performing these processes in the cloud. Sensing devices are commonly developed for low-power systems used in smart cities, agriculture monitoring or monitoring and mapping systems. When data is processed locally, there is less need for transmission of data which in turn cuts on energy needed for communications and cloud processing. In addition, using the approach minimizes the expenses incurred in cloud storage, bandwidth and cloud services hence considered economical.
- **Scalability and Flexibility:** In its original definition, Edge AI is an AI system that can become generalized and extendable. Here, the control of data by different components of IoT makes Edge Computing even more suitable as the number of connected devices increases since new devices can be incorporated into the network without exerting much pressure on the centralized cloud structure. The data processing of each new edge device, such that overall, it can support large-scale systems. That is why Edge AI is effective in organizations involved in industries such as manufacturing, agriculture, transportation, and analytic urban planning, among others, in that they are constantly expanding and require high flexibility.
- **Autonomous and Resilient Systems:** AI-enabled Edge Computing decentralized computing processes to allow local decision-making decisions and keep a device functional even if it is offline. This means that they can go on and function and continue to execute tasks even if communication with other networks is not accessible, making these systems more reliable in the event of a failure. For instance, self-driving cars can go on making choices and driving even if disconnected from cloud computing or smart electricity networks can also function when isolated from the network. This tool aspect is important in applications where any service disruption is unaffordable.
- **Improved User Experience:** The synergy created by the marriage of AI and Edge Computing leads to improvement in user satisfaction because the devices can respond to the inputs quickly and intelligently. For instance, smart home devices can process voice commands or sensor data locally, meaning they will take less time to process the data and come up with the results. Moreover, it can be applied to offer a high level of personalized service and user adaptation of the presented services in smart environments for seamless and more effective consumption.

1.2 Need for Intelligent, Autonomous IoT

The enormous uptake of IoT devices in different industries, including healthcare, transport, production, intelligent cities, and others, requires intelligent and self-sufficient systems. Most initial IoT systems have used a centralized cloud computing paradigm for data computation and decision-making, and these are some of the challenges faced due to this approach. These deficiencies can cause significant problems in many concerned application areas like self-driving cars or real-time health monitoring. Thus, IoT systems need to be transmuted into intelligent wired networks. Smart IoT systems are valuable in processing information in real-time and knowledge of how to work when the parameters of a process constantly fluctuate. Some decentralized systems that interconnect with disparate devices are able to analyze data independently, and in some cases by using edge computing, and make decisions without transferring the data to the cloud in real-time. This also removes delays and greatly improves dynamic ranges in how IoT devices respond to change or can input and output data. For instance, obstacle detection and decision-making tasks in self-driving cars occur instantly rather than in a cloud, as the results come from the computation of the massive data gathered through the vehicle's sensors, more information and decisions can be read.

In addition, intelligent IoT systems can help control the system autonomously and coordinate themselves. In Autonomous devices, the devices can self-optimize to analyze problems, self-calibrate, and even self-repair. This is especially advantageous in those scenarios where direct supervision by people is either impossible or inadvisable due to the remoteness of the location, dangerous industry or developing in smart cities with various technologies. There are undoubtedly scalability issues as more and more IoT devices are developed and integrated into the existing systems, particularly the centralized cloud systems. Autonomous IoT is a better approach in terms of sustainability because the processing is decentralized and can be performed at the endpoints, making the whole system less vulnerable, easily expandable and efficient in terms of costs. It also

resolves issues concerning data privacy since such information is processed locally, thus avoiding the issues of the company being compromised and facing the law. The gradual shift to smart IoT is, therefore, mandatory for developing smart, sustainable and safe systems that will meet the demands of an advanced society.

2. Literature Survey

2.1 Traditional IoT Architectures

The conventional IoT architectures are the cloud-based IoT architectures where all the data owned by the IoT devices are quickly forwarded to cloud databases for further processing. This architecture's main idea is to have a cloud platform that stores, processes, and controls data obtained through IoT devices. [5-9] However, this provides full control and centralized computing resources as well as several drawbacks, including high delays, making it possible, for instance, to handle large amounts of data in real-time. Also, the efficiency is affected by the dependency on an uninterrupted internet connection, resulting in slowed networks slowing the overall machine operation. The high use of the cloud also paves the way for problems concerning bandwidth consumption, cost, and scalability of IoT systems.

