iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76740

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Role of AI in Edge Computing

Ms. Mohini Suresh Fulzele[1], Prof. Roshni Sonsare[2], Mr Sandeep Ramesh Sonaskar[3], Ms Neema Ukani[4]
Department of Computer Technology, YCCE Nagpur

Abstract: ***Edge computing has developed into an essential approach for addressing the shortcomings of cloud-based architectures by facilitating data processing near the sources of that data. The integration of Artificial Intelligence (AI) greatly improves edge computing by allowing for real-time data analysis, smart decision-making, and the operation of autonomous systems in environments with limited latency and bandwidth. This review article examines the significance of AI within edge computing, focusing on its architectures, supporting technologies, challenges, and various applications. A practical case study from the industry is presented to showcase the efficacy of AI-driven edge systems. Edge computing offers a crucial solution to the limitations of traditional cloud architectures by processing data closer to its source. The power of Edge computing is significantly enhanced through the incorporation of Artificial Intelligence (AI), enabling real-time data analysis, intelligent decision-making, and the operation of autonomous systems in bandwidth- and latency-constrained environments. This review article investigates the vital role of AI in edge computing. Specifically, it explores the foundational architectures, enabling technologies, inherent challenges, and diverse applications of AI-driven edge systems. An industry-based case study is also included to demonstrate the practical effectiveness of these systems.***

Keywords: Edge Computing, Artificial Intelligence, Edge AI, Federated Learning, Internet of Things.

I. INTRODUCTION

The exponential growth of Internet of Things (IoT) devices has resulted in massive data generation at the network edge. Traditional cloud-based processing models suffer from high latency, bandwidth constraints, and privacy issues. Edge computing addresses these challenges by processing data near its source. Artificial Intelligence (AI), when integrated with edge computing, enables intelligent, real-time decision-making without dependency on centralized cloud infrastructures. This integration, commonly referred to as Edge AI, is gaining attention across domains such as smart manufacturing, healthcare monitoring, autonomous vehicles, and smart cities.The proliferation of Internet of Things (IoT) devices has catalyzed an unprecedented surge in data generation, with this massive volume of information originating predominantly at the periphery of the network—the "edge." Conventional data processing paradigms, which rely on centralized cloud infrastructures, are increasingly proving inadequate for handling this deluge. These models are inherently constrained by several critical drawbacks, including high transmission latency, which is detrimental to time-sensitive applications; significant bandwidth consumption, leading to escalating operational costs; and considerable privacy and security vulnerabilities associated with moving sensitive data off-site. Edge computing has emerged as a transformative architectural solution specifically designed to circumvent these limitations. By positioning computational, storage, and networking resources closer to the data source, edge computing minimizes the geographical distance data must travel. This immediate proximity enables ultra-low latency processing, conserves network bandwidth, and inherently improves data privacy by keeping sensitive information localized. The true paradigm shift occurs with the seamless integration of Artificial Intelligence (AI) into this localized processing environment. When AI capabilities are deployed directly at the network edge, the system evolves beyond mere data processing to enable truly intelligent, real-time, and autonomous decision-making. This powerful synergy, universally recognized as Edge AI, represents a crucial technological frontier. Edge AI systems operate with a high degree of independence from centralized cloud infrastructures for day-to-day operations, allowing for immediate responsiveness even in environments with intermittent or poor network connectivity. The compelling advantages offered by Edge AI are driving its rapid adoption and specialized application across a diverse spectrum of high-stakes and performance-critical domains. In smart manufacturing, Edge AI enables predictive maintenance, real-time quality control via vision systems, and autonomous robotic coordination, maximizing uptime and efficiency. In healthcare monitoring, it facilitates continuous, on-device analysis of patient biometrics for immediate detection of critical events, such as cardiac anomalies, offering life-saving responsiveness. For autonomous vehicles, Edge AI is indispensable, as instantaneous processing of sensor data (LIDAR, cameras) for navigation, collision avoidance, and decision-making is a non-negotiable requirement for safety. Finally, within the framework of smart cities, Edge AI enhances public safety, optimizes traffic flow, manages energy consumption in real-time, and enables responsive utility services by processing environmental and infrastructure data where it is generated.

