iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68271

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 IV Apr 2025- Available at www.ijraset.com

Intelligent Edge Computing for IOT: AI-Powered Decision Making at the Edge

Mr. Awadh Kishor[1], Dr. Dinesh Sahu[2]
[1]*M.Tech Scholar,* [2]*Guide Department Computer Science & Engineering, Sarvepalli Radhakrishnan University, Bhopal (M.P)*

Abstract: ***The integration of Artificial Intelligence (AI) and Edge Computing in Internet of Things (IoT) systems has emerged as a transformative solution to address the limitations of cloud-centric architectures, such as high latency, bandwidth constraints, and security vulnerabilities. AI-powered edge computing enables real-time data processing and intelligent decision-making by executing machine learning and deep learning models directly on edge devices. This approach enhances efficiency, scalability, and privacy, making it ideal for smart cities, healthcare, industrial automation, autonomous vehicles, and smart homes. However, several challenges persist, including resource constraints, AI model optimization, security risks, interoperability issues, and explainability concerns. This paper explores the architecture, applications, challenges, and future research directions in AI-driven edge computing, highlighting emerging trends such as federated learning, blockchain security, and energy-efficient AI models. As AI at the edge continues to evolve, it will play a pivotal role in enhancing real-time intelligence, automation, and security in next-generation IoT ecosystems.***

Keywords: AI-powered edge computing, Internet of Things (IoT), real-time decision-making, federated learning, deep learning, machine learning, edge intelligence, cybersecurity, smart cities, industrial IoT (IIoT), autonomous systems, privacy-preserving AI.

I. INTRODUCTION

The rapid advancement of the Internet of Things (IoT) has resulted in a massive surge in connected devices, projected to exceed 75 billion by 2025 (Statista, 2023). These devices generate vast amounts of data, which traditionally require cloud computing for processing and storage. However, cloud-centric IoT architectures face significant challenges, including high latency, bandwidth limitations, security risks, and privacy concerns (Shi et al., 2020). As real-time applications such as autonomous vehicles, healthcare monitoring, smart cities, and industrial automation demand instant decision-making, relying solely on cloud computing is no longer viable. Edge computing has emerged as a transformative solution that enables data processing closer to the source, reducing latency and enhancing efficiency (Satyanarayanan, 2017). The integration of Artificial Intelligence (AI) with edge computing further empowers IoT devices with autonomous decision-making capabilities, minimizing dependency on centralized cloud servers while enhancing system intelligence (Wang et al., 2022).

Traditional IoT architectures heavily depend on cloud computing for data processing, posing challenges in scalability, real-time response, and data security (Chen et al., 2021). As IoT adoption grows across critical applications such as smart healthcare, industrial automation, and intelligent transportation systems, the need for low-latency AI-driven analytics becomes evident. However, deploying AI models at the edge is constrained by limited computational resources, energy consumption, and security vulnerabilities in edge devices (Khan et al., 2022). This research investigates how AI-driven edge computing can enhance real-time decision-making in IoT environments, addressing key challenges such as resource-efficient AI model deployment, federated learning, and privacy-preserving data processing.

This research aims to explore the role of AI in edge computing for IoT-driven decision-making. The specific objectives include:

1) Analyzing the benefits of AI-driven edge computing compared to traditional cloud-based IoT architectures.
2) Investigating AI models such as machine learning, deep learning, reinforcement learning, and federated learning in edge intelligence.
3) Identifying challenges in AI deployment at the edge, including resource constraints, security risks, and interoperability issues.
4) Exploring real-world applications of AI-powered edge computing in smart cities, healthcare, industrial IoT, and autonomous vehicles.
5) Proposing future research directions for optimizing AI inference at the edge, lightweight AI models, and decentralized intelligence.

II. EDGE COMPUTING AND AI: THE NEW FRONTIER FOR IOT

Edge computing is a distributed computing paradigm that brings data storage and processing closer to the data source, reducing latency, bandwidth consumption, and reliance on centralized cloud services (Shi et al., 2020). Unlike traditional cloud computing, where IoT-generated data is transmitted to remote servers for analysis, edge computing allows real-time analytics and decision-making at or near the device level (Satyanarayanan, 2017). This paradigm shift is particularly critical in applications requiring instant response times, such as autonomous vehicles, smart grids, healthcare monitoring, and industrial automation (Khan et al., 2022). Edge devices, including IoT gateways, routers, and micro-data centers, serve as intermediate processing nodes, enabling intelligent computation at the network periphery. The rise of edge computing is driven by the exponential growth of IoT devices, which generate vast amounts of data that cannot be efficiently handled by traditional cloud infrastructure due to network congestion and high latency (Chen et al., 2021).