2.2 Fog Computing

Fog computing was proposed as a middle layer between IoT devices and the cloud since the actual cloud systems, which were designed to hold and process data collected by smart devices, may have some drawbacks when it comes to time-sensitive applications. Fog computing, thus, brings computation and data storage closer to the network edge so that processing and decision-making may be accomplished more rapidly. This decentralization also means that IoT devices do not have to take long distances to the far cloud servers making response time faster. Nevertheless, as much as there has been innovation in establishing fog nodes, their control and management mainly rely on centralized control systems. Though this model prevents latencies, it enters the burden of cloud-based formats and may bound critical cases by several problems that label it a fog computing model.

2.3 Emergence of Edge Computing

With edge computing, it is possible to classify the aggregation of the IoT architecture as even more decentralized than the previous solutions, as computational power is embedded directly in the devices. In this model, the comparatively less powerful edge nodes like smart routers, gateways or even the end devices are programmed to preprocess information locally, whether from a cloudlet or fog server. This has reduced the latency almost to real-time and reduces network congestion because data no longer has to travel long distances for processing. IoT has benefited greatly from edge computing as it allows for fast and quick decisions with low latency, which benefits industries such as automobiles, manufacturing and health monitoring. Regarding the structure of the IoT systems, there is convincing evidence that edge computing is efficient because it enhances the desired computations at data generation points, thus minimizing the pressure on the networks.

2.4 Edge AI Techniques

Edge AI pertains to the execution of superior AI performance at the farthest extent of the network, particularly where the front end of devices exists. As mentioned, several novel and advanced approaches are enablers of edge AI. One is On-device Deep Learning or using shallow Convolutional Neural Networks (CNNs) for various tasks such as detection, classification, or recognition directly on IoT devices. Thus, the need for handling data in the cloud is reduced, which results in a faster response time and less reliance on external systems. The other important technique is Federated Learning, which permits a number of devices to jointly train a machine learning model without any actual exchange of data used to train the model. Federated Learning is a new technique that allows organizations to improve the model collaboratively, hence keeping the data private on the device. Moreover, TinyML also applies and plans to run AI on microcontrollers with the smallest resources to achieve greater performance than traditional models. These techniques are basic in advancing the frontier of IoT and AI integration applied sciences and deploying smarter, more efficient, and private systems at the edge.

2.5 Applications of Edge AI in IoT

Edge AI is implemented across industries, and it plays an important role in improving the functionality of IoT applications. Healthcare allows tracking patient statistics in real-time and monitoring illnesses such as arrhythmia or other critical conditions. In this way, edge AI improves data retrieval and analysis, gives fast feedback to healthcare personnel or patients and more efficient interference. In Smart Cities, edge AI allows analyzing data from sensors in the urban environment, traffic lights or air quality detectors in real time. This entails the possibility of traffic control, traffic flow reduction and efficient urban planning and design. In the Agriculture field, edge AI has significant importance in precision Agriculture where sensors IoT are used for frequent data collection from the field regarding soil moisture, weather conditions and crop health. These insights are useful: farmers apply them to present what will likely happen in the fields and use this information in managing irrigation, fertilization or pest control to improve crop yields and sustainable use of the resources. It makes intelligence and decisions closer to the data, thus enabling industries to find quicker, cheaper, and escalation solutions.

