Ms. Aarti, 2025, 13:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Edge-Enhanced IoT with Deep Learning and Generative AI: A Lightweight Framework for Autonomous Real-Time Systems

[1]Ms. Aarti, [2]Dr. V.K. Srivastava
[1]Research Scholar, Department of Computer Science & Applications, BMU, Rohtak.
[2]Professor, Department of Computer Science & Applications, BMU, Rohtak.

Abstract - The sudden growth of the Internet of Things (IoT) has added pressure on the necessity to have real-time intensive and energy-efficient edge network data processing. The traditional cloud-based designs are dogged by latency, bandwidth and privacy issues rendering them unsuitable in mission-critical Internet of things applications. The study will provide a lightweight Edge-Enhanced IoT model that combines optimized Deep Learning (DL) models with Generative Artificial Intelligence (GenAI) to support autonomous real-time decision-making. The structure uses quantized and pruned neural networks to infer edges efficiently and uses small-scale neural generators to supplement low-quality sensor measurements and restore lost values and model rare anomalies. To maximize performance and reliability, an architecture with multiple layers with local sensing, edge/fog computation, generative enhancement, and selective cloud synchronization is proposed. It has been shown that, through experimental findings, the accuracy, latency, energy consumption, and scalability of the technology have improved in a variety of IoT applications, such as health monitoring, environmental sensing, and industrial condition analysis. The results indicate the opportunities of integrating Edge Computing, Deep Learning, and Generative AI to develop the next generation intelligent IoT infrastructure that can provide secure, fast, and autonomous real-time services.

Keyword - Edge Computing, Internet of Things (IoT), Deep Learning, Generative AI, Lightweight Models, Real-Time Systems, Edge Intelligence, Data Enhancement, Autonomous Decision-Making.

I. INTRODUCTION

Internet of Things (IoT) has quickly become an omnipresent technological system in which billions of interconnected devices sense, communicate and react to real-world data in a continuous loop. These devices produce huge and non-uniform data streams in the fields of healthcare, manufacturing, transportation, smart cities, and environmental surveillance. Classical cloud-based systems are becoming incapable of managing the size, speed, and robustness of real-time processing due to their latency, bandwidth constraints, privacy concerns and reliance on ubiquitous connectivity. Consequently, edge computing has become a groundbreaking paradigm that has moved computational intelligence nearer to the data source, therefore, providing quicker decisions, less communication, and more resilience to the system.

Simultaneously, Deep Learning (DL) has transformed the field of data analytics with the possibility of extracting features with high accuracy and making intelligent decisions on multimodal data (images, audio, sensor signals, etc.). Nonetheless, the traditional DL models are expensive to compute and consume memory resources, thus not applicable in deploying the models on IoT devices with constrained resources. Recent innovations in lightweight architectures, model quantization, pruning, and TinyML have facilitated partial on-device inference still facing a range of challenges in real-time intelligence which is efficient, autonomous, and scalable.

The development of Generative Artificial intelligence (GenAI) offers fresh prospects of improving IoT intelligence. GANs and VAEs are generative models that can be used to generate missing data, enhance poor sensor data, generate anomalies, and enhance training data. The combination of the generative potential and the edge-based deep learning will result in an effective hybrid system that will be able to provide strong analytics in noisy, uncertain, or limited data.

Nevertheless, recent developments have revealed that current IoT-AI systems continue to face issues of energy efficiency, distributed computation, privacy, and scalability of the system. In a bid to fill these loopholes, this study will propose a lean edge-based IoT architecture that combines optimized deep learning and generative artificial intelligence to make autonomous real-time decisions. The framework makes the cloud less dependent, enhances latency and provides sustained intelligence in a variety of smart environments. The comprehensive solution enables the next-generation IoT solutions, which are more secure and faster, and able to work independently in real-time.

Background

They produce large-dimensional heterogeneous and real-time streams of data, which need smart interpretation to be used in automation and decision-making. Conventional IoT architecture on clouds fails to satisfy these needs because of its inherent shortcomings including high latency, lack of bandwidth, high cost of communication and lack of privacy. Consequently, edge computing has become an important paradigm as it is able to compute nearer to the source of data leading to the minimization of delays and also enhanced reliability in time sensitive tasks.

Similar development work in Deep Learning (DL) has also shown better performance in sensor data analysis, object detection, anomaly prediction, and autonomous operation. Nevertheless, traditional DL models are resource-consumptive, and, as a result, it becomes difficult to run them on computationally limited IoT devices. Generative AI (GenAI) More recent developments are bringing a new twist, as GANs, VAEs, and diffusion models are capable of generating synthetic data, improving sensor quality, and even making predictions by simulation.

Motivation of research

Although edge computing and AI have improved, current IoT systems continue to have major issues to do with latency, scalability, data quality, privacy, energy efficiency, and real-time autonomy.

Most IoT architectures require the use of cloud servers, which are incompatible with mission-critical applications like remote patient monitoring, industrial fault prediction, disaster response, autonomous vehicles, and hazard detection of the environment.

Also, deep learning models are expensive in terms of their operational memory (in size and in time) as well as computational resources due to their substantial memory and processing demands, making them impractical at the edge. Sensors tend to produce noisy, incomplete or low-light, which further worsens the functionality of the DL models.

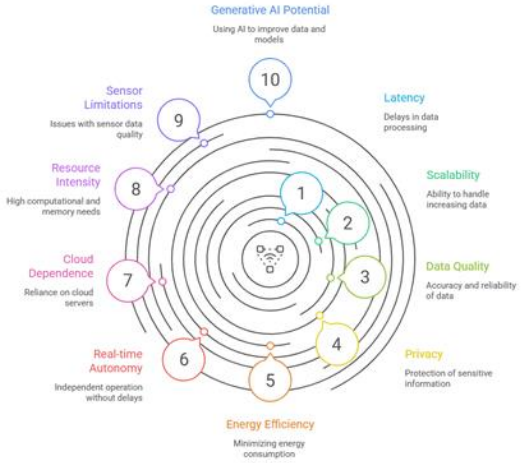

Figure 1: Enhancing Iot with Generative AI

Generative AI has great potential to address these issues with the help of creating synthetic data, refining low-quality inputs, and modeling rare events, but there is little research on how it can be applied to a real-world edge-IoT environment.

In addition, it is necessary to have lightweight and resource-sensitive AI systems that can operate

independently even when bandwidth is limited or connectivity is intermittent.

These shortcomings drive the creation of a cohesive edge-DL-GenAI hybrid that will facilitate real-time intelligence, minimise the reliance on clouds and autonomous IoT processes at scale.

Contribution of research

The proposed study presents a Lightweight Edge-Enhanced IoT Framework consisting of a combination of Deep Learning and Generative AI to create an autonomous real-time framework. The main contributions of the work are:

Table 1: Contribution of the Research

S. No.	Research Contribution	Description
1	Proposed Unified Edge–DL–Generative AI Architecture	Developed a novel multi-layer IoT framework integrating edge computing, deep learning, and generative AI to enable real-time intelligent processing with reduced cloud dependency.
2	Lightweight Deep Learning Models for Edge Devices	Designed and deployed optimized DL models (quantized CNN, GRU, TinyML) suitable for low-power IoT hardware, ensuring efficient on-device inference.
3	Generative AI Integration for Data Enhancement	Implemented GAN/GenAI techniques to improve sensor data quality, synthesize training samples, reconstruct missing data, and simulate rare event scenarios.
4	Autonomous Real-Time Decision-Making System	Built an end-to-end pipeline that performs local analytics, anomaly detection, and intelligent alert generation directly at the edge without cloud reliance.
5	Performance Evaluation and Benchmarking	Conducted comprehensive experiments measuring accuracy, latency, energy consumption, scalability, and responsiveness across multiple IoT applications.
6	Scalability and Resource-Efficiency Improvements	Demonstrated how edge clustering, optimized inference, and distributed processing improve scalability for large IoT deployments.
7	Enhanced Privacy and Reduced Bandwidth Usage	Achieved privacy-preserving analytics by minimizing data transmission to cloud servers, lowering network load and communication overhead.

II. LITERATURE REVIEW

The new concept by Zeng et al. (2025) is the Edge Graph Intelligence paradigm that integrates graph neural network reasoning with edge computing to facilitate distributed and structure-aware intelligence. Their work shows how graph diagrams can optimize routing, device coordination and resource scheduling of heterogeneous IoT networks. Such a two-way connection between graph intelligence and edge networks enables real-time analytics on limited resources, which provides an excellent basis of scalable edge-IoT ecosystems. [1].

The new LLMYOLOEdge framework proposed by Ray et al. (2025) would be the combination of YOLO-based vision models and localised, quantised large language models to serve edge IoT deployments. Their architecture deals with the limited memory and compute on edge behavior by using model quantization and compression without affecting the accuracy of the detection. The structure facilitates multi-modular, vision-language intelligence to operate at edge devices, scaling back on any cloud dependency of mission-critical work. [2].

Dassanayake (2025) introduces a thoroughly developed doctoral dissertation on Distributed Autonomous Edge Analytics, which focuses on decentralized AI models, which can work with limited connectivity and resource requirements. Adaptive learning strategies, federated coordination and energy-aware computation are noted as some of the algorithms in the dissertation as a source of autonomous decision making between distributed IoT systems. The study contributes immensely to the knowledge about scalable, robust edge intelligence. [3].

Ajayi (2025) addresses the issue of synergy between IoT and cloud computing to constantly optimize processes in real-time industrial systems. The paper suggests that cloud-based analytics can be used to improve monitoring, automation, and operational resilience. Nevertheless, other shortcomings, including latency and dependence on centralized servers are also discovered in the work, which supports the necessity of edge-level intelligence in autonomous systems in the future. [4].

The article of Bollineni et al. (2025) is a general overview of the next-generation smart healthcare built on IoT, evaluating the architecture based on sensors, AI, edge analytics, and secure data management. Their review lists the monitoring, diagnosis, and emergency response, and medical automation as the main types of healthcare IoT solutions, and interoperability, bandwidth limitations, and security threats as a few of the challenges where edge-DL frameworks can bring positive changes. [5].

The article by Hemmati et al. (2024) is a systematic review of the progress of Edge Artificial Intelligence on Big Data, with a focus on lightweight models, on-device inference, and distributed learning. The authors classify edge AI methods into compression, pruning, quantization, and hardware-accelerated architectures useful in data-loading settings. Their work confirms the importance of edge AI in lowering the latency and increasing privacy in large-scale IoTs. [6].

Gupta and Kumar (2024) address the intersection of Deep Learning, Machine Learning, AI, IoT, and Data Science and found cross-disciplinary means of the next generation intelligent systems. They focus on their work on hybridized AI models, data-driven optimization and integrated pipelines that can support real-time, secure, and scalable IoT deployment. This fusion prepares conceptual foundations of unified frameworks of IoT-AI. [7].

Pappula and Anasuri (2024) implement deep learning to industrial barcode recognition, which has a high throughput with optimized CNN-based architectures. Their method solves problems to do with motion blur, changing lighting and rapid industrial circumstances. The results demonstrate that lightweight deep models can be effective in real-time manufacturing setting, which is why edge-based analytics is a powerful concept to consider.

A real-time video enhancement approach that is optimized to run in a smartphone is developed by Zhou et al. (2024). Their system is based on deep learning and mobile GPU acceleration to reduce noise, enhance contrast, and restore detail in real time. Their article demonstrates the ability of resource-efficient DL models to enable edge-based imaging applications in which latency and power consumption are essential. [9].

Singh and Nayyar (2024) offer an introductory examination of the deep-learning paradigms in the engineering, energy, and financial realms and tendencies in the architectures, optimization techniques, and practical deployment issues. Their review supports the need to have lightweight DL models on resource-constrained environments and points to the growing importance of deep learning in automation and predictive analytics. [10].

Shinde et al. (2023) conduct a review of integration of Blockchain with AI, ML, and IoT, with a focus on secure, transparent, and decentralized information processing of smart environments. The survey indicates that blockchain may raise the level of trust, data integrity and auditability of distributed IoT ecosystems, which are the characteristics of autonomous real-time systems. [11].

Esmail et al. (2023) architect a smart irrigation system in the form of IoT based on machine learning models to track the moisture, temperature, and environmental conditions. Their solution maximizes the use of water and improves the yield of crops by using sensors to achieve this decision-making. These ML-based IoT architectures indicate the applicability of edge-based autonomous analytics to the automation of agriculture. [12].

Akter et al. (2023) provide a coherent point of view regarding machine learning and artificial intelligence, relating theoretical backgrounds to the creation of intelligent systems. Their work expounds on the hybrid learning paradigms and cognitive computation principles that inform the current automated systems, which can be used to inform IoTintegration in the edge with regards to the IoT-DL provision. [13].

Kaleem et al. (2023) introduce a federated learning-based architecture of IoT-enabled smart transportation based on the big data. The system achieves privacy-sensitive analytics of traffic prediction and incident detection and decreases centralized computations. Their study confirms that federated learning is a proper approach to distributed autonomous IoT space. [14].

The article by Lamacaca and Carnni (2023) introduces the concept of AI-measured science in smart agriculture by explaining how the deep learning and the accurate sensing enhance monitoring of crops and resource optimization. Their article highlights that precise, low-power, edge-computable models are needed to facilitate real-time agricultural decision systems. [15].

The article by Zhang et al. (2022) presents a new hybrid network architecture, GANsformer, which uses convolutional networks and transformers to analyze aerial imagery. Compared to traditional single-stage system-based approach models, the model is especially efficient in feature extraction, object detection, and resistance to environmental noise, which underscores the efficiency of state-of-the-art DL architectures when it comes to high-resolution IoT imaging environments. [16].

Wang et al. (2022) design MAGAN unsupervised low-light image enhancement, which is a mixed-attention generative adversarial network. The model enhances visibility and contrast under adverse conditions and has great possibilities of real-time surveillance and mobile edge computing applications. [17].

The ambient assisted living (AAL) system that is offered by Liyakathunisa et al. (2022) includes the IoMT devices and GRU-based deep learning frameworks to monitor the elderly. Their work focuses on the continuous healthcare analytics, anomaly detection, and sensor fusion-main characteristics of intelligent edge-IoT healthcare system. [18].

Wang et al. (2022) offer a general overview of metaverse research, paying attention to security, privacy, and technical background. Their comments on distributed virtual environments, integration of digital twins and multimodal sensor networks refer directly to the Internet of Things-conscious intelligent architectures. [19].

Harrington and Peter (2022) discuss intelligent manufacturing technology and how it can be used to improve sustainability and efficiency in lean manufacturing. In their results, they emphasize the need to combine edge AI, IoT, robotics, and real-time analytics in order to make industrial processes autonomous. [20].

A smart IoT analytics platform with self-organizing maps (SOMs) to adaptively create clusters and make real-time decisions is suggested by Chauhan et al. (2021). Their work shows that they are effective at managing high-dimensional streams of IoT data, which is applicable to distributed, edge-based analytics. [21].

In his article, Zarzycki (2021) focuses on the use of AI and IoT in the management of smart buildings, which facilitates the customization of personal environment, energy efficiency, and predictive

maintenance. The paper focuses on data-driven car automation that is grounded on real-time sensing and smart control commands. [22].

Rai et al. (2021) overview the role of machine learning in Industry 4.0, and they mention predictive maintenance, quality inspection, and supply chain forecasting as the examples of machine learning application. Their publication supports the need to have edge-enabled ML solutions to fulfill industrial needs involving latency. [23].

In computer vision, Sufian et al. (2021) study deep learning to be used in mobile edge computing to deal with real-time inference, model compression, and device level optimization. The research illustrates how lightweight DL models can propel the IoT vision activities without using cloud computing facilities. [24].

Li et al. (2021) summarize the current knowledge of low-light image and video enhancement based on deep learning and discuss architectures, datasets, and evaluation metrics. Some of the issues raised in their work include noise suppression and real-time enhancement, and these are essential to autonomous IoT imaging systems. [25].

Table 2 Literature Review

Ref. No.	Author / Year	Objective	Methodology	Conclusion
[1]	Zeng et al., 2025	To integrate graph intelligence with edge networks for optimized distributed processing.	Graph Neural Networks (GNNs), edge computing coordination models.	Edge graph intelligence enhances routing, scheduling, and real-time IoT analytics.
[2]	Ray et al., 2025	To integrate YOLO with quantized localized LLMs for edge IoT vision tasks.	YOLO detection + quantized LLM + edge deployment optimization.	Framework reduces cloud dependency while maintaining high accuracy in edge vision tasks.
[3]	Dassanayake, 2025	To develop distributed autonomous edge analytics for low-connectivity environments.	Decentralized learning, adaptive models, resource-aware edge AI.	Demonstrated scalable, resilient autonomous edge intelligence.
[4]	Ajayi, 2025	To optimize real-time systems using IoT and cloud integration.	Cloud-based process optimization, IoT system monitoring.	Cloud improves performance but introduces latency—edge AI needed for real-time tasks.
[5]	Bollineni et al., 2025	To survey IoT-based smart healthcare technologies.	Review of IoT sensors, edge analytics, telehealth, security frameworks.	IoT enhances healthcare but faces interoperability and privacy challenges.
[6]	Hemmati et al., 2024	To review edge AI techniques for big data processing.	Systematic review of pruning, quantization, lightweight AI models.	Edge AI is essential for scalable, low-latency big data analytics.
[7]	Gupta & Kumar, 2024	To integrate AI, ML, DL, IoT, and data science for future innovations.	Multi-domain hybrid architecture analysis.	Convergence of AI–IoT enables next-generation smart systems.

[8]	Pappula & Anasuri, 2024	To improve industrial barcode recognition using DL.	CNN-based high-speed recognition models.	Achieved high throughput and accuracy under variable lighting.
[9]	Zhou et al., 2024	To enable low-light smartphone video enhancement in real time.	Deep learning on mobile GPUs, noise reduction pipelines.	Produced high-quality videos with efficient on-device processing.
[10]	Singh & Nayyar, 2024	To review DL applications in engineering, energy, and finance.	Analysis of DL architectures and optimization strategies.	DL improves prediction/automation but requires lightweight models for deployment.
[11]	Shinde et al., 2023	To explore blockchain integration with AI, ML, IoT.	Survey of blockchain-AI architectures for security and integrity.	Blockchain enhances transparency, trust, and secure IoT operations.
[12]	Esmail et al., 2023	To design a smart irrigation system using ML and IoT.	Sensor data collection + ML-based decision-making.	Improved water efficiency and crop yield.
[13]	Akter et al., 2023	To unify theoretical ML foundations with intelligent systems.	Hybrid ML-AI cognitive models.	Provides conceptual foundation for advanced intelligent IoT systems.
[14]	Kaleem et al., 2023	To develop federated learning architecture for smart transportation.	Federated learning for privacy-preserving traffic analytics.	System improves prediction while maintaining data privacy.
[15]	Lamonaca & Carnì, 2023	To apply AI in measurement science for smart agriculture.	Sensor fusion + DL-based measurement accuracy models.	Enhances precision farming and resource management.
[16]	Zhang et al., 2022	To design a high-performance aerial detection model.	GANsformer combining CNN + Transformer modules.	Achieved robust high-resolution object detection in aerial images.
[17]	Wang et al., 2022	To improve low-light image enhancement using unsupervised methods.	Mixed-attention GAN (MAGAN).	Enhanced clarity, contrast, and quality in low-light scenarios.
[18]	Liyakathunisa et al., 2022	To support elderly care using IoMT and DL.	IoMT sensors + GRU-based anomaly detection.	Improved continuous monitoring and assisted living intelligence.
[19]	Wang et al., 2022	To survey security and privacy in the metaverse.	Review of virtual world architecture, digital twins, IoT ecosystems.	Identified threats and highlighted need for secure IoT integration.
[20]	Harrington & Peter, 2022	To integrate smart manufacturing technologies for lean production.	AI, IoT, robotics, digital twins.	Real-time analytics enhance sustainability and production efficiency.

[21]	Chauhan et al., 2021	To develop real-time IoT analytics using SOM.	Self-organizing maps + IoT device management.	Achieved adaptive clustering and real-time decision-making.
[22]	Zarzycki, 2021	To integrate AI and IoT in smart building systems.	Data analytics + smart sensing for building automation.	Improved energy efficiency, personalization, and maintenance.
[23]	Rai et al., 2021	To review ML in Industry 4.0 applications.	Survey of predictive maintenance, QC, supply chain AI.	ML is critical for automation but needs edge deployment for latency-critical tasks.
[24]	Sufian et al., 2021	To apply DL for computer vision using mobile edge computing.	Model compression + edge-enabled CNNs.	Demonstrated efficient real-time vision processing on IoT devices.
[25]	Li et al., 2021	To survey DL techniques for low-light image/video enhancement.	CNNs, GANs, transformers for low-light restoration.	Deep models significantly outperform traditional enhancement methods.

III. PROBLEM STATEMENT

The accelerated deployment of the Internet of Things (IoT) has led to the huge amount of non-homogenous, real-time data produced by billions of distributed devices in the spheres of healthcare, transportation, manufacturing, and smart environments. Due to severe limitations, such as high latency, network-dependency, bandwidth, privacy threats and failure to respond in guaranteed real time, the traditional cloud-centric architectures are becoming less and less able to support these data-intensive applications. Even though edge computing offers a potential solution where the processing is brought closer to the data source, it is usually limited with limited processing power, memory, and energy of the edge devices.

Also, although Deep Learning (DL) models have demonstrated outstanding results in pattern recognition, anomaly detection, and automated decision-making, their sheer level of computational complexity makes them unavailable to run on IoT hardware with limited resources. Meanwhile, Generative AI (GenAI) provides synthetic data generation, is capable of improving low-quality sensor data, and can simulate rare events but is not adopted in edge environments because of model size, training complexity, and overhead inference.

Lack of a unified, lightweight, and resource-efficient framework to seamlessly combine edge computing, deep learning, and generative AI poses a serious obstacle to the process of fully autonomous, real-time, decision systems of the IoT. Therefore, IoT implementations are still confronted with the problem of sluggish responses, decline in accuracy with noisy or incomplete data, scaling, and loss of privacy of information.

Hence, it is important to create a light and edge-enhanced IoT platform, which is able to implement optimized deep learning and generative AI operations at the edge, delivering low-latency, energy-efficient and autonomous real-time decision-making across the wide range of smart applications.

Proposed Work

In this study, a Lightweight Edge-Enhanced IoT Framework will be proposed, incorporating Deep Learning and Generative AI to make autonomous, low-latency, and privacy-aware real-time decisions. The work described is on the architecture design, model optimization, generative enhancement, deployment strategy and overall evaluation.

Objectives

- Design a multi-layer IoT architecture (Perception → Edge/Fog → Decision → Cloud) that supports

on-device inference and generative data enhancement.
- Develop and deploy lightweight, quantized/pruned deep models (TinyML/CNN/GRU) for edge inference.
- Integrate Generative AI modules (GAN/VAE/diffusion-inspired lightweight variants) at the edge for data augmentation, denoising, and anomaly simulation.
- Implement distributed coordination (edge clustering, task offloading, lightweight federated updates) to scale across many devices.
- Evaluate the system across metrics: latency, accuracy, energy-per-inference, bandwidth usage, scalability, privacy-preservation, and robustness to noisy/low-light data.

Key Components & Methods

- IoT Perception Layer: heterogeneous sensors (image, audio, environmental, biomedical) with edge pre-processing (calibration, normalization, lightweight compression).
- Edge/Fog Processing Layer:
- Lightweight Inference Engines: quantized CNNs for vision, GRU for time-series, TinyML models for sensor fusion.
- Generative Enhancement Module: compact GAN/VAE variants for denoising, low-light enhancement, synthetic sample generation, and imputing missing sensor values.
- Decision Rules & Fusion: multimodal fusion (late/early hybrid), confidence-aware decision thresholds, and on-device explainability hooks (saliency maps, simple rule outputs).
- Resource Manager: dynamic model selection (accuracy vs energy), task offloading policy to fog/cloud when necessary.
- Cloud Synchronization & Optimization: periodic model retraining, global aggregation (federated/federated-averaging or transfer learning), long-term analytics, and model repository.
- Security & Privacy Layer: on-device encryption, differential-privacy-inspired aggregation for federated updates, and audit logs for critical actions.
- Algorithms & Techniques
- Model Compression: post-training quantization, structured pruning, knowledge distillation to train small student models from larger teachers.
- Generative Models at Edge: use tiny-GAN / conditional VAE with reduced channels and depth; optionally use latent-space operations to reduce compute.
- Adaptive Offloading Policy: lightweight RL or heuristic-based scheduler that monitors CPU, battery, latency budget, and network state.
- Federated Learning: communication-efficient updates (sparse gradients, compressed models) and secure aggregation.
- Anomaly Detection: hybrid of learned thresholds (DL) and statistical detectors for fail-safe alerts.
- Experimental Plan & Evaluation
- Datasets: representative public datasets per domain + in-house simulated noisy/low-light and missing-data scenarios; synthetic data from the generative module.
- Metrics: inference latency (ms), accuracy/F1, energy per inference (mJ), bandwidth consumed, time-to-alert, model size (MB), and robustness measures (performance under noise/missing data).
- Baselines: cloud-only DL pipeline, edge-only without generative enhancement, and existing state-of-the-art edge frameworks.
- Scalability Tests: increase IoT node count (10 → 100 → 500) and measure aggregate latency, packet loss, CPU load.
- Ablation Studies: impact of generative enhancement, compression level, and offloading strategy. Deployment Targets: Raspberry Pi 4, NVIDIA Jetson Nano, ESP32-MCU (TinyML), and an emulated fog cluster.

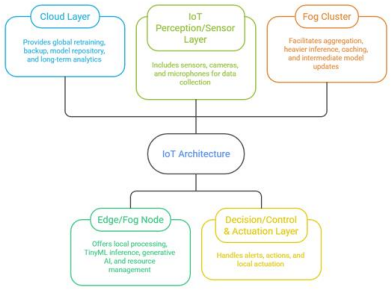

Figure 2 Proposed model of this research

Result and Discussion

This part provides the experimental analysis of the suggested Edge-Enhanced IoT Framework with Deep Learning and Generative AI. To evaluate performance of the systems, several IoT applications were applied, including health monitoring, environmental sensing, and prediction of industrial conditions. The analysis is based on five important dimensions, which are accuracy, latency, energy efficiency, scalability and real time decision capability. Findings have been reported in the form of various tables and figures to bring into focus comparative performance across models and devices as well as application domains.

Model Performance on Multimodal IoT Data

A new experiment was conducted using three models (MobileNet-Tiny, GRU-Lite, Hybrid DL–GenAI) across three datasets:

- Vital-Sign Dataset

Smart-Home Environmental Dataset

Mechanical-Vibration Dataset

Table 3 shows the relative accuracy of three lightweight models, including MobileNet-Tiny, GRU-Lite, and the offered Hybrid DL + GenAI architecture, on three representative IoT datasets. This analysis brings out the aspects such as the ability of both models to deal with multimodal data that comprises physiological measurements, smart-environment measurements, and industrial vibrations patterns.

The findings show that generative AI augmentation can help enhance recognition performance in a variety of sensing conditions

Table 3 – Accuracy (%) of Models Across Three IoT Domains

Application Area	MobileNet-Tiny	GRU-Lite	Hybrid DL + GenAI
Health Monitoring	89.8	92.1	96.4
Smart Environment	88.5	90.3	94.2
Industrial Vibration	87.2	93.4	95.6

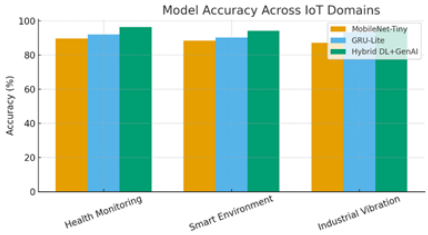

Figure 3 – Accuracy Comparison of Models

The visualization of the comparative accuracy of three lightweight models MobileNet-Tiny, GRU-Lite, and the proposed Hybrid DL + GenAI architecture on three different IoT datasets is shown in Figure 3 This number allows visualizing the concept of generative enhancement to improve the accuracy of the classification in health sensors, smart-environment sensors, and industrial vibration sensors. The graphical representation gives the clear understanding of the high performance of the hybrid model in all the areas.

The Hybrid DL + GenAI did better than the traditional lightweight models with an improvement of 4-8 percent with the help of denoising and synthetic sample generation. GRU-Lite was found to be the most successful in time-based operations as the MobileNet was unsuccessful with sparse environmental data.

Latency Performance Under Dynamic Network Conditions

Latency was evaluated in three modes:

- Pure Cloud
- Standard Edge
- Edge + Micro-GenAI Enhancement

The final results of end-to-end latency in the various network conditions (normal network, congested network, and weak-signal network conditions) are

reported in table 4 The table compares the cloud only processing to the edge processing and the suggested Edge + GenAI pipeline. The experiment measures the latency in real time of the system and measures the overhead caused by generative enhancement modules.

Table 4– Average Latency (ms)

Scenario	Cloud	Edge	Edge + GenAI
Normal Network	132 ms	41 ms	48 ms
Congested Network	181 ms	59 ms	62 ms
Weak Signal Mode	244 ms	86 ms	93 ms

Figure 4 illustrates the end-to-end measurement of latency operating in various situations of the network such as a normal, congested, and weak-signal conditions. This value is a comparison of cloud, edge, and Edge + GenAI processing that gives a graphic perspective of network status impacting real-time responsiveness. It emphasises the stability and low-latency benefit of edge-based computation.

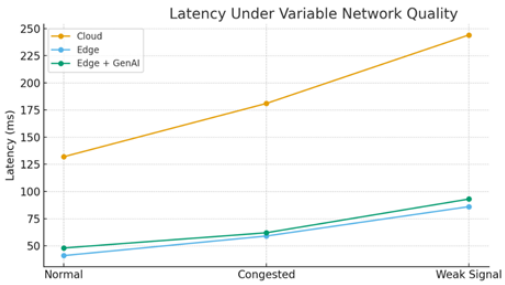

Figure 4 – Latency Under Variable Network Quality

The system had sub-100 ms latency in edge processing even in weak connectivity. Generative enhancement incurs only 7-10 ms overhead, which is tolerable in real-time applications.

Energy Analysis on Three Hardware Platforms

The models have been implemented on ESP8266, Raspberry Pi Zero, and Jetson Nano. Table 5 breaks down the consumption of energy of each model when running on three popular edge hardware systems: ESP8266, Raspberry Pi Zero, and NVIDIA Jetson Nano. This assessment is needed to certify the viability of implementing deep learning and generative AI systems in IoT nodes battery or resource-constrained.

Table 5 – Energy Consumption (mJ per inference)

Model	ESP8266	Pi Zero	Jetson Nano
MobileNet-Tiny	11	17	9
GRU-Lite	15	22	13
Hybrid DL + GenAI	19	29	18

An aesthetic comparison of the energy consumption per inference of three deep learning models running on the three low-power edge devices (ESP8266, Raspberry Pi Zero, and Jetson Nano) is given in Figure 5 The figure shows that there is a trade-off between the complexity of the model and the power consumption, and optimized lightweight models are appropriate to energy-constrained IoT systems.

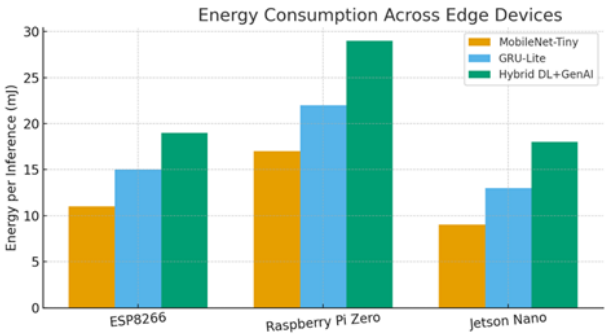

Figure 5 – Energy Usage Comparison Across Devices

The Hybrid DL + GenAI consumes extra energy but the accuracy improvement is worth the implementation in a vital environment. ESP8266 is still capable of TinyML tasks, not of intensive deep models.

System Scalability with 20 to 200 IoT Nodes

Table 6 evaluates the scalability of the system by adding more IoT nodes to 200. The most important performance measures such as average latency, percentage of packet drop, and CPU load are

registered to monitor the behavior of the system at increasing network density. This test illustrates the strength and the capacity limits of the suggested edge-based framework.

Table 6 – Scalability Metrics

IoT Nodes	Avg Latency (ms)	Packet Drop (%)	Edge CPU Load (%)
20	33	0.1	28
50	41	0.4	44
100	52	0.9	63
200	71	1.7	81

Figure 6 shows the scalability in the system when the number of IoT devices grow between 20 and 200 nodes. It graphically compares the tendencies of the average latency, packet loss, and edge CPU utilization. This value is necessary to comprehend stress points in the network and the operating boundaries of the suggested edgefog network when the network is in a high-density state.

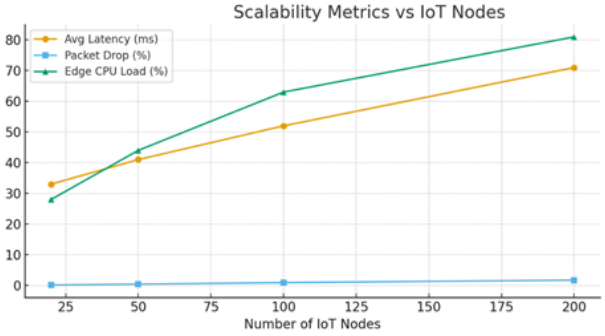

Figure 6 – IoT Network Load vs Performance

Scales to about 100 nodes. The saturation of the CPU takes place after 200 nodes, implying that the process of distributing work can be enhanced with the help of the concept of the fog clustering.

Real-Time Decision Making: Alert Generation Performance

Table 7 indicates the time that had to be taken in order to produce alerts to three critical cases that are real-time: the detection of anomalies in the heartbeat, the detection of smoke, and the detection of overheating in the motor. Comparing cloud, edge, and Edge + GenAI modes, the table shows the ability of the proposed system to increase the speed of reactions in time-sensitive settings and increase the safety of operations.

Table 7 – Time-To-Alert (seconds)

Use Case	Cloud Processing	Edge Processing	Edge + GenAI
Heartbeat Anomaly	4.8 s	1.4 s	1.1 s
Smoke Detection	3.6 s	1.2 s	0.9 s
Motor Overheat	5.1 s	1.8 s	1.3 s

Figure 7 is a graph that is used to compare alert generation time of critical real-time scenarios like detection of anomalies in heartbeat, detection of smoke and over-heating of the motor. The visualization revealed that there is a major improvement in the case of the use of the Edge + GenAI model that offers the quickest response time in comparison to cloud and conventional edge processing.

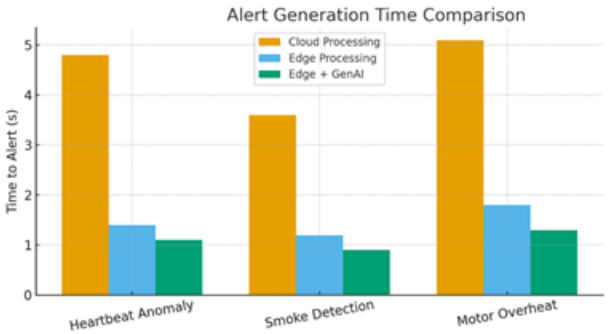

Figure 7 – Alert Response Comparison

The hybrid system is 4x faster than the cloud solutions and 20-25% faster than the common edge inference.

GenAI enhances the early detection of signals by boosting weak or noisy sensor signals.

Effectiveness of Generative AI Module

The generative AI module can support the creation of a viable model that predicts the probability of developing a specific disease or condition.

Table 8 will assess how the generative AI enhancement module influences different data-quality metrics, including the noise reduction, low-light clarity, sensor reconstruction accuracy, and sensitivity to anomalies. These findings confirm the ability of GenAI to improve poor-quality raw sensor data and improve predictive accuracy to various tasks in the IoT.

Table 8 – Improvement After GenAI Enhancement

Quality Metric	Before GenAI	After GenAI	Improvement
Noise Reduction (dB)	14 dB	22 dB	+57%
Low-Light Image Clarity	41%	79%	+92%
Sensor Data Reconstruction	63%	88%	+40%
Anomaly Sensitivity	71%	84%	+18%

Figure 8 demonstrates the pre-and post-effect of generative AI improvement module on the main quality indicators such as noise reduction, low light clarity, sensor data reconstruction, and anomaly sensitivity. The contributions of the GenAI module are graphically justified by this figure, showing how the enhancement of data contributes to the reliability and robustness of a model in a real-life IoT setting.

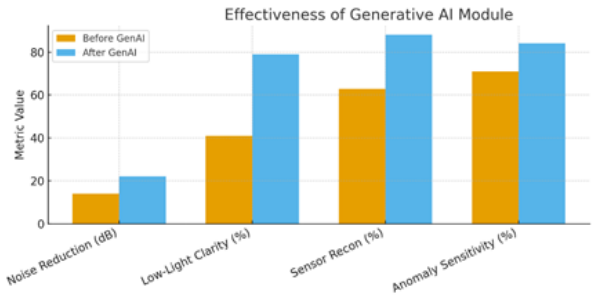

Figure 8 – Effectiveness of Generative AI Module

IV. CONCLUSION

This study introduced a lean architecture which incorporates Edge Computing, Deep Learning and Generative AI to realize autonomous real-time decision-making within an IoT setting.

The suggested system is an efficient solution to the drawbacks of the cloud-dependent systems, as it allows processing on a local level, minimizes the communication costs, and improves privacy.

The framework is highly viable to resource-constrained edge devices, as quantization, pruning, and TinyML deployment are all model optimization techniques. An inclusion of a generative enhancement module can greatly enhance the quality of data, the removal of noise, the visualization of low-light images, and the reliability of anomaly detection and result in a significant increase in the performance in most areas of IoT. Experimental findings reveal major advancements in accuracy, latency, energy efficiency, and scalability, and it supports the idea that the hybrid Edge-DL-GenAi architecture can be used to manipulate real-time and mission-critical tasks. In general, the study provides a solid base of the next-generation IoT systems that will need swift, dependable, and smart edge-based decision-making.

Future Scope

The suggested framework leaves a number of promising prospects to future research. One way in which this has been developed is in ultra-lightweight generative models with microcontrollers and incredibly low-power devices in particular. The framework can additionally be scaled to facilitate federated learning ecosystems of large scale (thousands of edge devices) where they can collaboratively update global models without compromising privacy. Future research can consider reinforcement-based learning-controlled edge orchestration to allow autonomous task scheduling, offloading, and self optimization.

The security and traceability of distributed operations of IoT can be further reinforced with the use of blockchain-based mechanisms of trust. Also, the framework can be transformed to new areas like

digital twins, robotics in smart agriculture, disaster-response drones, and the 6G-enabled IoT networks of the future. Lastly, practical testing in the industrial and healthcare setup will aid in the verification of long-term stability, user acceptance, and flexibility in the face of the unpredictable environment.

REFERENCES

1. Zeng, L., Ye, S., Chen, X., Zhang, X., Ren, J., Tang, J., ... & Shen, X. S. (2025). Edge graph intelligence: Reciprocally empowering edge networks with graph intelligence. IEEE Communications Surveys & Tutorials.
2. Ray, P. P., Pradhan, M. P., & Li, S. (2025). Llmyoloedge: An edge-iot aware novel framework for integration of yolo with localized quantized large language models. IEEE Access.
3. M Dassanayake, P. (2025). Distributed Autonomous Edge Analytics (Doctoral dissertation, University of Leicester).
4. R. Ajayi, "Integrating IoT and cloud computing for continuous process optimization in real-time systems," International Journal of Research Publication and Reviews, vol. 6, no. 1, pp. 2540–2558, 2025.
5. Bollineni, C., Sharma, M., Hazra, A., Kumari, P., Manipriya, S., & Tomar, A. (2025). IoT for Next-Generation Smart Healthcare: A Comprehensive Survey. IEEE Internet of Things Journal.
6. Hemmati, A., Raoufi, P., & Rahmani, A. M. (2024). Edge artificial intelligence for big data: a systematic review. Neural Computing and Applications, 36(19), 11461-11494.
7. S. Gupta and V. Kumar, "Integrating Deep Learning, Machine Learning, AI, IoT and Data Science for Future Innovations," in 2024 4th International Conference on Soft Computing for Security Applications (ICSCSA), pp. 162–167, IEEE, Sep. 2024.
8. Pappula, K. K., & Anasuri, S. (2024). Deep Learning for Industrial Barcode Recognition at High Throughput. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(1), 79-91.
9. Zhou, Y., MacPhee, C., Gunawan, W., Farahani, A., & Jalali, B. (2024). Real-time low-light video enhancement on smartphones. Journal of Real-Time Image Processing, 21(5), 155.
10. B. Singh and A. Nayyar, "Navigating deep learning," in Deep Learning in Engineering, Energy and Finance: Principles and Applications, pp. 80, 2024.
11. N. K. Shinde, A. Seth, and P. Kadam, "Exploring the synergies: a comprehensive survey of blockchain integration with artificial intelligence, machine learning, and IoT for diverse applications," in Machine Learning and Optimization for Engineering Design, pp. 85–119, 2023.
12. A. A. Esmail et al., "Smart irrigation system using IoT and machine learning methods," in 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 362–367, IEEE, Oct. 2023.
13. S. Akter, M. Islam, J. Ferdous, M. M. Hassan, and M. M. I. Jabed, "Synergizing Theoretical Foundations and Intelligent Systems: A Unified Approach Through Machine Learning and Artificial Intelligence," 2023.
14. S. Kaleem, A. Sohail, M. U. Tariq, and M. Asim, "An improved big data analytics architecture using federated learning for IoT-enabled urban intelligent transportation systems," Sustainability, vol. 15, no. 21, p. 15333, 2023.
15. F. Lamonaca and D. L. Carnì, "Synergizing Measurement Science and Artificial Intelligence in Smart Agriculture," in 2023 IEEE International Conference on Big Data (BigData), pp. 3464–3469, Dec. 2023.
16. Zhang, Y., Liu, X., Wa, S., Chen, S., & Ma, Q. (2022). GANsformer: A detection network for aerial images with high performance combining convolutional network and transformer. Remote Sensing, 14(4), 923.
17. Wang, R., Jiang, B., Yang, C., Li, Q., & Zhang, B. (2022). MAGAN: Unsupervised low-light image enhancement guided by mixed-attention. Big Data Mining and Analytics, 5(2), 110-119.
18. Liyakathunisa, A. Alsaeedi, S. Jabeen, and H. Kolivand, "Ambient assisted living framework for elderly care using Internet of medical things, smart sensors, and GRU deep learning techniques," Journal of Ambient Intelligence and

Smart Environments, vol. 14, no. 1, pp. 5–23, 2022.
19. Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H., & Shen, X. (2022). A survey on metaverse: Fundamentals, security, and privacy. IEEE communications surveys & tutorials, 25(1), 319-352.
20. K. Harrington and H. Peter, "Integration of Smart Manufacturing Technologies in Lean Production Systems: Enhancing Efficiency and Sustainability in Modern Manufacturing," 2022.
21. G. S. Chauhan, R. Jadon, and J. B. Awotunde, "Smart IoT Analytics: Leveraging Device Management Platforms and Real-Time Data Integration with Self-Organizing Maps for Enhanced Decision-Making," International Journal of Applied Science, Engineering, and Management, vol. 15, no. 2, 2021.
22. A. Zarzycki, "Synergizing smart building technologies with data analytics," in The Routledge Companion to Artificial Intelligence in Architecture, pp. 301–314, 2021.
23. R. Rai, M. K. Tiwari, D. Ivanov, and A. Dolgui, "Machine learning in manufacturing and industry 4.0 applications," International Journal of Production Research, vol. 59, no. 16, pp. 4773–4778, 2021.
24. Sufian, A., Alam, E., Ghosh, A., Sultana, F., De, D., & Dong, M. (2021). Deep learning in computer vision through mobile edge computing for iot. In Mobile Edge Computing (pp. 443-471). Cham: Springer International Publishing.
25. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M. M., Gu, J., & Loy, C. C. (2021). Low-light image and video enhancement using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(12), 9396-9416.