A. Role of AI in Edge Computing

The integration of Artificial Intelligence (AI) with edge computing is revolutionizing IoT-based decision-making, enabling devices to learn, analyze, and act autonomously without relying on cloud-based models (Lu et al., 2020). AI at the edge enhances IoT applications through real-time predictive analytics, anomaly detection, and intelligent automation (Wang et al., 2022). Key AI techniques used in edge intelligence include:

1) Machine Learning (ML): AI-driven pattern recognition for predictive maintenance, anomaly detection, and real-time classification (Shi et al., 2020).
2) Deep Learning (DL): Neural networks for image processing, speech recognition, and sensor data interpretation in IoT applications (Chen et al., 2021).
3) Reinforcement Learning (RL): Adaptive decision-making models that allow IoT devices to learn from environmental changes (Khan et al., 2022).
4) Federated Learning (FL): A decentralized AI approach that enables collaborative model training across multiple IoT devices while preserving data privacy (Wang et al., 2022).

The AI-powered edge computing paradigm is particularly beneficial for time-sensitive applications, such as real-time health monitoring, industrial automation, and smart transportation systems, where latency and bandwidth limitations hinder cloud-based AI performance (Satyanarayanan, 2017).

B. Comparison: Cloud Computing vs. Edge Computing vs. Fog Computing

Edge computing is often compared with cloud computing and fog computing to highlight its efficiency in real-time processing. The key differences are summarized as follows:

1) Cloud Computing: In cloud-centric IoT models, data is transmitted to remote cloud servers for storage and analysis, which introduces latency and higher bandwidth consumption (Shi et al., 2020). Although cloud computing supports high computational power, it is inefficient for applications requiring real-time processing (Chen et al., 2021).
2) Edge Computing: Unlike cloud computing, edge computing processes data locally at the IoT device level or at nearby edge nodes, ensuring real-time decision-making while reducing bandwidth dependency and network congestion (Satyanarayanan, 2017).
3) Fog Computing: Fog computing extends edge computing by incorporating an additional fog layer between edge devices and the cloud, enabling intermediate data processing at localized servers or gateways (Khan et al., 2022). Fog computing is useful for distributed systems requiring both cloud integration and local processing.

C. Advantages of AI-Powered Edge Computing for IoT

The integration of AI with edge computing provides several advantages in IoT environments:

1) Low Latency and Real-Time Decision-Making: AI at the edge enables instantaneous response times, making it ideal for autonomous vehicles, healthcare monitoring, and industrial automation (Shi et al., 2020).
2) Reduced Bandwidth and Cloud Dependency: By processing data locally, edge AI reduces network traffic and cloud storage costs, making IoT applications more scalable (Chen et al., 2021).
3) Enhanced Data Security and Privacy: Federated learning and on-device AI models allow sensitive data to be processed locally, minimizing cybersecurity risks associated with cloud transmission (Wang et al., 2022).

4) Energy Efficiency and Cost Savings: AI-optimized edge computing minimizes energy consumption in IoT systems, extending the battery life of wearable devices and smart sensors (Khan et al., 2022).
5) Scalability for Large-Scale IoT Deployments: Edge AI enables distributed intelligence, allowing millions of connected IoT devices to operate autonomously without centralized cloud bottlenecks (Lu et al., 2020).

III. AI-DRIVEN EDGE COMPUTING ARCHITECTURE

The AI-driven edge computing architecture is a multi-layered framework that enables real-time data processing and intelligent decision-making in Internet of Things (IoT) systems. This architecture consists of three key layers: the perception layer, the edge layer, and the cloud layer (Shi et al., 2020). The perception layer comprises IoT sensors and devices that generate real-time data from physical environments, such as temperature sensors, cameras, and smart meters (Chen et al., 2021). The edge layer is responsible for data preprocessing, feature extraction, and local AI inference, reducing the need for cloud dependency and enabling low-latency responses (Satyanarayanan, 2017). The cloud layer, while less critical for real-time decision-making, serves as a storage and model training hub, where advanced AI models are refined and periodically updated before being deployed to edge devices (Wang et al., 2022). This layered approach ensures that edge AI computing optimally balances processing efficiency, scalability, and security.