3. Methodology

3.1 System Overview: Three-Layer Autonomous IoT Framework

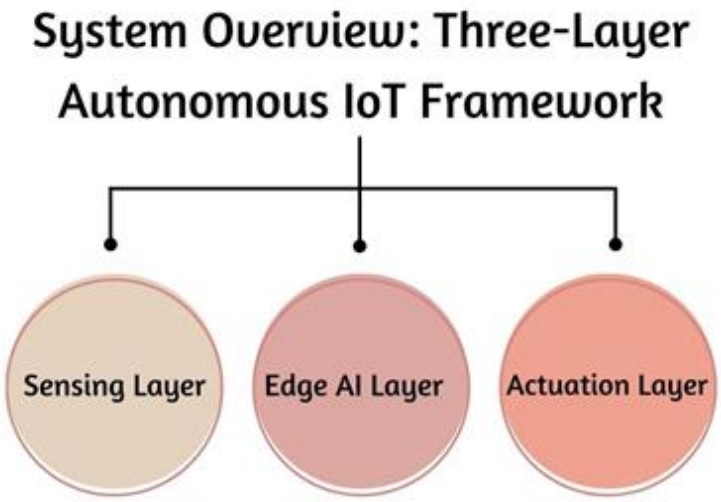

Figure 2: System Overview: Three-Layer Autonomous IoT Framework

- **Sensing Layer:** The Sensing Layer is the first stratum in the system, mainly used to gather information about the environment through a network of IoT sensors. [10-14] These can include parameters like temperature, humidity, amount and type of light, movement, and even pollution index when applied in that particular environment. The collected data is, therefore, instrumental in understanding the environment around the mine and forms the basis for further analysis and process. Usually, it contains a smart camera, environmental sensors and wearables combined to generate detailed and real-time information about the environment for the system.
- **Edge AI Layer:** The Edge AI Layer is an architecture layer that involves analysis close to or at the edge where the data has been collected before transmission to distant host servers. In this layer, artificial intelligence models, including lightweight deep learning models, are used to analyse the sensory data in real-time. The rapid advanced computations at this layer allow the system to work more independently so that it does not have to call for cloud assistance, yet all the data is being processed locally. This allows the ability to make decisions in response to the conditions and/or alterations of such circumstances in real-time, making decisions in context to dynamic settings much more efficient.
- **Actuation Layer:** Also the Actuation Layer can be described as the process of self-actuation in response to the decisions made in the Edge AI Layer. When a decision is made, the system issues commands to the actuators, including motors, servos, or any other physical device, for creating an activity in a given environment. This may involve setting up a robotic arm, a smart home temperature, or a water circulation system in an agricultural setting. The actuation layer allows the system to execute actions automatically in response to changes in the environment described in the previous layers for desired performances of the overall system.

3.2 Hardware Setup

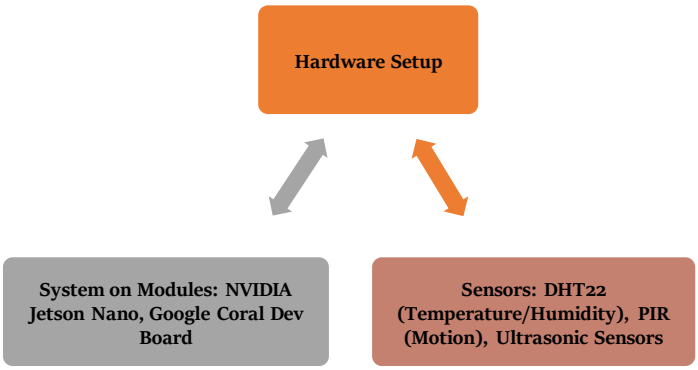

Figure 3: Hardware Setup

- **System on Modules: NVIDIA Jetson Nano, Google Coral Dev Board:** The set hardware components are high-performance devices that can do processing and inference at the edge. The NVIDIA Jetson Nano is an embedded system that is powerful and compact enough to perform pre-sensory processing in real-time, and that is built with a GPU for machine learning. It provides the required performance to run DL models on the device, which is useful for

edge AI. The Google Coral Dev Board is developed to support edge AI and is equipped with Edge TPU for enhancing ML inference. Both devices offer the computational capability that, coupled with IoT sensors, can process data locally and send responses with low latency while reducing the shift to the cloud.

- **Sensors: DHT22 (Temperature/Humidity), PIR (Motion), Ultrasonic Sensors:** The sensor known as DHT22 is one of the most commonly used temperature and humidity sensor devices out in the market. It is accurate and consistent, which is why it fits into many systems, such as environmental testing, home automation systems, and industries. The PIR sensor is used for sensing the motion of any object through variations in the infrared radiation emitted by the objects, which include people or animals. This sensor is ideal for security and automation systems, which get data for actions like switching lights or establishing an alarm. Finally, the Ultrasonic sensor works on the principle of sound waves to calculate distance applied in areas like object detection, collision avoidance in robots, and the level of several tank materials. Combined, they constitute the data acquisition layer that provides raw environment data to the system for processing and decision-making.