©IJRASET: All Rights are Reserved

II. BACKGROUND OF EDGE COMPUTING AND AI

Edge computing represents a fundamental shift in distributed computing, strategically extending the capabilities traditionally housed in centralized cloud data centers to the periphery of the network. This network edge encompasses a diverse range of devices, from simple gateways and routers to sophisticated industrial controllers and embedded systems. By processing data closer to its source of generation, edge computing dramatically improves system responsiveness, leading to reduced latency and enhanced real-time interaction. Furthermore, this localized processing significantly bolsters system reliability and autonomy, particularly in environments with intermittent or high-latency network connectivity.

The integration of Artificial Intelligence (AI), specifically through powerful techniques like machine learning (ML) and deep learning (DL), introduces transformative capabilities, enabling these edge systems to learn from operational data, recognize complex patterns, and make informed predictions or autonomous decisions. This synergy is termed Edge AI. However, deploying complex AI models in resource-constrained edge environments presents formidable technical challenges. The primary constraints revolve around the limited computation power available on typical edge devices, the restricted memory and storage capacity for housing large model parameters and training data, and the crucial need for energy efficiency, especially in battery-powered or remotely deployed systems.

Despite these constraints, the domain of Edge AI has seen rapid, innovative advancement. Recent progress in developing lightweight AI models—such as pruned, quantized, or knowledge-distilled models—significantly reduces their computational and memory footprint without severe loss of accuracy. Concurrently, the emergence of specialized hardware accelerators, including Application-Specific Integrated Circuits (ASICs), Field-Programmable Gate Arrays (FPGAs), and specialized AI-focused processors (like NPUs or TPUs), is purpose-built to execute AI inference tasks with high efficiency and low power consumption. These parallel innovations are collectively making the practical and widespread deployment of sophisticated Edge AI applications across diverse sectors, including smart cities, industrial IoT, and autonomous vehicles, not only feasible but increasingly commonplace.

III.ROLE OF AI IN EDGE COMPUTING

AI plays a vital role in enhancing the functionality of edge computing systems. It enables real-time data processing and inference at the edge, reducing reliance on cloud communication. AI-based algorithms optimize resource utilization through intelligent scheduling and load balancing. Security is improved through anomaly detection and intrusion prevention mechanisms deployed locally. Furthermore, AI enables context-aware and personalized services by learning from localized data patterns.Artificial Intelligence (AI) is a cornerstone technology for significantly enhancing the capabilities and performance of modern edge computing systems. This integration transforms edge nodes from simple data collectors into intelligent processing units, fundamentally changing how data is handled and utilized.

A. Real-Time Data Processing and Inference:

A primary benefit of integrating AI into edge computing is the enablement of real-time data processing and inference directly at the source. By deploying lightweight, optimized machine learning models (like deep neural networks) onto edge devices, systems can analyze sensor data, video streams, and device logs instantly. This local intelligence dramatically reduces the latency associated with transmitting vast amounts of raw data to a centralized cloud for analysis, thereby making critical applications—such as autonomous vehicles, industrial automation, and remote patient monitoring—feasible and reliable. The result is quicker decision-making and immediate system response to dynamic environmental changes.

B. Optimization of Resource Utilization:

AI-based algorithms are instrumental in optimizing resource utilization across the distributed network of edge devices. Through continuous learning and predictive analytics, AI implements intelligent scheduling and load balancing techniques. For instance, it can predict future traffic patterns or resource needs and dynamically allocate computing power, memory, and bandwidth across the edge nodes. This prevents bottlenecks, ensures efficient power management (critical for battery-operated IoT devices), and maximizes the throughput of the entire edge infrastructure, leading to lower operational costs and improved energy efficiency.