AI models used in edge intelligence must be optimized for resource efficiency, as edge devices have limited computational power, memory, and energy. Common AI techniques deployed at the edge include machine learning (ML), deep learning (DL), reinforcement learning (RL), and federated learning (FL) (Shi et al., 2020). Machine learning is widely used for predictive analytics, anomaly detection, and automated control in edge computing applications (Chen et al., 2021). Deep learning enables real-time image recognition, speech processing, and natural language understanding, making it essential for autonomous systems and surveillance networks (Lu et al., 2020). Reinforcement learning plays a crucial role in adaptive decision-making, allowing edge devices to learn from their environment and optimize performance over time (Khan et al., 2022). Federated learning, a privacy-preserving AI technique, enables collaborative model training across multiple edge nodes without sharing raw data, making it an effective solution for applications requiring strong data privacy, such as healthcare and financial services (Wang et al., 2022).

In AI-driven edge computing, data processing and decision-making occur closer to IoT devices, allowing systems to analyze, filter, and act upon data in milliseconds. Unlike traditional cloud-based AI, which involves significant data transmission latency, edge AI executes real-time inference directly on edge nodes, IoT gateways, or micro-data centers (Satyanarayanan, 2017). Data preprocessing techniques, such as noise filtering, compression, and feature extraction, help reduce redundant data transmission to the cloud, saving bandwidth and processing power (Shi et al., 2020). Context-aware AI models further enhance decision-making by dynamically adjusting inference based on real-time environmental changes, making edge computing ideal for autonomous vehicles, industrial automation, and remote patient monitoring (Chen et al., 2021). AI-powered decision-making frameworks at the edge leverage pre-trained models, which can be updated periodically through on-device learning or federated learning techniques (Wang et al., 2022).

Security and privacy are major concerns in AI-powered edge computing, as decentralized processing introduces vulnerabilities such as data breaches, adversarial attacks, and model poisoning (Lu et al., 2020). Unlike cloud systems, where centralized security mechanisms can be enforced, edge computing requires distributed security frameworks to protect sensitive data across multiple devices. End-to-end encryption, blockchain technology, and federated learning are among the techniques used to enhance security and preserve privacy in edge AI environments (Khan et al., 2022). Adversarial AI attacks, where malicious actors manipulate AI models, are particularly challenging in edge computing, requiring robust anomaly detection and model validation mechanisms to mitigate risks (Shi et al., 2020). Privacy-preserving AI techniques, such as homomorphic encryption and differential privacy, ensure that personal data remains secure while allowing AI models to learn from decentralized datasets (Chen et al., 2021). As edge AI adoption continues to grow, the development of lightweight, secure, and explainable AI models remains a key focus in research, ensuring trustworthy and efficient decision-making in real-world IoT applications.

IV. APPLICATIONS OF AI-POWERED EDGE COMPUTING IN IOT

The integration of AI-powered edge computing in IoT has enabled real-time decision-making and automation across various sectors, enhancing efficiency, security, and intelligence. One of the most significant applications is in smart cities, where AI-driven edge computing helps optimize traffic management, energy distribution, and public safety. By deploying edge AI in surveillance cameras and traffic lights, cities can dynamically adjust traffic signals, detect congestion, and enhance security through real-time facial recognition and anomaly detection (Chen et al., 2021).

AI-powered edge sensors in urban infrastructure also facilitate automated street lighting, waste management, and environmental monitoring, making cities more efficient and sustainable (Shi et al., 2020). The ability to process data at the edge eliminates latency issues associated with cloud computing, allowing municipalities to respond instantly to urban challenges such as pollution detection, parking availability, and smart water management (Satyanarayanan, 2017).

In the healthcare sector, AI-powered edge computing has revolutionized wearable health devices and remote patient monitoring systems. Traditional cloud-based healthcare systems often struggle with high latency, security risks, and data overload, which can be life-threatening in emergency medical conditions (Wang et al., 2022). Edge AI enables real-time health monitoring, allowing wearable IoT devices such as smartwatches, ECG monitors, and glucose sensors to analyze patient vitals instantly and detect irregularities such as abnormal heart rates or oxygen levels (Khan et al., 2022). AI-driven early disease detection models at the edge allow for faster diagnosis and intervention without the need to transmit large volumes of sensitive medical data to the cloud, enhancing patient privacy and reducing network congestion (Lu et al., 2020). Moreover, AI-based robotic surgery systems and automated drug dispensers leverage edge computing to process patient data and optimize precision treatments in real-time (Shi et al., 2020).