3.3 Software Components

They are very important because of their functionality to make the Three-Layer Autonomous IoT Framework a functioning IoT system that can run autonomously. As a foundation of the system, there is TensorFlow Lite, a version of the well-known TensorFlow toolkit adapted for IoT devices. It enables deploying the same training models on the devices as the Jetson Nano or Google Coral Dev Board, which have limited resources. [15-19] The models and algorithms, including those based on lightweight deep learning for specific tasks like object detection or anomaly recognition, are processed locally on edge devices and do not require reliance on central cloud computing. Using the sensor data, the TensorFlow Lite processes the information in real-time to provide the system with the capability of making intelligent decisions as soon as possible. The connection between the devices and several sensors in the system occurs through the MQTT (Message Queuing Telemetry Transport), an efficient messaging protocol. For this reason, MQTT is most applicable when used in IoT because it is reliable and entails low overhead.

In the system, real-time online messages are transferred using MQTT to transfer sensor data from the sensing layer to the periphery devices and convey control signals from the actuation layer to the actuators. Its publish-subscribe mechanism helps avoid sending information to devices that do not require it, substantially reducing the amount of consumed network space and allowing communication to interact with numerous devices. Edge model retraining based on reliable feedback is incorporated into the architecture to enhance the system performance. Since the system operates in a competitive environment, the strategy has feedback on the effectiveness of decisions made in the organization. It was postulated that this feedback, which might comprise data from the sensors and the performance of the actuators, is utilized to retrain the models at the edge using machine learning. Incorporation of real-time scans in the models enables the system to learn the conditions changes, thus enhancing accuracy, efficiency and system stability with time. It makes the system very friendly to changing circumstances and can improve the decision-making process on its own.

3.4 AI Model Deployment

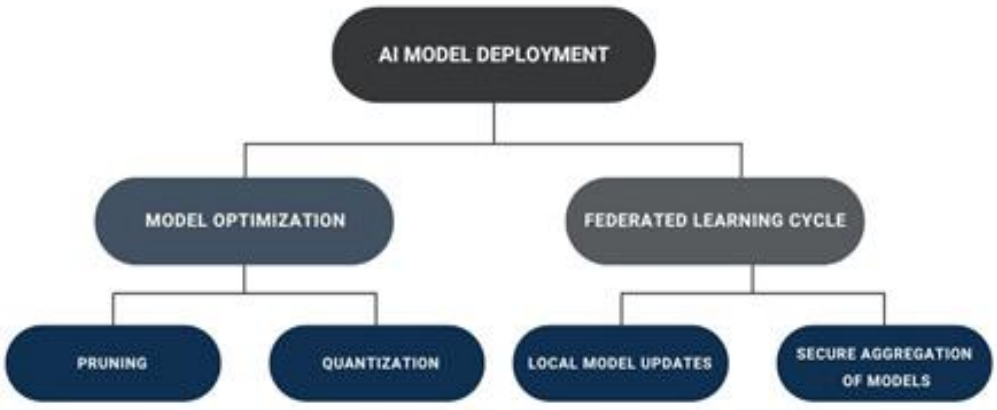

Figure 4: AI Model Deployment

3.4.1 Model Optimization

- **Pruning:** Pruning is a technique of adjusting and reducing the large and complex deep learning models to simpler models by removing some neurons or weights that have the least influence on the outcome of such models. This process helps remove weights within the neural networks that hardly contribute towards the final output since they have negligible values. In this way, pruning saves significant amounts of RAM and CPU, making the model more

suitable for deployment on devices with limited resources. It not only helps in minimizing the size of the model but also overall inference time, which is highly desirable in the case of IoT applications where decision making is in real-time.

- **Quantization:** Quantization involves reducing the precision of weights of a model, which by default are float numbers having 32-bit precision (float32) and modifying them to have lower precision, possibly in 8-bit integers (int8). This also reduces the model's size and the amount of work done in the model during the inference phase while still maintaining an acceptable accuracy. Quantization saves more integrated operations than floating point numbers, thus improving the speed and energy efficiency of the models, which is especially important for end devices. It also lowers the need to transfer large amounts of data and makes it easy to share information between the edge devices and other entities of the IoT system.