C. Enhanced Security Mechanisms:

The security posture of edge computing is substantially improved through the deployment of AI-driven anomaly detection and intrusion prevention mechanisms deployed locally.

Traditional security relies on signature-based detection, but AI models, trained on normal operational data, can identify subtle deviations in network traffic, system calls, or device behavior that signify a zero-day attack or an insider threat. By running these detection algorithms at the edge, potential threats can be identified and mitigated instantly, isolating compromised devices before the attack can propagate to the wider network or the core cloud infrastructure.

D. Context-Aware and Personalized Services

AI enables edge systems to offer context-aware and personalized services by effectively learning from localized data patterns. Since AI models process data generated by specific users, environments, or machines locally, they can develop highly specific and accurate insights that would be obscured or averaged out in a massive cloud dataset. This leads to tailored experiences, such as personalized recommendations in retail environments, predictive maintenance schedules specific to individual machinery in a factory, or adaptive traffic light control based on real-time local vehicle flow, significantly boosting user experience and system efficiency.

IV. ENABLING TECHNOLOGIES FOR EDGE

AI at the edge, often referred to as Edge AI, represents a paradigm shift where Artificial Intelligence processing moves from centralized cloud servers to the devices and sensors at the network's periphery. This decentralized approach offers numerous benefits, including reduced latency, enhanced data privacy, lower bandwidth consumption, and greater system reliability. The successful deployment of AI at the edge is critically supported by several interconnected technological innovations:

A. Model Optimization Techniques:

These methods are fundamental to shrinking the resource footprint of sophisticated AI models, making them viable for resource-constrained edge devices.

1) Pruning: This technique removes redundant weights and connections in a neural network, reducing the model's size and computational load without significant loss in accuracy. Both structured (removing entire channels or layers) and unstructured (removing individual weights) pruning are employed.

2) Quantization: This process reduces the precision of the numerical representations of weights and activations, typically moving from 32-bit floating-point numbers (FP32) to lower precision formats like 8-bit integers (INT8). This dramatically decreases memory usage and speeds up inference on hardware optimized for integer arithmetic.

3) Knowledge Distillation: A "teacher" model (a large, high-performing model) trains a smaller, more efficient "student" model. The student learns to mimic the teacher's outputs, achieving comparable performance with a significantly smaller model size, suitable for edge deployment.

B. Lightweight AI Architectures:

These are novel neural network designs inherently optimized for efficiency, aiming for high performance with minimal computational cost and power consumption.

1) MobileNet: Developed by Google, MobileNet models utilize depthwise separable convolutions to drastically reduce the number of parameters and computations compared to standard convolutional networks, making them ideal for mobile and embedded vision applications.

2) ShuffleNet: This architecture further enhances efficiency by employing pointwise group convolutions and a channel shuffle operation to reduce computational cost while maintaining representation power, often used in low-power mobile scenarios.

3) Other Examples: Architectures like EfficientNet and various forms of SqueezeNet also fall into this category, focusing on balancing model size, speed, and accuracy through compound scaling and efficient building blocks.

C. Edge Hardware Accelerators:

Specialized hardware is crucial for achieving high-speed inference and energy efficiency on edge devices, overcoming the limitations of general-purpose CPUs.

1) Dedicated Neural Processing Units (NPUs): These are chips specifically designed to accelerate the matrix multiplications and activation functions that dominate neural network computation. They offer superior power efficiency for AI tasks compared to CPUs and GPUs.

2) Graphics Processing Units (GPUs): While powerful, miniature GPUs are often used in high-end edge devices (like autonomous vehicles) to handle complex parallel processing tasks for computer vision and large model inference.
3) Tensor Processing Units (TPUs): Developed by Google, while primarily known for cloud acceleration, customized, low-power variants are increasingly being adapted for edge applications, particularly those focused on TensorFlow workloads.
4) FPGAs (Field-Programmable Gate Arrays): These offer a balance of flexibility and performance, allowing for custom hardware acceleration pipelines tailored to specific AI models.

D. Federated Learning (FL):

This distributed machine learning approach enables collaborative model training across a multitude of edge devices without requiring the raw data to be centralized.

1) Preserving Data Privacy: Only locally computed model updates (gradients or weights) are shared with a central server, ensuring that sensitive user data remains on the device, addressing critical privacy and regulatory concerns (e.g., GDPR, HIPAA).
2) Collaborative Training: FL allows the global model to benefit from the diverse, real-world data generated by millions of edge devices, leading to a more robust and generalized final model.
3) Communication Efficiency: Smart aggregation strategies are used to minimize the amount of data transferred, which is a significant advantage in environments with limited or unreliable bandwidth.

E. Edge Orchestration Platforms

These software frameworks are essential for managing the lifecycle and complexity of deploying AI across a vast, distributed, and often heterogeneous collection of edge nodes.

1) Model Deployment and Lifecycle Management: They handle the seamless deployment of optimized AI models to the correct edge devices, manage version control, and facilitate over-the-air updates.
2) Scaling and Monitoring: These platforms provide tools for monitoring the performance, health, and energy consumption of AI models on various edge nodes, allowing for dynamic scaling and resource allocation.
3) Heterogeneous Edge Node Management: They abstract away the differences between various hardware and operating systems (e.g., Linux, Android, specialized RTOS), allowing developers to manage the entire fleet through a unified interface.

The edge computing devices used in the industry as shown in Fig. 1. which sends the data to cloud server for AI analysis

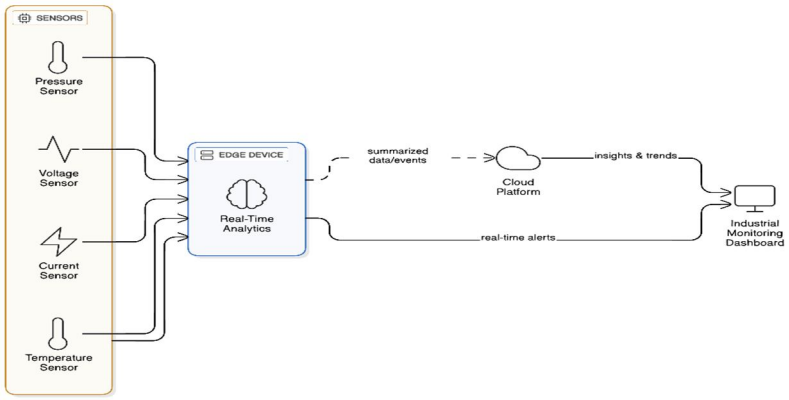


Fig. 1 A Edge computing System flow in the industry

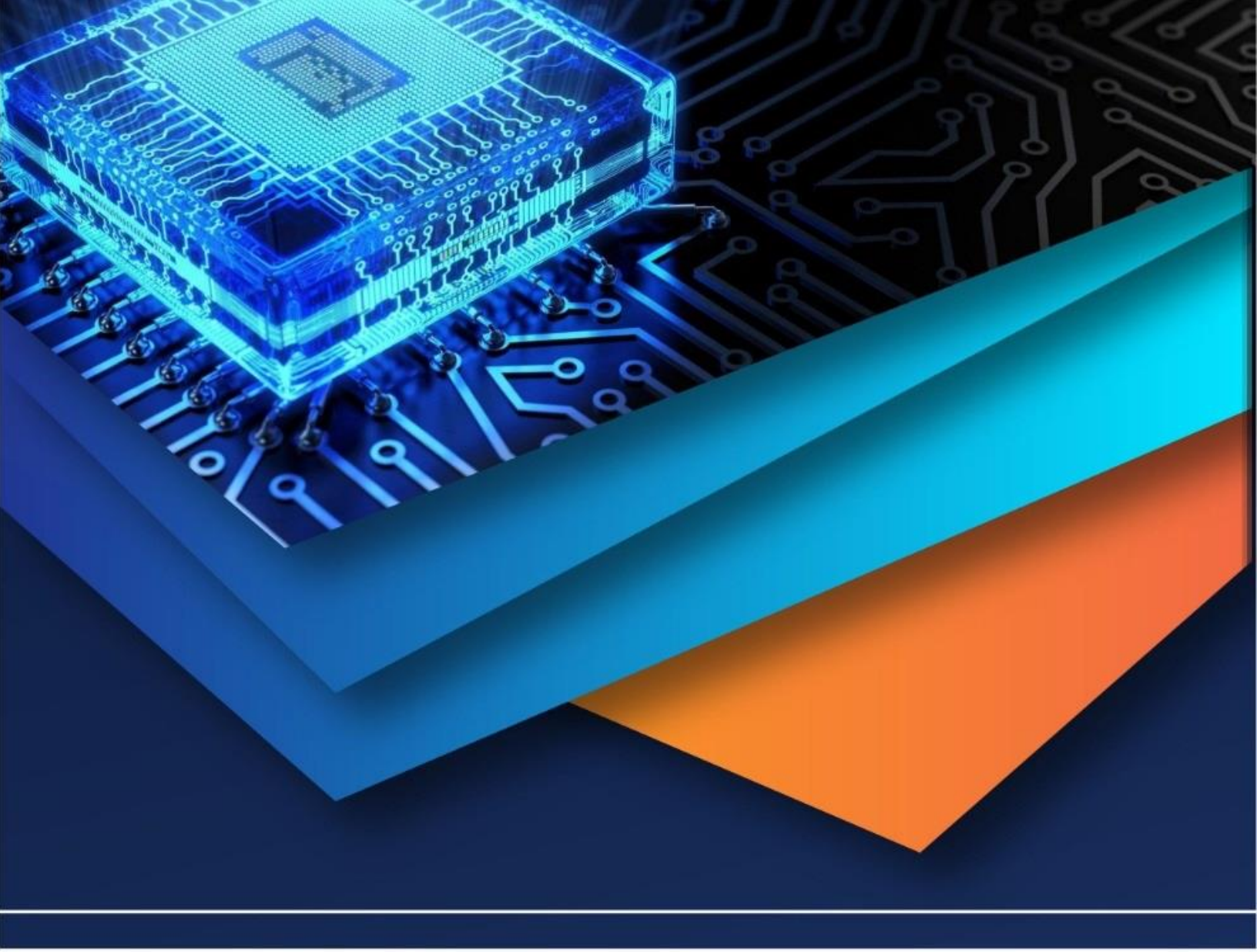
V. CONCLUSION

AI is vital for maximizing edge computing's potential, enabling intelligent, low-latency, and privacy-aware applications. This review examined AI's role in edge computing, enabling technologies, challenges, and an industrial case study. Edge AI's continued evolution will significantly impact next-generation intelligent systems.

REFERENCES

[1] W. Shi et al., "Edge computing: Vision and challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, 2016.
[2] M. Satyanarayanan, "The emergence of edge computing," IEEE Computer, vol. 50, no. 1, pp. 30–39, 2017.
[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
[4] C. Zhang et al., "Edge intelligence: Paving the last mile of artificial intelligence," Proc. IEEE, vol. 107, no. 8, pp. 1738–1762, 2019.

[5] S. Han et al., "Deep compression," ICLR, 2016.
[6] A. Howard et al., "MobileNets: Efficient CNNs for mobile vision applications," arXiv:1704.04861, 2017.
[7] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," AISTATS, 2017.
[8] X. Chen and X. Ran, "Deep learning with edge computing," Proc. IEEE, vol. 107, no. 8, pp. 1655–1674, 2019.
[9] M. Chiang et al., "Fog and IoT: An overview of research opportunities," IEEE Internet of Things Journal, vol. 3, no. 6, pp. 854–864, 2017.
[10] S. Teerapittayanon et al., "Distributed deep neural networks over the cloud, the edge and end devices," IEEE ICDCS, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)