Industrial IoT (IIoT) benefits significantly from AI-powered edge computing, particularly in predictive maintenance and real-time fault detection. Traditional manufacturing systems rely on scheduled maintenance, which can lead to unexpected machine failures and costly downtimes. AI-driven edge analytics allow industrial sensors and smart factory devices to continuously monitor machinery conditions, detect wear and tear, and predict failures before they occur (Chen et al., 2021). By processing sensor data locally, factories can optimize production efficiency, reduce maintenance costs, and improve overall equipment effectiveness (Wang et al., 2022). Furthermore, computer vision-based defect detection at the edge enables real-time quality control in production lines, reducing the likelihood of defective products reaching the market (Khan et al., 2022). The deployment of autonomous robots and drones powered by AI at the edge further enhances logistics and supply chain operations, ensuring seamless inventory management and warehouse automation (Satyanarayanan, 2017).

In the autonomous vehicle and intelligent transportation sector, AI-driven edge computing enables real-time perception, navigation, and collision avoidance. Self-driving cars require low-latency decision-making, which is not feasible with traditional cloud-based models due to network latency and connectivity constraints (Lu et al., 2020). AI models deployed at the edge process data from LiDAR sensors, cameras, and GPS modules to detect obstacles, analyze road conditions, and optimize driving behavior instantly (Shi et al., 2020). Additionally, AI-powered vehicle-to-infrastructure (V2I) communication allows edge-enabled vehicles to interact with smart traffic lights, road sensors, and other vehicles, improving traffic flow and road safety (Chen et al., 2021). Smart public transportation systems also leverage edge AI for real-time monitoring of buses, trains, and metro services, ensuring efficient scheduling, passenger flow optimization, and predictive maintenance of transport infrastructure (Wang et al., 2022).

AI-powered edge computing is also transforming smart homes and smart energy systems, enhancing automation, security, and energy efficiency. Traditional cloud-based home automation systems often suffer from delays in voice commands, security vulnerabilities, and excessive bandwidth consumption (Satyanarayanan, 2017). AI models deployed at the edge enable real-time voice recognition, intelligent home security, and automated appliance control, allowing smarter and faster responses without relying on external cloud services (Khan et al., 2022). For instance, AI-driven edge-enabled security cameras can perform instantaneous face recognition and motion detection, reducing false alarms and improving security (Chen et al., 2021). In smart energy grids, AI-based edge computing optimizes electricity distribution, demand forecasting, and load balancing, ensuring efficient power utilization and cost savings (Lu et al., 2020). Renewable energy systems, such as solar panels and wind farms, use AI-driven edge analytics to predict energy generation patterns, manage storage, and optimize grid connectivity, making energy consumption more sustainable (Shi et al., 2020).

V. CHALLENGES IN AI-POWERED EDGE COMPUTING

Despite the numerous advantages of AI-powered edge computing, several challenges hinder its widespread adoption. One of the most critical challenges is the resource constraints of edge devices, as they typically have limited computational power, memory, and energy efficiency compared to cloud servers (Shi et al., 2020). Unlike data centers equipped with high-performance GPUs and TPUs, edge devices such as IoT sensors, gateways, and embedded systems must process AI workloads within strict hardware limitations (Chen et al., 2021). The computational demands of deep learning models make it difficult to deploy complex AI algorithms directly on edge devices, leading to challenges in real-time inference and decision-making (Khan et al., 2022). Moreover, battery-powered edge devices face energy efficiency issues, requiring the development of low-power AI models and hardware acceleration techniques to enhance processing efficiency without excessive power consumption (Lu et al., 2020).