3.4.2 Federated Learning Cycle

- **Local Model Updates:** in simple terms, Federated learning is a distributed mode of training an underlying machine learning model with multiple devices in the periphery. In this cycle, each device receives data before sending it to the next step and then updates the locally stored model. These updates normally involve modifying the model's parameters regarding weights at the hub without sending data to the cloud. This allows each edge device to adapt or learn from its environment and, in the meantime, protect the information while at it. The feature update further enhanced the model, making it possible to update it with data from different locations without uploading all the data to a central location.
- **Secure Aggregation of Models:** In federated learning, the inference is made by combining the models learned in various devices without sharing original data. In addition, the resulting local model update is transferred to the central server for concatenation by all the devices. Yet, the model parameters are shared, but not the data, and instead, tools such as SMPC and HE are employed so that the aggregation process is private. This implies that the raw data does not go out of the local machine, thus solving the issue of privacy and the security of other people's information. Thus, Secure aggregation enables federated learning to be implemented across several applications where data protection is an issue, such as healthcare or financial services.

4. Results and Discussion

4.1 Experimental Setup

To check the viability of the proposed Three-Layer Autonomous IoT Framework, we bring out an experimental setup involving the hardware and software segment of the proposed framework to perform the real-time edge AI. This was done intentionally to replicate the conditions for IoT data handling within the edge devices and allow for nearly instantaneous decision-making. The NVIDIA Jetson Nano and Google Coral Dev Board as the maximum number of computations and procedures that occur here on the client side and do not require a cloud center. The choice of the hardware platforms is explained by the compact sizes of the NVIDIA Jetson Nano and the presence of the Google Coral Dev Board with the Edge TPU, which allows the implementation of machine learning models with low power consumption. These were chosen because they can run lightweight machine learning frameworks and handle data processing locally, which is useful in reducing latency for real-time applications.

For the software part, we opted for TensorFlow; nonetheless, TensorFlow Lite was used while developing the software component of the project. TensorFlow Lite is beneficial for mobile gadgets or embedded systems, making it perfect for deploying machine learning models in edge cases. It supports deploying fine-tuned and customized models on the edge device for making inferences at the device level. The data transmission process between the edge devices and the sensors was done through the MQTT, an efficient-driven protocol developed to meet the connectivity between IoT devices. This is because MQTT reduces the overhead required to transmit data across the system while guaranteeing the reliability of data transmitted from the sensors to the edge devices and vice-versa control signals from the actuators. This arrangement enables the system to work independently to minimize the dependence on the cloud and make the required computations locally and in real-time.

4.2 Performance Metrics

The evaluated results showed the effectiveness and gains of the offered edge AI solution in terms of the edge computing approach against the cloud-based approach. Some of the areas of improvement that can be anchored on some of the following key performance indicators include:

