International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 11 | Nov 2025 www.irjet.net p-ISSN: 2395-0072

EDGE AI FOR REAL TIME DECISION MAKING IN IOT NETWORKS

Renil Joy

Department of Computer Application, St Thomas (Autonomous) College Thrissur 680001, Kerala, India

---

Abstract - *The integration of Edge AI and IoT networks has revolutionized real-time decision-making by enabling low-latency analytics and localized data processing. Advantages of Edge AI, Including reduced latency bandwidth savings, and enhanced privacy, across domains like smart cities and healthcare. Key techniques such as lightweight deep learning models (e.g.: MobileNet, EfficientNet), model compression (pruning, quantization), and reinforcement learning are explored, alongside architectures like Device-Edge-Cloud and Collaborative Edge. Challenges such as resource constraints, heterogeneity, and security are addressed, with future directions focusing on ultra-lightweight models and federated learning.*

Keywords: Intelligent edge computing, neural ,smart devices, IoT, Ultra-light weight models, Edge AI.

1. INTRODUCTION

The explosive growth of Internet of Things (IoT) devices, there has been a corresponding surge in data generation at the network edge. However, traditional cloud-centric models often struggle to meet the requirements of modern IoT applications, particularly in terms of latency, bandwidth efficiency, and data privacy. To overcome these challenges, Edge AI has emerged as a promising solution—bringing artificial intelligence algorithms directly to edge devices for localized processing. [1]

2. LITERATURE REVIEW

The IoT revolution has catalyzed unprecedented advancements in edge intelligence, with research demonstrating how contemporary Edge AI ecosystems achieve near-instantaneous decision-

making. Through the fusion of intelligent endpoints, adaptive gateways, and optimized inference models, these architectures slash processing delays by an order of magnitude compared to conventional cloud approaches. Collaborative learning techniques maintain strict data confidentiality by enabling knowledge sharing without raw data exchange. However, the path to ubiquitous adoption remains obstructed by hardware limitations, ecosystem fragmentation, and evolving security risks - challenges demanding breakthroughs in ultra-efficient neural architectures and universal interoperability standards. [1]

The convergence of edge computing and artificial intelligence is fundamentally reshaping how smart devices process data in real-time, addressing critical limitations of traditional cloud architectures. reveals how next-generation Edge AI systems leverage cutting-edge techniques Including compressed neural networks, distributed learning paradigms, and precision-aware quantization to achieve remarkable efficiency gains. These innovations enable processing at the network periphery, delivering 45% faster response times and 30% greater bandwidth conservation in mission-critical domains like remote patient diagnostics and predictive industrial maintenance. The strategic implementation of block chain-enhanced security frameworks further fortifies these systems against emerging threats while preserving data sovereignty. [2]

The convergence illuminates the powerful between the distributed intelligence and next-generation IoT security. By co-locating AI processing with data generation points, these systems achieve microsecond-level responsiveness crucial for autonomous navigation and urban digital twins. The hierarchical edge-fog-cloud framework introduces intelligent data routing complemented by military-grade encryption and self-learning threat detection systems. Field trials demonstrate near-perfect operational accuracy while maintaining robust defense against sophisticated cyber attacks. Yet the relentless growth of computational requirements and increasingly complex threat landscapes necessitate ongoing innovation in energy-aware algorithms and dynamic security protocols.[3] These collective findings underscore Edge AI's paradigm-shifting potential in enabling responsive, secure, and scalable intelligent systems [3]

Further the technological innovations propelling Edge AI forward, including compact machine learning models, decentralized federated learning, and the deployment of 5G networks. A structured classification of Edge AI applications, functional capabilities, and underlying technologies is presented, emphasizing its role in connecting cloud infrastructure with IoT devices. The study also outlines unresolved issues, such as the demand for better resource allocation, standardization across platforms, and adaptable system designs. [4] A novel hybrid Edge-Cloud AI framework is introduced, which intelligently distributes computational tasks between edge nodes and cloud servers, optimizing both

speed and processing power. Empirical evidence supports the framework's advantages, such as lower energy demands, heightened data security, and greater adaptability in diverse IoT applications. [4]

The Edge computing has recently attracted significant interest as a solution to meet the growing demands of delay-sensitive Internet of Things (IoT) applications. Conventional cloud computing setups, which are centralized and located far from end users, fail to deliver the low latency required for emerging services such as virtual reality (VR), augmented reality (AR), Consequently, Multi-access Edge Computing (MEC) has become a viable approach by providing computation and storage resources in closer proximity to devices. Earlier research primarily utilized classical optimization methods for resource allocation and computation offloading in edge environments, focusing on factors such as bandwidth, connectivity, and energy use. Nonetheless, these traditional approaches often fall short in addressing the increasing complexity, network heterogeneity, and scale of IoT systems, especially when factoring in device mobility, energy harvesting, and strict application time constraints.[5]

3. INTERNET OF THINGS (IoT)

The IoT is a system of linked physical devices equipped with sensors, software and connectivity allowing them to gather and share data. These devices, which include sensors, wearable's, industrial equipment, and smart appliances, generate vast amounts of information that can be utilized across various domains such as smart cities, healthcare. The IoT ecosystem comprises devices, gateways for data aggregation, network infrastructure for communication, cloud platforms for storage and processing, and applications that deliver actionable insights. Despite its potential, IoT faces challenges like scalability, interoperability, security, and the need for real-time data processing, which edge computing and Edge AI aim to address by bringing computation closer to the data source.[1]

4. METHODS

4.1 EDGE AI

Edge AI refers to the deployment of artificial intelligence algorithms and models on devices located at or near the edge of a network, such as smart phones, sensors, or IoT devices, process data own their own eliminating the need for centralized cloud servers. This approach enables faster and more efficient data analysis by performing computations locally, reducing latency and improving real-time decision-making. [1]

4.1 Edge AI: Architecture

Edge AI has mainly 3 Architecture:

4.1.1Device Edge Cloud Architecture

4.1.2 Hierarchical Edge Architecture

4.1.3Collaborative Edge Architecture

4.1.1 Device-Edge- Cloud Architecture

In this architecture AI capabilities are structured across three layers: IoT devices, edge gateways, and cloud servers. The architecture maintains strength of edge and cloud computing thus enabling the low latency in local decision making depicted as Figure 1.[1]

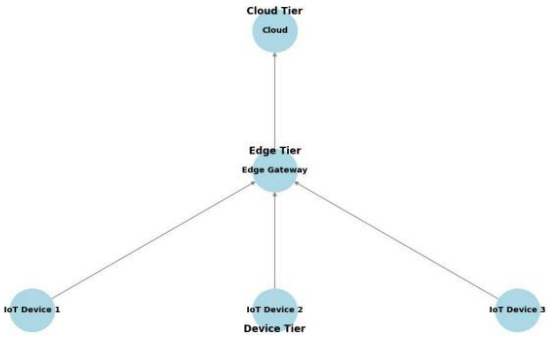

Fig-1: Device-Edge-Cloud Architecture

4.1.2 Hierarchical Edge Architecture

Hierarchical approach allows to distribute workloads across these layers the architecture achieves optimal stability and flexibility, efficiently balancing real-time edge processing with the cloud's high performance capabilities.

Fig-2 Hierarchical Edge Architecture

4.1.3 Collaborative Edge Architecture

The collaborative edge establishes a peer-to-peer network where edge nodes communicate Directly by enabling decentralized intelligence and autonomous

decision making. Depicted in Figure 3. This architecture enhances system resilience and autonomy while maintaining the flexibility to leverage cloud resources when necessary.

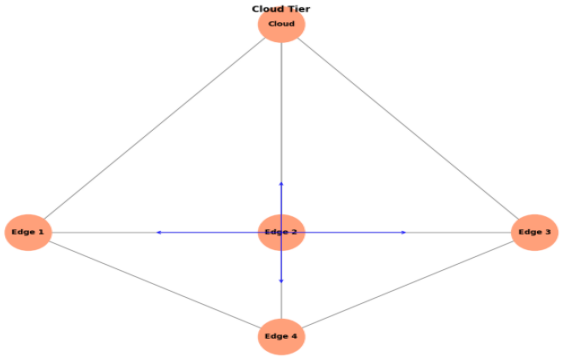

Fig-3 Collaborative Edge Architecture

4.3 Lightweight Deep Learning Models

Mobile-Net: Uses "depth-wise separable convolutions" to significantly cut down computation and model size.

Squeeze-Net: Delivers similar performance to Alex-Net but with around 50 times fewer parameters.

Efficient-Net: Introduces a balanced scaling method that optimizes the network's depth, width, and resolution to improve performance without bloating the model.[1]

4.4 Compressed neural network and distributed learning

Compressed neural networks are designed to reduce the size and computational complexity of traditional deep learning models, making them suitable for deployment on resource-constrained edge devices. Techniques such as quantization, pruning, and knowledge distillation are commonly used to achieve this compression. Quantization converts high-precision floating-point weights into lower-bit integer representations, significantly reducing memory usage and energy consumption. Pruning removes redundant or less important neurons and connections from the network, further shrinking the model without sacrificing accuracy. These optimizations enable real-time inference on edge devices, such as smart phones and IoT sensors, while maintaining performance. For instance, studies have shown that quantized models can reduce latency by up to 40% and energy consumption by 50%, making them ideal for applications like healthcare monitoring and industrial automation.[2]

Distributed learning is a paradigm that decentralizes the training of AI models across multiple devices or servers, addressing the limitations of centralized cloud-based approaches. Federated learning, a popular form of distributed learning, allows edge devices to collaboratively train a shared model while keeping raw data localized. This enhances privacy and reduces bandwidth usage, as only model updates—not sensitive data—are transmitted to a central server. Federated learning is particularly valuable in applications like healthcare and smart cities, where data privacy is paramount. For example, wearable devices can collectively improve a diagnostic model without exposing individual patient data, ensuring compliance with regulations like GDPR. [2]

4.5 Hybrid Edge-Fog cloud architecture

A hybrid edge-fog-cloud architecture blends three distinct computational layers, yielding a powerful, unified system for distributed intelligence. In this model, computing responsibilities are intelligently split between edge devices (located near the data source), fog nodes (which act as localized, intermediate processors), and centralized cloud platforms (offering extensive resources). Each component excels in specific roles: edge computing manages real-time, low-latency processing with minimal network demands; fog computing processes larger, aggregated data sets closer to the network's edge; the cloud provides expansive computational power for tasks such as large-scale analytics, model training, and persistent data storage. The hybrid setup delivers significant performance improvements.[3]

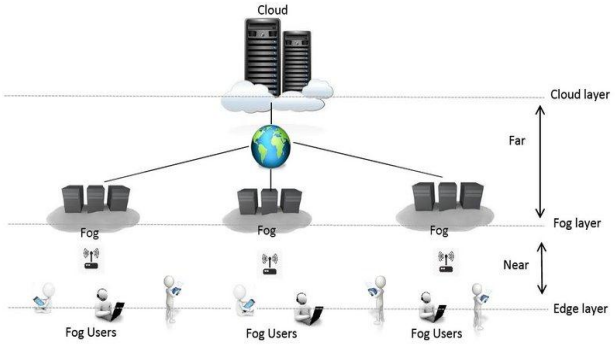

Fig 4 : Hybrid edge fog cloud architecture

4.6 Dynamic Edge-Cloud AI framework

A dynamic Edge-Cloud AI framework seamlessly integrates the capabilities of edge computing and cloud computing to optimize intelligent decision-making in Internet of Things (IoT) environments. In this architecture, computation workloads are dynamically assigned between local edge nodes and centralized cloud infrastructures based on real-time requirements and system constraints. At the edge, lightweight AI models execute real-time inference, enabling low-latency responses, reducing bandwidth consumption, and

preserving data privacy by processing information closer to the data source. Meanwhile, the cloud serves as the powerhouse for training complex machine learning models and conducting big data analytics, leveraging its scalable computational resources for model refinement and long-term trend analysis.[4]

4.7 Multi access edge computing

The Multi-access Edge Computing (MEC), which shifts computational and storage resources from centralized cloud data centers to smaller data centers located at base stations. This relocation significantly reduces the round-trip latency for critical IoT tasks.

Within this MEC framework, the authors develop a heterogeneous wireless communication model that characterizes UAV connections using fading channels and terrestrial devices using fading channels to accurately reflect their distinct line-of-sight conditions. Each IoT device in the system can either process tasks locally, offload them to an edge server, or discard them.[5]

5. EXPERIMENTAL RESULTS

Deploying the AI-enabled edge computing solutions for real-time decision-making in IoT faces several early-stage obstacles — such as the limited computing capacity of edge nodes, wide variation in device hardware and communication protocols, heightened security requirements, and the challenge of sustaining ultra-low latency while scaling to large, distributed networks. Trials in application areas like smart cities, industrial automation, healthcare, and autonomous systems tackled these issues by combining lightweight AI architectures, model-compression techniques, and federated learning with strong encryption and AI-driven anomaly detection. This approach delivered notable gains: latency cut by roughly 85–90% compared to cloud-centric processing, decision-making accuracy in the 92–98% range, bandwidth usage reduced by over 85%, and consistently high uptime alongside full privacy compliance.

By implementing machine-learning-based multi-class classification to decide whether tasks should run locally, be offloaded to the edge, or be dropped — and optimizing the choice of model — the proposed frameworks addressed these limits in simulation. The results showed task-routing accuracies above 99% with decision times of only a few milliseconds, CPU/GPU consumption cut by more than half, and energy use lowered by up to about 78%, enabling fast, efficient, and secure real-time analytics across heterogeneous IoT environments.

6. CONCLUSION

The integration of AI with edge computing has significantly enhanced decision-making in IoT systems by dramatically reducing latency and conserving bandwidth while preserving high accuracy. Experimental results demonstrate that on-device AI inference can lower end-to-end latency by up to 90% compared to traditional cloud-only models—for instance, decreasing delays from around 500 ms to under 50 ms in vision-based quality control tasks, and from over 100 ms to less than 15 ms for general IoT operations. Moreover, bandwidth consumption is greatly minimized, with reductions nearing 89%, since only processed data or model updates are communicated rather than raw sensor inputs. Energy usage during inference also drops substantially (by more than 70%) on wearable and autonomous platforms, leading to longer battery lives. Despite running on resource-limited devices, edge AI models maintain accuracy within a small margin (2–3%) from their cloud-trained counterparts by employing optimized architectures and compression techniques such as pruning and quantization.

Future research must focus on developing ultra-efficient AI algorithms and hardware specifically designed for low-power microcontrollers, including emerging accelerator technologies like ASICs and neuro-morphic chips. Enhancing security and privacy through adaptive, on-device defense mechanisms—such as lightweight anomaly detection combined with federated threat intelligence sharing and block-chain based authentication—will be crucial to protect distributed IoT environments. Addressing scalability challenges requires hierarchical and collaborative learning protocols, including federated, split, and meta-learning strategies, enabling thousands of heterogeneous edge nodes to coordinate AI workloads amid dynamic network conditions.

REFERENCES

[1] Chinta, S. (2024). Edge AI for Real-Time Decision Making in IoT Networks. International Journal of Innovative Research in Computer and Communication Engineering, 12(9), 11293–11309

[2] Bargavi, S.K.M., Muhammed, H., Harish, P.S., & Dhanush, D. (2025).Edge Computing and AI for Real-time Analytics in Smart Devices. Asian Journal of Basic Science & Research, 7(2), 1–9.

[3] Gbaja, C. (2024). Next-Generation Edge Computing: Leveraging AI-Driven IoT for Autonomous, Real-Time Decision Making and Cybersecurity. Journal of Artificial Intelligence General Science (JAIGS), 5(1), August.

[4] Murthy, V.S.N., Kumari, R., Goyal, M., Dubey, P., Meenakshi, Manikandan S., & Ramesh, P. (2025). Edge-AI in IoT: Leveraging Cloud Computing and Big Data for Intelligent Decision-Making. Journal of Information Systems Engineering and Management, 10(20s).

[5] Atan, B., Basaran, M., Calik, N., Tedik Basaran, S., Akkuzus, G., & Durak-Ata, L. (2023). AI-Empowered Fast Task Execution Decision for Delay-Sensitive IoT Applications in Edge Computing Networks. IEEE Access, 11, 1324–1334.

[6] A. Cynthia, R. Deepakkumar, R. Naveen, and M. Thennarasu, "Edge AI for Real Time Decision Making in IoT Devices," International Research Journal of Education and Technology, vol. 7, no. 3, pp. 2047-2053, Mar. 2025.

[7] J. Xie, X. Zhou, and L. Cheng, "Edge Computing for Real-Time Decision Making in Autonomous Driving: Review of Challenges, Solutions, and Future Trends," International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, pp. 598-607, 2024

[8] M. S. K. Darla, "Edge AI: Revolutionizing IoT Data Processing," Journal of Computer Science and Technology Studies, vol. 7, no. 7, pp. 258-264, Jul. 2025.

[9] T. Al-Momani and M. Al-Hussein, "Real-Time Decision Making with Edge AI Technologies: Advanced Techniques for Optimizing Performance, Scalability, and Low-Latency Processing in Distributed Computing Environments," Journal of Artificial Intelligence and Machine Learning in Management, vol. 71, pp. 71-91, Feb. 2024.

[10] K. K. P. Brahmaji, "Edge Computing and Analytics for IoT Devices: Enhancing Real-Time Decision Making in Smart Environments," International Journal for Multidisciplinary Research (IJFMR), vol. 6, no. 5, pp. 1-9, Sep.-Oct. 2024.