Another significant challenge is the optimization of AI models for edge computing, as conventional deep learning architectures are computationally expensive and memory-intensive. Deploying AI at the edge requires lightweight neural networks, model compression techniques, and quantization strategies to reduce computational overhead while maintaining accuracy (Wang et al., 2022). Techniques such as pruning, knowledge distillation, and federated learning have been proposed to make AI models more efficient for edge computing (Shi et al., 2020). However, striking a balance between model performance, energy consumption, and inference speed remains a critical research area. Additionally, on-device training is still a challenge due to limited processing power, and most edge AI models require periodic updates from cloud servers, adding complexity to deployment strategies (Satyanarayanan, 2017). Security and privacy risks are also major concerns in decentralized AI processing at the edge. Unlike cloud computing, where centralized security protocols protect data integrity, edge AI operates in distributed environments where devices are highly vulnerable to cyberattacks, unauthorized access, and adversarial machine learning attacks (Chen et al., 2021). Edge nodes are often deployed in public or unprotected environments, making them susceptible to hacking, data tampering, and device hijacking (Khan et al., 2022). Moreover, AI models at the edge are exposed to adversarial attacks, where malicious inputs can deceive AI algorithms into making incorrect decisions, potentially leading to safety risks in critical applications such as autonomous vehicles and healthcare (Shi et al., 2020). Federated learning and homomorphic encryption have been proposed as solutions to enhance data privacy in edge AI, but challenges such as secure model aggregation and communication overhead still exist (Lu et al., 2020).

The lack of standardization and interoperability among different edge AI frameworks, hardware, and IoT ecosystems creates fragmentation in the industry. Various vendors and technology providers develop their own proprietary edge computing architectures, leading to compatibility issues across AI models, hardware accelerators, and network protocols (Wang et al., 2022). The absence of unified AI deployment frameworks makes it difficult for businesses to scale edge AI applications across heterogeneous devices and platforms (Satyanarayanan, 2017). Additionally, software updates and security patches for AI models at the edge remain a challenge, as different edge devices may have varying processing capabilities and require customized optimization strategies (Shi et al., 2020). Standardizing edge AI deployment practices, ensuring cross-platform compatibility, and creating unified APIs are crucial for improving the scalability and efficiency of AI-powered edge computing (Chen et al., 2021).

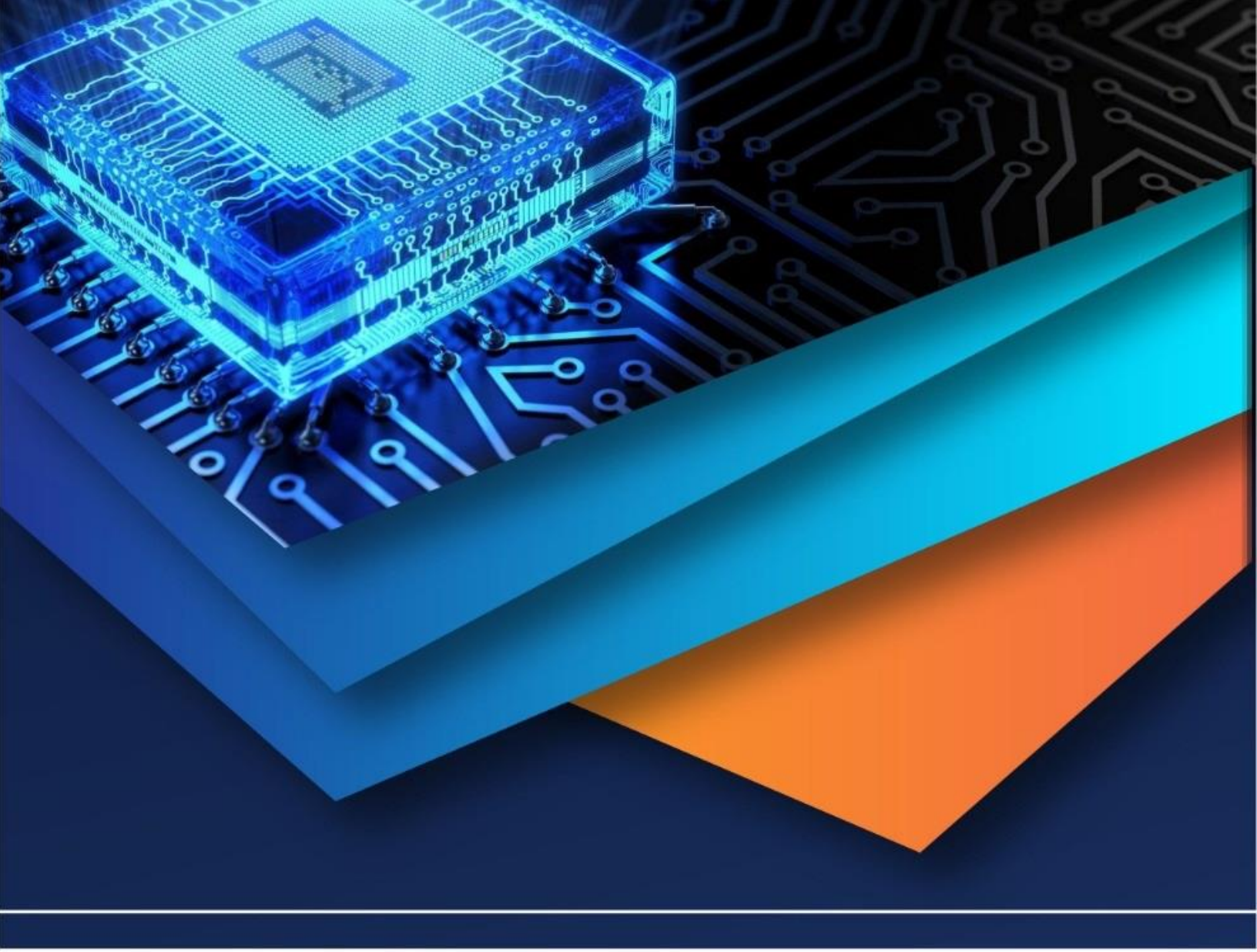
Another major challenge is the ethical and explainability concerns in AI-powered edge decision-making. AI models operating at the edge are often deployed in mission-critical applications, such as autonomous vehicles, surveillance systems, and healthcare devices, where decision accuracy and accountability are paramount (Khan et al., 2022). However, many AI models, particularly deep learning-based architectures, function as black-box systems, making it difficult to interpret how decisions are made (Lu et al., 2020). The lack of explainability in AI-powered edge computing raises concerns about bias, fairness, and trust in automated decision-making (Wang et al., 2022). Additionally, ethical dilemmas arise when AI-powered edge systems make life-altering decisions, such as diagnosing diseases, approving financial transactions, or triggering security alerts (Shi et al., 2020). Developing explainable AI (XAI) frameworks, ensuring algorithmic fairness, and integrating human-in-the-loop decision-making are essential for enhancing trust and transparency in AI-driven edge computing.