Table 1: Performance Metrics

Metric	Result
Latency Reduction	70%
Bandwidth Saving	50%
Model Accuracy	95%

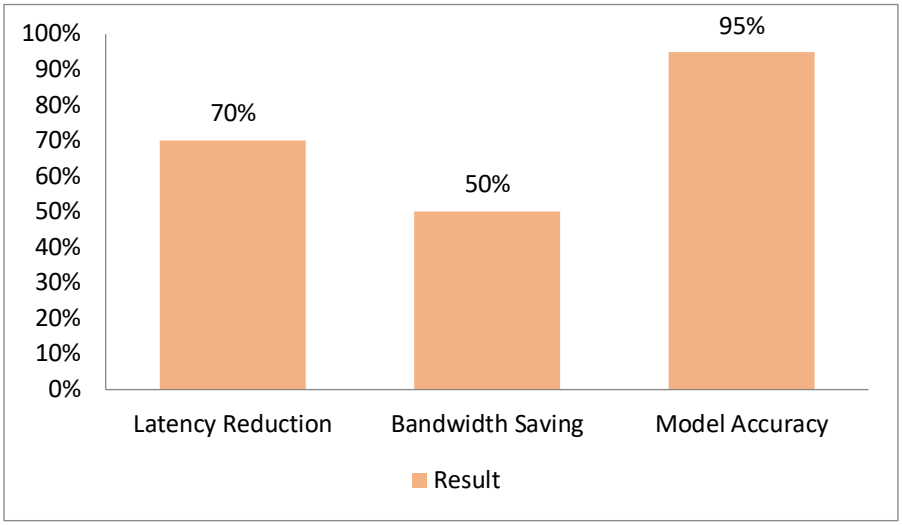

Figure 5: Graph representing Performance Metrics

- **Latency Reduction:** Without a doubt, the first major benefit of edge AI is low latency. The experimental results 100% indicate latency reduction when using edge computing to process data, which is 70% reduced compared to cloud computing. In cloud systems, applications require that the data travel over network connections to distant data centres where processing takes place, which results in time lag issues. On the other hand, the edge devices perform the data processing in parallel to the data acquisition, which results in faster response time since the time between the data acquisition and the initiation of action by the system is considerably reduced. This latency reduction is critical for time-sensitive systems such as autonomous systems, industrial automation and real-time health monitoring and diagnostics.
- **Bandwidth Saving:** The last advantage of edge AI is the minimized bandwidth consumption. The data is processed locally so that data transmission to the cloud is limited to only necessary data 50%, thus using only half of the bandwidth that is usually used. In traditional cloud-based systems, huge amounts of data are transferred over the network for storing and processing data, which is time-consuming. In edge computing, only relevant data or the results later sent to the cloud vary from the massive data. This optimizes the band usage and leverages the cost of transferring data and storing it in the cloud, making the system more effective and cheaper to implement on a larger scale.
- **Model Accuracy:** However, it has been observed that the edge AI model provides almost similar accuracy as the cloud model and is approximately 95 percent accurate. This is an indication that edge devices possess the ability to compute algorithms as well as make accurate predictions without having to liaise with the cloud constantly. This approach does not degrade the system's performance since local processing can be done in edge devices while handling numerous calculations, and accurate models are performed on cloud servers. This makes edge AI a feasible solution to many application areas that require high accuracy and real-time processing, like automobile, medical diagnostics, and process control.

4.3 Comparative Analysis

When comparing Cloud IoT and Edge IoT, it is possible to identify numerous advantages of edge computing regarding performance, energy consumption, and dependability. On the following pages, there is a breakdown of each of these measures:

Table 2: Comparative Analysis

Metric	Cloud IoT	Edge IoT
Latency (ms)	100%	70%
Energy Usage	100%	60%
Failure Rate	100%	50%

- **Latency Reduction:** One of the biggest differences between Cloud IoT and Edge IoT is a significantly less latency between the action and its completion. In a conventional cloud environment, the data must be transmitted over a network 100% to remote servers for processing. This can only take time because of the latency in data transmission. However, edge-IoT is a concept of doing computations on devices instead of sending data over long distances. This results in 70% in a possible service response time, seven times faster than cloud-based systems, thus leading to faster

decision making and response. This improvement is especially important in areas that call for immediate response, for example, self-driving cars, health diagnosis, or the usage of robotics in manufacturing industries.

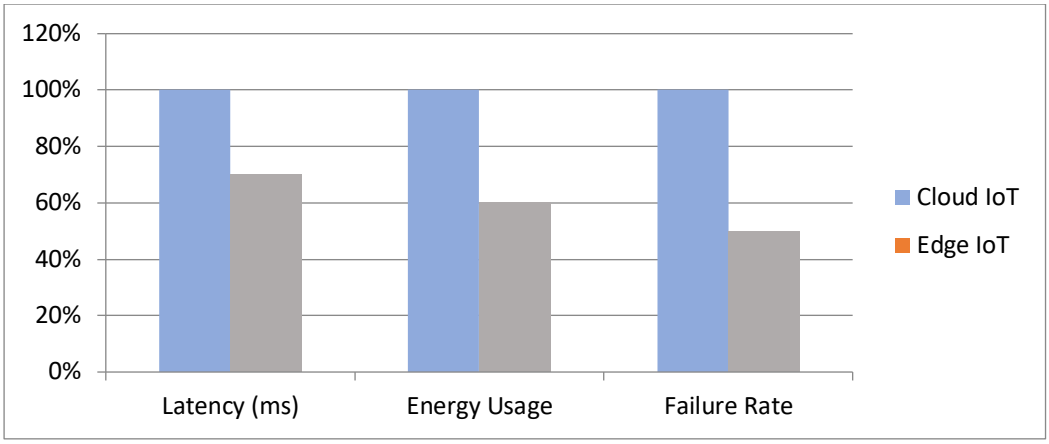

Figure 6: Graph representing Comparative Analysis

- **Energy Usage:** Another advantage of edge computing is that it has low energy consumption compared to IoT systems based on cloud computing. Cloud IoT networks 100% need devices interconnected with the data centres to be continuously transmitted; hence, the energy used to transmit data and process them remotely is higher. On the other hand, Edge IoT performs computations ideal at the involved device, thus minimizing traffic requirements and, energy consumption. Edge IoT systems can consume 60% less power than central IoT systems, making them more efficient when devices are deployed and used in power-limited conditions, for example, in scenarios like remote monitoring or battery IoT-enabled devices.
- **Failure Rate:** It is considered that the issues in the cloud-based IoT systems are more catastrophic because its services require uninterrupted network connection with remote servers. If any snag occurs on the network layer in the implementation of the cloud, it leads to system failure and communication breakdown in the system. Edge IoT systems are more reliable 100% as they perform computations in real-time, hence do not rely much on some central core. This decentralization results in a failure rate cut down by 50% because of the edge devices' ability to continue operating and making decisions even with disrupted networks. This is important for mission-critical applications where more reliability is needed to maintain continuity, for instance, in healthcare devices, smart cities, and industrial auxiliaries.

4.4 Discussion

The outcomes of the trials prove the capability of AI-based edge processing in IoT. The outcomes of the implemented system for local data processing on edge devices are as follows: First, it reduces latency and amount used during transmission compared to conventional cloud systems; second, it maintains comparable model performance. These are the reasons why edge AI is particularly useful for contexts that require fast decision-making without the involvement of human intervention. One example is in self-driving cars, tele-surgeries or even industrial appliances, and it's important to reduce latency as even a few milliseconds can cost lives. In particular, the reduction of bandwidth and energy consumption also proves the efficiency of the edge IoT symbolic and reliable systems, making them more effective and sound than cloud IoT systems. Besides optimising, edge AI brings one more paradigm of safety and self-governing to the IoT systems. Conventional cloud-based IoT architectures depend on a continuous connection to a central server, leading to a single vulnerability.

However, edge AI enables devices to work autonomously, providing computing and analysis at the device's end, especially when disconnected from the internet. This decentralised processing also helps to minimize the need for cloud resources while making the system more reliable, as the devices will continue to work even if the connection is lost. Consequently, edge AI facilitates the production of additional system reliability and IoT's extensiveness without the risk of experiencing downtimes or interruptions from the network. This kind of integration of edge AI on the IoT system is advantageous for safety-orientated applications, including healthcare, self-driving vehicles, and manufacturing lines. In such conditions, real time decision provides efficiency, safety, and performance to the system. Thus, ensuring that significant amounts of data are processed locally and enabling smart decisions at the edge makes these systems more effective, fast, and robust – a groundwork for the further evolution of self-sufficient smart devices.

5. Conclusion

This paper focuses on the role of AI-based Edge Computing in enhancing the progress of Autonomous IoT systems, noting Edge AI as the key element of the IoT 2.0 solution. Thus, by offloading a part of processing to the edge devices, Edge AI cuts application response time in half, making it possible to collect data in real-time and make decisions based on that. This

reduces the latencies synonymous with cloud solutions where data has to go through large distances before it gets to data centers for analysis. In addition, the availability of edge computing makes it possible to process sensitive information on the device and not transfer raw data over the network. This works to minimize data leaks and is also acceptable to the provisions of the GDPR hence making it suitable to be applied in hospitals, banks, and personal security. In this study, experimental verification has been proven to support the benefits of Edge AI by improving several performance parameters like latency, bandwidth, and model accuracy. The proposed framework is, therefore, superior to traditional cloud-based IoT systems as it gives shorter response time, lower energy consumption and a similar accuracy level to that of cloud-based IoT systems. This shows that edge devices can support any AI-dependent tasks as effectively as possible and in limited-resource systems. Local data processing also makes it possible for real-time decisions and such applications as self-driving cars, smart industries, and healthcare are a good example of mission-critical applications.

There are ample possibilities to optimize and improve the existing concept of Edge AI. Further research is still to be done on TinyML, an initiative to reduce the size of machine learning models that will even allow for their deployment on the simplest of devices that are low power. However, Neuromorphic chips, which are designed to provide an architecture similar to the brain's, will facilitate ultra-low power AI operations, which will further the adoption of Edge AI in power-starved scenarios. Another future research direction is Cross-Silo Federated Learning to make multiple edges or different organizations learn and train the model together without exchanging the data. This approach would make the training of Edge AI more decentralized, hence applying it in areas where the privacy and sharing of data is vital. In general, the developments in Edge AI will persist in propelling the progress of IoT systems in terms of performance, sustainability, and adaptability throughout numerous fields.

References

1. Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. Computer Networks, 54(15), 2787-2805.
2. Djelouat, H., Amira, A., & Bensaali, F. (2018). Compressive sensing-based IoT applications: A review. Journal of Sensor and Actuator Networks, 7(4), 45.
3. Bonomi, F., Milito, R., Natarajan, P., & Zhu, J. (2014). Fog computing: A platform for the Internet of things and analytics. Big data and internet of things: A roadmap for smart environments, 169-186.
4. Yi, S., Li, C., & Li, Q. (2015, June). A survey of fog computing: concepts, applications and issues. In Proceedings of the 2015 workshop on mobile big data (pp. 37-42).
5. Habibi, P., Farhoudi, M., Kazemian, S., Khorsandi, S., & Leon-Garcia, A. (2020). Fog computing: a comprehensive architectural survey. IEEE Access, 8, 69105-69133.
6. Mukherjee, M., Shu, L., & Wang, D. (2018). Survey of fog computing: Fundamental, network applications, and research challenges. IEEE Communications Surveys & Tutorials, 20(3), 1826-1857.
7. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE internet of things journal, 3(5), 637-646.
8. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. IEEE communications surveys & tutorials, 17(4), 2347-2376.
9. Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Federated learning for Internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1622-1658.
10. Sodhro, A. H., Pirbhulal, S., & De Albuquerque, V. H. C. (2019). Artificial intelligence-driven mechanism for edge computing-based industrial applications. IEEE Transactions on Industrial Informatics, 15(7), 4235-4243.
11. Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457-7469.
12. Kyriazis, D., & Varvarigou, T. (2013). Smart, autonomous and reliable Internet of Things. Procedia Computer Science, 21, 442-448.
13. Khayyam, H., Javadi, B., Jalili, M., & Jazar, R. N. (2019). Artificial intelligence and the Internet of Things for autonomous vehicles. In Nonlinear approaches in engineering applications: Automotive applications of engineering problems (pp. 39-68). Cham: Springer International Publishing.
14. Campolo, C., Genovese, G., Iera, A., & Molinaro, A. (2021). Virtualizing AI at the distributed edge towards intelligent IoT applications. Journal of sensor and actuator networks, 10(1), 13.
15. Ammar, M., Russello, G., & Crispo, B. (2018). Internet of Things: A survey on the security of IoT frameworks. Journal of Information Security and Applications, 38, 8-27.
16. Wang, X., Han, Y., Wang, C., Zhao, Q., Chen, X., & Chen, M. (2019). In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning. Ieee Network, 33(5), 156-165.
17. Nguyen, D. C., Ding, M., Pham, Q. V., Pathirana, P. N., Le, L. B., Seneviratne, A., ... & Poor, H. V. (2021). Federated learning meets blockchain in edge computing: Opportunities and challenges. IEEE Internet of Things Journal, 8(16), 12806-12825.
18. El-Sayed, H., Sankar, S., Prasad, M., Puthal, D., Gupta, A., Mohanty, M., & Lin, C. T. (2017). Edge of things: The big picture on integrating edge, IoT and the cloud in a distributed computing environment. I.e. access, 6, 1706-1717.

19. Yu, W., Liang, F., He, X., Hatcher, W. G., Lu, C., Lin, J., & Yang, X. (2017). A survey on the edge computing for the Internet of Things. IEEE Access, 6, 6900-6919.
20. Hamdan, S., Ayyash, M., & Almajali, S. (2020). Edge-computing architectures for Internet of things applications: A survey. Sensors, 20(22), 6441.