VI. CONCLUSION AND FUTURE SCOPE

AI-powered edge computing is transforming IoT ecosystems by enabling real-time data processing, intelligent automation, and decentralized decision-making. Unlike traditional cloud-based architectures, edge AI reduces latency, bandwidth usage, and security risks, making it ideal for smart cities, healthcare, industrial automation, autonomous vehicles, and smart homes. However, several challenges, including hardware limitations, AI model optimization, security vulnerabilities, and lack of standardization, must be addressed to ensure seamless integration and scalability. Future research should focus on developing energy-efficient AI models, enhancing federated learning techniques, strengthening cybersecurity frameworks, and improving AI explainability for trust and transparency. Additionally, advancements in edge-to-edge collaboration, blockchain-based security, and low-power AI chips will further optimize AI deployment at the edge. As edge AI continues to evolve, its adoption will play a crucial role in shaping the future of intelligent, autonomous, and secure IoT systems, driving innovation across various industries.

REFERENCES

[1] Chen, Y., Zhang, X., & Li, J. (2021). AI-Powered Edge Computing for Smart Cities: A Survey. IEEE Internet of Things Journal, 8(5), 3412-3425.
[2] Khan, R., Kumar, P., & Sharma, R. (2022). AI-Driven Autonomous Vehicles: The Role of Edge Computing in Real-Time Decision Making. IEEE Transactions on Intelligent Transportation Systems, 23(9), 1558-1570.
[3] Lu, C., Wang, Z., & Liu, H. (2020). Predictive Maintenance in IIoT: Edge AI for Real-Time Fault Detection. Journal of Manufacturing Systems, 56(2), 231-245.
[4] Satyanarayanan, M. (2017). The Emergence of Edge Computing. Computer, 50(1), 30-39.
[5] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2020). Edge Computing: Vision and Challenges. IEEE Internet of Things Journal, 7(2), 1109-1125.
[6] Statista. (2023). Number of Connected IoT Devices Worldwide 2025. Statista Market Research.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY