IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Edge Intelligence: AI at the Network Edge

[1]Dr.RomaSaxena, [2]Dr. Anuj Kumar
[1]Assistant Professor, [2]Professor
[1]Department of Computer Science, Bareilly College Bareilly

Abstract: In this research paper, a detailed overview of Edge Artificial Intelligence (Edge AI) is given in which processing capabilities of AI are embedded into edge devices near the sources of data generation. This paper discusses the history, processing models, and implementations of Edge AI in a variety of industries and how it can support real-time and low-latency decision-making as well as minimize bandwidth reliance and improve data privacy. The paper also explores resource limitation, security vulnerability and scalable issue as some of the major challenges that limit the implementation of Edge AI at present. Moreover, the future research directions are outlined, such as improved specialized hardware, lightweight, and adaptable AI models, federated learning, strong security frameworks, and edge-cloud infrastructure coordination. It will be used to guide continued innovation and deployment of Edge AI technologies to develop smarter, more efficient, and privacy-vaulting intelligent systems, and revolutionize industries and make AI-driven applications ubiquitous.

Index Terms - EdgeAI, Edge Computing, Artificial intelligence, Internet of things(IoT), Network Edge

I. INTRODUCTION

Edge Artificial Intelligence (Edge AI) represents an advanced computing paradigm that enables the deployment and execution of AI workflows distributed across a continuum—from centralized cloud infrastructures to decentralized network endpoints. In this context, the network "edge" refers to the outermost points within a system, including user-end devices and Internet of Things (IoT) nodes (Adeoye, 2025).

Unlike the conventional approach, where AI models are primarily developed, trained, and executed within centralized cloud environments, a framework often referred to as Cloud AI, Edge AI integrates the principles of Artificial Intelligence and Edge Computing. Edge Computing focuses on relocating computation and data processing closer to the physical sources of data generation. This architectural shift reduces latency, optimizes bandwidth utilization, enhances privacy, and improves real-time responsiveness in distributed systems.

Artificial Intelligence (AI) encompasses a broad domain of computer science dedicated to designing systems that can perform tasks requiring cognitive functions traditionally associated with human intelligence, such as learning, reasoning, and decision-making. Within this domain, Machine Learning (ML) constitutes a specialized subfield that focuses on developing algorithms capable of enabling machines to infer patterns and make autonomous decisions based on data (Dwivedi et al, 2019).

Through Edge AI, machine learning models and inference mechanisms are deployed directly on edge devices, allowing data analysis, inference, and decision-making to occur locally rather than in centralized cloud or private data center infrastructures. This decentralized processing model enhances performance for latency-sensitive and bandwidth-constrained applications while fostering greater autonomy and scalability across distributed intelligent systems (Laroui,2021).

II. Background and Evolution

Recent advancements in hardware have significantly optimized Edge AI deployment, with modern Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) engineered for superior performance, lower energy consumption and space-efficient form factors. High-performance GPUs, such as those found in NVIDIA Jetson and RTX series, enable robust real-time inference for tasks like video analytics and predictive maintenance, while maintaining energy efficiency suitable for edge environments where power and thermal budgets are strict. To manage these constraints, model quantization techniques (using 8-bit or 16-bit precision) and dynamic performance scaling effectively reduce resource demands, enhance battery life, and improve inference speeds, achieving latency reductions from 50–200 ms (cloud) to 1–10 ms (edge) (Khan et al., 2019). TPUs, particularly Google's Edge TPU used in Coral devices, deliver high-speed neural network inference with minimal power draw, making them ideal for embedded AI applications. Edge TPUs are designed as application-specific integrated circuits for accelerating neural network tasks, especially tensor operations, with significant performance-per-watt advantages and tight integration into platforms for easy deployment in smart cities, healthcare, and retail. TPUs are optimized for TensorFlow-based workloads and excel at large-scale inference for deep learning models, while retaining efficient operation in power-constrained edge scenarios (Carvalho, 2021).

Choosing between GPU and TPU hardware for Edge AI involves evaluating the application's requirements for real-time inference, energy consumption, and integration complexity. GPUs offer broad compatibility with many AI frameworks and excel in flexible, high-performance computing. In contrast, TPUs are specialized for efficient, high-throughput deep learning model execution and are particularly advantageous in embedded environments requiring low power and rapid inference. Developers often optimize AI models with frameworks like TensorFlow Lite, ONNX Runtime, and leverage architectural innovations such as MobileNet or EfficientNet to ensure optimal performance on edge hardware (Iftikhar, 2022).

III. CORE FEATURES OF EDGE AI

Edge AI exhibits a distinct set of technological and functional features that make it suitable for applications requiring low latency, autonomy, and localized intelligence. These features are critical to its architecture and use in modern research and enterprise systems.

- **3.1 Low Latency and Real-Time Processing**: Edge AI processes data directly on local devices, enabling instant analysis and decision-making without waiting for cloud responses. This is essential for high-speed applications like autonomous driving, robotics, or medical monitoring.
- **3.2 Reduced Bandwidth Consumption**: Only essential or summarized data is sent to the cloud, significantly lowering network traffic. This feature is vital in environments with limited or expensive bandwidth and reduces dependency on cloud infrastructure.
- **3.3 Enhanced Data Privacy and Security**: Since most processing occurs locally, sensitive data such as biometric and location information remains on the device, minimizing exposure to interception or cyberattacks during transmission.
- **3.4 Offline Functionality and Reliability**: Edge AI devices continue to operate effectively even without internet connectivity, which is valuable in remote or industrial sites where cloud access is unreliable. This autonomy increases operational continuity and resilience.
- **3.5 Energy Efficiency**: Modern edge processors and AI accelerators are optimized for low-power inference, reducing energy consumption compared to continuous cloud interaction. This feature extends device battery life and supports sustainable computing practices.
- **3.6 Context Awareness and Personalization:** Localized processing allows Edge AI systems to adapt dynamically to user behavior, local conditions, or environmental changes—enabling context-responsive intelligence, such as tailored notifications or motion-based automation.
- **3.7 Scalability and Integration Flexibility:** Edge AI frameworks support seamless scaling across diverse devices and applications—from IoT sensors to industrial robots—using modular and interoperable AI interfaces, fostering integration with smart city, healthcare, and manufacturing ecosystems.
- **3.8 Continuous Learning and Model Optimization:** Many edge AI systems incorporate incremental learning at the device level, allowing models to improve autonomously over time through fine-tuning or federated learning without compromising privacy.

3.9 Cost Efficiency: Processing data locally reduces operational expenses associated with cloud computation and data transfer. Organizations benefit from lower infrastructure costs while achieving faster insights and control (Yu, 2017).

IV. ARCHITECTURE OF EDGE AI

The architecture of Edge AI represents the structural and functional design that enables artificial intelligence to operate close to the data source rather than relying solely on centralized cloud resources. It integrates hardware, software, and communication technologies to support distributed intelligence, real-time decision-making, and efficient resource use. A standard Edge AI system comprises three primary layers that interact hierarchically to execute AI tasks efficiently

4.1 Device Layer (Edge Devices or Endpoints): Edge devices are computing entities located at the periphery of a network, responsible for the acquisition, generation, and initial processing of data closer to its source. These devices act as intelligent nodes that interact directly with the physical environment and contribute to distributed data management across the network. Common examples include industrial sensors, surveillance cameras, autonomous vehicles, and smartphones, all of which play critical roles in capturing and analyzing data in real time (Calo, 2017).

- **Computational Power:** Edge devices are equipped with embedded processors or AI accelerators capable of executing machine learning algorithms locally. This local execution ensures reduced latency, lower bandwidth consumption, and enhanced responsiveness for time-sensitive applications.
- **Storage:** These devices incorporate sufficient local memory and persistent storage mechanisms to support caching, temporary data retention, and model parameter storage. Adequate local storage capacity enables efficient data handling without constant reliance on centralized systems or cloud-based repositories.
- **Connectivity:** Seamless connectivity is achieved through Wide Area Network (WAN) technologies such as Wi-Fi, Ethernet, Bluetooth, and 4G/5G cellular links. This network layer facilitates synchronization with other edge nodes, gateways, or cloud platforms, ensuring real-time data exchange and system interoperability (Banoth, 2025).

4.2 Edge Layer (Edge Gateway / Edge Server): Edge gateways function as intermediary nodes that bridge communication between edge devices and centralized data centers or cloud platforms. Positioned strategically within the network, they aggregate, filter, and process data streams from multiple devices while enforcing security and compatibility protocols. By managing interoperability and ensuring minimal latency, edge gateways play a pivotal role in maintaining an efficient and secure edge-to-cloud operation. Core functions of Edge Gateways are:

- **Data Aggregation:** Gateways integrate data from numerous edge devices, executing initial pre-processing and normalization operations. This aggregation minimizes redundancy and ensures that only relevant information is transmitted to upper-level systems for advanced analytics.
- **Protocol Translation:** Edge gateways support heterogeneous communication protocols, translating and harmonizing data exchanges among devices that utilize different standards, thereby promoting seamless interoperability across the ecosystem.
- **Security Management:** Robust encryption, authentication, and data integrity controls are implemented to safeguard data during transmission. These measures prevent unauthorized access, tampering, or data leakage within the distributed edge network (Chen, 2019).

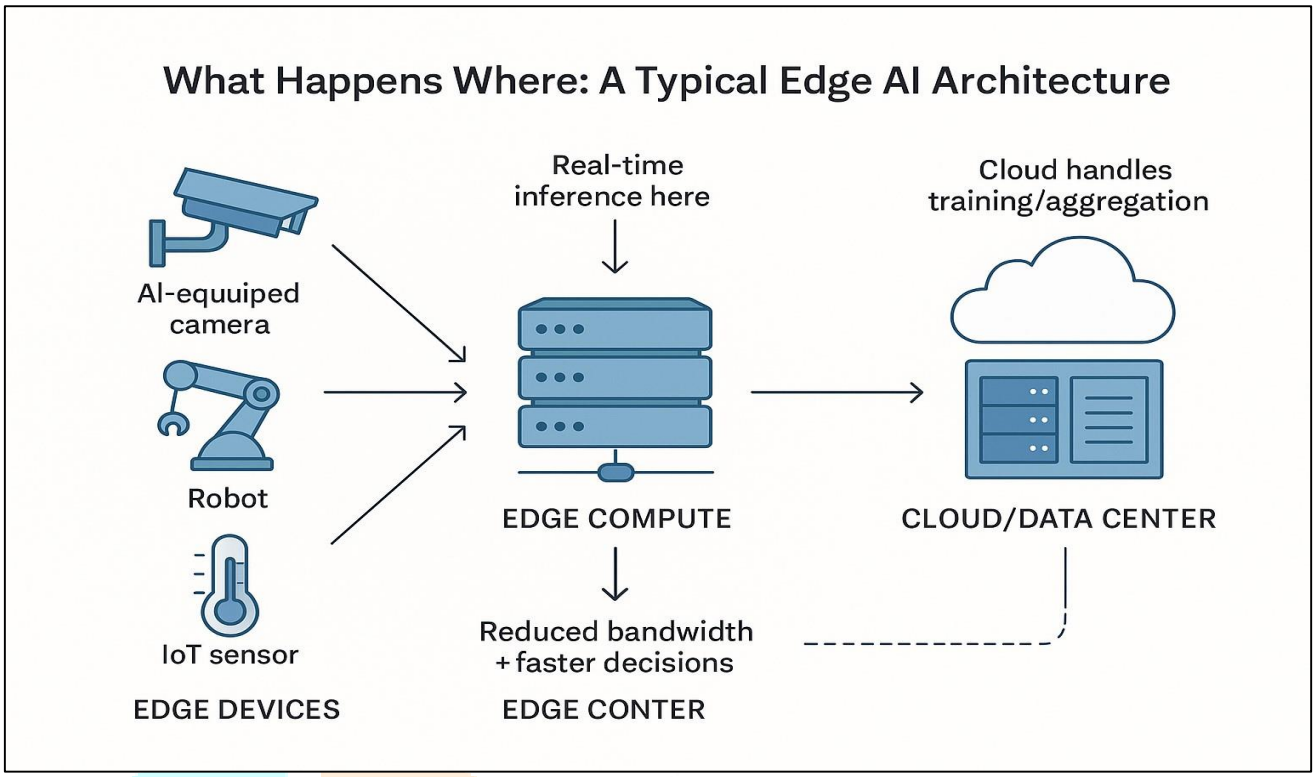

Figure: 4.1 Edge AI Architecture

4.3 Cloud Layer: The Cloud Layer in Edge AI architecture acts as the central control and intelligence hub. While edge devices handle real-time tasks locally, the cloud manages data storage, training, model updates, and coordination across all connected edge systems. The cloud layer provides the computational power, storage capacity, and advanced analytics that are difficult to perform on edge devices due to limited resources. It connects multiple edge nodes and ensures that all AI systems remain updated and optimized. This layer Collects and stores large volumes of data sent from various edge devices and provides centralized storage for long-term use and future analysis. The cloud performs AI model training using massive datasets that edge devices cannot handle. It uses high-performance GPUs, TPUs, or clusters to train deep learning models. Once trained, these models are compressed and sent to edge devices for local inference. This layer also performs big data analytics to discover global trends or patterns and integrates AI results from multiple sources to provide predictive insights. Cloud storage takes backups of edge data and models to prevent data loss and ensures disaster recovery and business continuity if edge devices fail (Singh & Gill, 2023).

4.4 Processing Steps: The processing steps of Edge AI describe the sequential operations through which data is collected, analyzed, and acted upon locally near its source, rather than relying solely on remote cloud servers. This workflow ensures low latency, efficient computation, and improved privacy while maintaining real-time intelligence. The following steps explain the Edge AI processing pipeline.

4.4.1 Data Generation and Collection: The first step involves gathering raw data directly from edge devices such as sensors, cameras, IoT nodes, or mobile devices. These devices continuously monitor their environment, capturing various inputs like temperature, motion, sound, or video streams. The collected data serves as the foundation for subsequent analysis and decision-making.

4.4.2 Local Data Pre-processing: Before applying AI models, the collected data undergoes pre-processing to clean, filter, and compress it. This step removes noise and irrelevant information, ensuring only significant and high-quality data is analyzed. It enhances accuracy while conserving bandwidth and computational resources at the device level.

4.4.3 Model Inference and Analysis: In this step, pre-trained AI or machine learning models, previously developed and optimized in the cloud—are deployed to edge hardware for inference. The models perform on-device analytics such as object recognition, anomaly detection, or speech analysis. Edge processors, equipped with AI accelerators, GPUs, or TPUs, execute these models within milliseconds, ensuring fast and context-aware results.

4.4.4 Real-Time Decision and Action: Based on the AI inference results, the system performs immediate actions at the device level without needing cloud intervention. These actions include triggering alerts, adjusting machinery operations, changing navigation paths, or activating cameras. This step is critical for safety- or time-sensitive applications such as autonomous driving, predictive maintenance, and healthcare monitoring.

4.4.5 Feedback and Learning: After local actions are executed, performance feedback is often collected to refine future responses. Many Edge AI systems use federated learning to update models collaboratively: devices share model parameters with a central server instead of raw data. This enhances global model accuracy while protecting user privacy.

4.4.6 Cloud Synchronization and Model Updating (Optional): In some cases, selective or summarized data is transmitted to the cloud for deeper analytics, centralized reporting, or retraining of AI models. The cloud periodically sends updated or optimized versions of the model back to edge devices, closing the learning loop and ensuring continuous improvement of local intelligence (Gill et al, 2022).

V. TAXONOMY OF EDGE AI

To offer a holistic approach to the analysis, comparison, and design decisions, a multidimensional taxonomy of edge intelligence is a classification of Edge AI technologies and systems based on several important dimensions. The fundamental dimensions normally involve the location of deployment, processing capacities, area of use, and hardware structure. The key taxonomy dimensions are:

- **Deployment Location**: Edge intelligence may be deployed onto devices (on-device), locally on edge nodes, fog nodes or regional data centers, which affects latency, privacy, and scalability.
- **Processing Capabilities**: This encompasses TinyML which is ultra-lightweight ML, federated learning which provides privacy-safe distributed training, conventional or deep learning approaches such as DRL. Processing methods are chosen depending on resource limitations and application needs.
- **Application Domain**: Edge AI is being utilised in a wide range of industries such as healthcare, transportation, smart cities, industrial IoT, and entertainment/metaverse systems. Both domains define requirements in terms of latency, privacy and regulatory compliance.
- **Hardware**: Architecture: Edge systems take advantage of CPUs, GPUs, FPGAs, neuromorphic chips, and custom accelerators. It depends on the computational requirements, energy and real-time considerations as a basis of selection.
- **Resource Management**: Provisioning, workload distribution, application placement, and resource allocation strategies are all included in this (resource management).
- **Security**: Frameworks should be capable of considering platform security, host security and network security to achieve data protection and dependability.
- **Model Management**: Model sizing (full vs. reduced), lifelong learning, and joint model updates without providing raw data (e.g. federated learning).
- **Scalability and Migration**: Discusses container scaling, task scheduling, and migration policies on cloud, fog, and edge[2].
- **Heterogeneity**: Solves the heterogeneity of hardware, platforms, and available resources at the edge and provides them to the edge [2].
- **Operational Issues**: These are energy efficiency, connection management, and power consumption limitations. (Zhou et al., 2019).

VI. APPLICATION DOMAINS OF EDGE AI

Edge AI powers a wide variety of real-world applications across many sectors, offering faster, more private, and more efficient data processing by performing AI analysis locally on edge devices. Edge AI applications span across multiple industries, bringing artificial intelligence directly to devices where data is generated. This localized intelligence supports faster computation, improved security, and real-time decision-making without relying on cloud infrastructures. The following sections provide a detailed exploration of its applications.

6.1 Manufacturing and Industrial Automation

- In manufacturing, Edge AI is central to Industry 4.0, driving predictive maintenance, quality inspection, and smart automation (Foukalas & Tziouvaras, 2021).

  - **Predictive Maintenance:** Sensors embedded in machinery continuously monitor parameters like vibration, temperature, and pressure. Edge AI detects anomalies and predicts failures before breakdowns occur, preventing costly downtime.
  - **Quality Inspection:** AI algorithms at the edge analyze visual and sensor data in real-time to identify defects, ensuring high product quality and minimizing waste.
  - **Process Optimization:** Edge systems dynamically adjust operational parameters to maintain peak efficiency, thereby enhancing throughput and reducing resource consumption (Golec et al., 2020).

6.2 Healthcare and Medical Devices

Edge AI transforms healthcare through real-time diagnostics and remote patient monitoring.

- **Remote Monitoring:** Wearables equipped with AI algorithms analyze vital signs locally, alerting physicians or patients about irregularities without relying on cloud connectivity.
- **Medical Imaging:** Edge-based analysis of X-rays or CT scans enables near-instant diagnostic insights, especially in remote or emergency settings.
- **Smart Hospitals:** Sensors powered by Edge AI streamline patient tracking, optimize energy use, and automate logistics like medicine delivery (Tuli et al., 2019).

6.3 Transportation and Autonomous Vehicles

The transportation sector relies heavily on Edge AI for safe, efficient, and autonomous mobility (Ke at al, 2020).

- **Autonomous Vehicles:** Edge AI processes sensor data—such as camera images, LiDAR, and radar inputs—to make split-second driving decisions.
- **Traffic Management:** Edge-enabled infrastructure analyzes real-time traffic patterns to optimize signal control and reduce congestion.
- **Fleet Optimization:** In logistics, Edge AI routes vehicles dynamically, minimizing delays and saving fuel (Shankar, 2024).

6.4 Agriculture and Smart Farming

Edge AI enhances precision agriculture, making farming operations more sustainable and efficient.

- **Crop Monitoring:** Drones and edge-enabled sensors collect and analyze soil moisture, weather, and crop health data in real time to provide actionable insights.
- **Predictive Irrigation and Fertilization:** AI automates water and nutrient distribution to optimize yield and minimize waste.
- **Pest Detection:** Local AI models detect pests early and trigger timely interventions, reducing crop loss and pesticide dependency (Liu et al., 2021).

6.5 Energy and Smart Grids

Edge AI optimizes **energy management** and sustainability through decentralized intelligence.

- **Smart Grids:** AI algorithms at substations analyze power demand and stabilize grids with renewable inputs like solar or wind energy.
- **Predictive Maintenance for Utilities:** Edge systems forecast equipment malfunction in turbines or power lines to prevent outages.

- **Energy Optimization:** Buildings employ AI-driven edge devices to regulate HVAC systems and minimize energy waste (Minh et al., 2022).

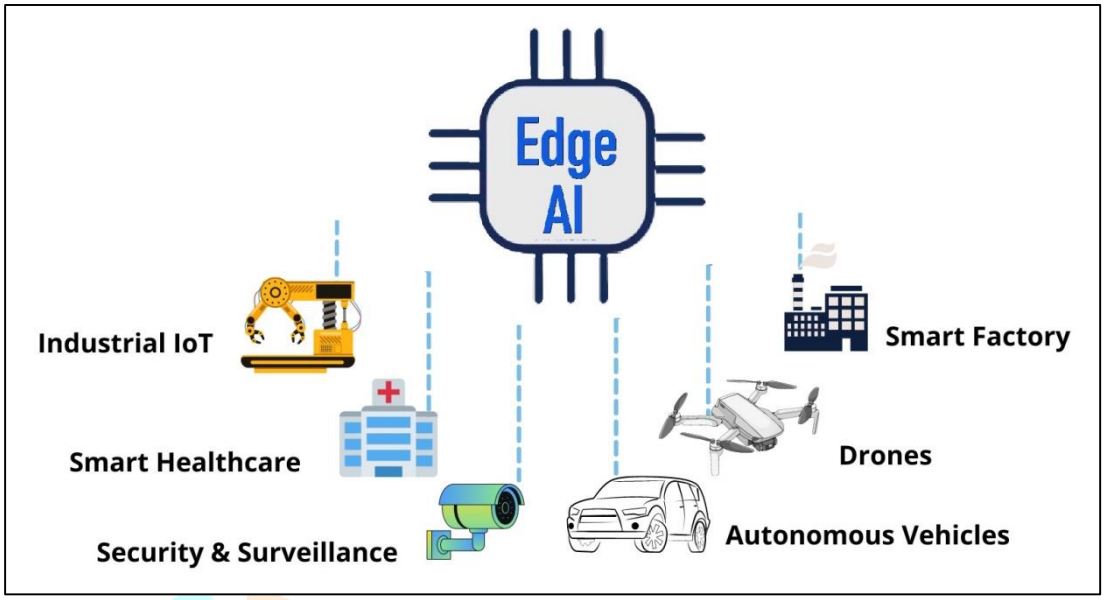

Figure 6.1 Application Domains of Edge AI

6.6 Retail and Smart Commerce

In retail, Edge AI enhances customer experience, inventory accuracy, and store operations.

- **Real-Time Analytics:** Cameras analyze shopper behavior for layout optimization and personalized promotions.
- **Inventory Management:** Smart shelves track stock levels and alert staff to replenish items.
- **Self-Checkout Systems:** Edge vision systems enable automated billing and theft detection without cloud dependency (Garcia et al., 2023).

6.7 Smart Cities and Public Infrastructure

Edge AI improves urban efficiency and safety through localized sensing and processing.

- **Public Safety:** Surveillance cameras equipped with AI detect suspicious activities or emergencies in real time.
- **Environmental Monitoring:** Sensors analyze pollution, noise, and waste collection patterns to improve city management.
- **Traffic and Lighting Optimization:** Intelligent edge nodes control traffic lights and street illumination based on activity data (Hu & Li, 2020).
- Smart Homes: AI-powered devices like voice assistants, security cameras, and thermostats optimize comfort and safety while preserving privacy (Khan et al., 2020).

VII. CHALLENGES AND OPEN ISSUES

The field of edge AI has significant challenges and open problems within the hardware, software, security and deployment, which restrict its broader application and reliability. Some of them are listed below:

- **Resource Constraints**: Edge devices may not be powerful in terms of CPU, memory, and storage as cloud servers are. Training complicated AI models needs approaches such as model pruning and quantization, and can decrease accuracy and functionality, particularly in large generative AI models.

- **Power**: Edge AI devices are typically battery-only devices that are unable to handle high-performance chips. One of the fundamental design issues is balancing output with minimal energy budget, heat dissipation and the affordability of sophisticated silicon.
- **Data and Model Management**: Data sampled on the edge can be fragmented or unresolved and huge-scale observations are hard to achieve. The process of updating models of thousands of distributed devices needs to operate well over-the-air, and may be vulnerable to bugs or version compatibility issues.
- **Security Vulnerabilities**: The edge devices, particularly IoT endpoints, are not secure by design and usually do not have strong cyber security that can withstand attacks and data breach. These are aggravated by the lack of a reliable automatic updates.
- **Data Privacy**: Although edge AI enhances data privacy by storing sensitive information in a local manner, it is however complicated to achieve privacy-preserving AI models, as well as comply with the regulations.
- **Scalability and Maintenance**: It is far more complicated to handle, monitor and update large quantities of edge devices than in centralized cloud systems. This adds operational costs and risks.
- **Model Robustness**: Models should be resistant to changes in the real world- alterations in light, noise, and device conditions can result in AI becoming worse in performance than expected.
- **High Upfront Cost**: Edge AI solutions can typically demand specialized hardware, staff training, and bespoke deployment tools that demand significant initial investment in contrast to cloud solutions.
- **Aggregated Analytics**: It is challenging to produce general insights of distributed devices due to the inconsistencies of datasets and the absence of centralized aggregation of the data due to dispersed devices (Letaief et al., 2021).

VIII. FUTURE RESEARCH DIRECTIONS

Future directions on edge AI are to address the existing weaknesses and enhance functionality to facilitate more efficient, robust, and scalable edge AI solutions. Such are hardware, software, connectivity, security, and AI model development.

- **Hardware Efficiency and Specialization**: Studies into low-power, high-performance AI accelerators and specialised chips (ASICs, TPUs) to support complex AI inferencing with few energy and computational resources on edge devices.
- **Lightweight and Adaptive AI Models**: Creation of lightweight, compressed and dynamically adaptable AI models, which are both accurate and able to fit into resource limited settings of edge devices.
- **Federated and On-Device Learning**: Federated learning advances to provide continuous learning that is privacy-preserving on the edge that requires minimal data transfer to centralized servers. On-device learning is required to support continuous, privacy-preserving model training and updates on the edge, which minimizes data transfer to centralized servers.
- **Edge AI Infrastructure and Orchestration**: Developing Java-rounded edge computing environments enabling an easy deployment, operation, and coordination of AI workloads distributed across different edge gadgets and hybrid cloud-edge settings
- **Security and Privacy**: Improving edge AI security with AI-based threat detection, zero trust, blockchain, and encrypted computations that guarantee data integrity and compliance at the edge
- **5G and Beyond Connectivity**: Utilizing 5G and 6G along with new wireless technology to offer ultra-low latency, high bandwidth, and reliable connectivity over real-time edge AI applications.
- **Real-World Adaptability and Robustness**: How to make AI more robust to new environmental factors, sensor noise, and hardware variation when deployed to edges to guarantee robust operation, in practice.
- **Expansion of Cross-Industry Use Cases**: Edge AI and other research should be extended to new domains of application such as precision agriculture, autonomous systems, smart cities, and healthcare diagnostics with an emphasis on domain-specific customization and scalability.

The directions are to make edge AI more energy-efficient, secure, flexible, scalable, and able to provide real-time intelligence at the data generation point in industries, as well as developing energy-efficient, high-performance special hardware and lightweight and adaptable AI models that are able to work with resource-

constrained edge devices. To allow privacy-sensitive updates to models without heavy reliance on the clouds, research is also steered towards federated and on-device learning strategies. It is important to design strong edge AI infrastructure and orchestration platforms that enable smooth deployment and scaling. The security is improved through AI-based threat detection, zero-trust, and blockchain integration to secure distributed edge devices. The development of 5G and other connectivity will facilitate real-time applications in ultra-low latency and high-bandwidth. Lastly, robustness to environmental variability and domain-specific edge AI applications in sectors like healthcare, agriculture, and smart cities are the main open problems that will lead to further innovation in the future.

IX. CONCLUSION

Edge Intelligence is rapidly transforming digital infrastructure by localizing AI-driven analytics and decision-making. The interplay of evolving hardware, innovative computing paradigms, and new software frameworks is driving applications once considered the exclusive domain of centralized clouds. Sustained research is necessary to realize the full promise of distributed, adaptive, and secure edge AI systems. Edge AI is a revolutionary development that allows real-time processing of data, immediate decision-making, and the ability to provide greater privacy without the extensive use of the centralized cloud service through the integration of artificial intelligence directly into edge devices. The paradigm shift resolves key issues of latency, bandwidth, and data security and enables industries, such as healthcare and manufacturing, and autonomous vehicles and smart cities to use Edge AI. However, the current problem of resource limits, security risks, and scalability issues limits the further research and innovation of Edge AI despite its potential as a transformative technology. High efficiency in hardware and adaptive AI models, federated learning, edge orchestration and strong security frameworks in the future will prove critical to realizing its full potential. Finally, Edge AI is set to transform the nature of intelligence deployment and implementation into wide range of real world uses, making technologies more intelligent, fast, and sustainable. The study reveals the prospects and difficulties in the future, informing the further explorations and progress in the fast changing environment of Edge AI.

References

1. Adeoye, S. (2025). Internet of Things (IoT): a vision, architectural elements and future directions. *Cognizance Journal of Multidisciplinary Studies*, *5*(1), 316–338. https://doi.org/10.47760/cognizance.2025.v05i01.027
2. Banoth, S., M, V., Punna, H. S., P, M., Prakash, V., & M, J. (2025). Edge computing architectures for Low-Latency data processing in Internet of Things applications. *ITM Web of Conferences*, *76*, 03003. https://doi.org/10.1051/itmconf/20257603003
3. Calo, S. B., Touna, M., Verma, D. C., & Cullen, A. (2017). Edge computing architecture for applying AI to IoT. *2021 IEEE International Conference on Big Data (Big Data).* https://doi.org/10.1109/bigdata.2017.8258272
4. Carvalho, G., Cabral, B., Pereira, V., & Bernardino, J. (2021). Edge computing: current trends, research challenges and future directions. *Computing*, *103*(5), 993–1023. https://doi.org/10.1007/s00607-020-00896-5

5. Chen, S., Wen, H., Wu, J., Lei, W., Hou, W., Liu, W., Xu, A., & Jiang, Y. (2019). Internet of things based smart grids supported by intelligent edge computing. *IEEE Access*, *7*, 74089–74102. https://doi.org/10.1109/access.2019.2920488
6. Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., . . . Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
7. Foukalas, F., & Tziouvaras, A. (2021a). Edge Artificial intelligence for Industrial Internet of Things applications: An Industrial Edge Intelligence solution. *IEEE Industrial Electronics Magazine*, *15*(2), 28–36. https://doi.org/10.1109/mie.2020.3026837
8. Garcia, A., De Barreana, T. F., Chacón, J. L. F., Oregui, X., & Etxegoin, Z. (2023). Edge Architecture for the Integration of Soft Models Based Industrial AI Control into Industry 4.0 Cyber-Physical Systems. In *Lecture notes in networks and systems* (pp. 67–76). https://doi.org/10.1007/978-3-031-42536-3_7
9. Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., Golec, M., Stankovski, V., Wu, H., Abraham, A., Singh, M., Mehta, H., Ghosh, S. K., Baker, T., Parlikad, A. K., Lutfiyya, H., Kanhere, S. S., Sakellariou, R., Dustdar, S., . . . Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things*, *19*, 100514. https://doi.org/10.1016/j.iot.2022.100514
10. Golec, M., Gill, S. S., Bahsoon, R., & Rana, O. (2020). BIOSEC: a biometric authentication framework for secure and private communication among edge devices in IoT and industry 4.0. *IEEE Consumer Electronics Magazine*, *11*(2), 51–56. https://doi.org/10.1109/mce.2020.3038040
11. Hu, B., & Li, J. (2020). An edge computing framework for powertrain control system optimization of intelligent and connected vehicles based on Curiosity-Driven Deep Reinforcement Learning. *IEEE Transactions on Industrial Electronics*, *68*(8), 7652–7661. https://doi.org/10.1109/tie.2020.3007100

12. Iftikhar, S., Gill, S. S., Song, C., Xu, M., Aslanpour, M. S., Toosi, A. N., Du, J., Wu, H., Ghosh, S., Chowdhury, D., Golec, M., Kumar, M., Abdelmoniem, A. M., Cuadrado, F., Varghese, B., Rana, O., Dustdar, S., & Uhlig, S. (2022a). AI-based fog and edge computing: A systematic review, taxonomy and future directions. *Internet of Things*, *21*, 100674. https://doi.org/10.1016/j.iot.2022.100674
13. Ke, R., Zhuang, Y., Pu, Z., & Wang, Y. (2020). A smart, efficient, and reliable parking surveillance system with edge artificial intelligence on IoT devices. *IEEE Transactions on Intelligent Transportation Systems*, *22*(8), 4962–4974. https://doi.org/10.1109/tits.2020.2984197
14. Khan, L. U., Yaqoob, I., Tran, N. H., Kazmi, S. M. A., Dang, T. N., & Hong, C. S. (2020). Edge-Computing-Enabled Smart Cities: A comprehensive survey. *IEEE Internet of Things Journal*, *7*(10), 10200–10232. https://doi.org/10.1109/jiot.2020.2987070
15. Khan, W. Z., Ahmed, E., Hakak, S., Yaqoob, I., & Ahmed, A. (2019a). Edge computing: A survey. *Future Generation Computer Systems*, *97*, 219–235. https://doi.org/10.1016/j.future.2019.02.050
16. Laroui, M., Nour, B., Moungla, H., Cherif, M. A., Afifi, H., & Guizani, M. (2021). Edge and fog computing for IoT: A survey on current research activities & future directions. *Computer Communications*, *180*, 210–231. https://doi.org/10.1016/j.comcom.2021.09.003
17. Letaief, K. B., Shi, Y., Lu, J., & Lu, J. (2021). Edge Artificial Intelligence for 6G: vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, *40*(1), 5–36. https://doi.org/10.1109/jsac.2021.3126076
18. Liu, J., Xiang, J., Jin, Y., Liu, R., Yan, J., & Wang, L. (2021). Boost Precision Agriculture with Unmanned Aerial Vehicle Remote Sensing and Edge Intelligence: A Survey. *Remote Sensing*, *13*(21), 4387. https://doi.org/10.3390/rs13214387
19. Minh, Q. N., Nguyen, V., Quy, V. K., Ngoc, L. A., Chehri, A., & Jeon, G. (2022). Edge computing for IoT-Enabled smart Grid: The future of energy. *Energies*, *15*(17), 6140. https://doi.org/10.3390/en15176140
20. Shankar, V. (2024). Edge AI: A Comprehensive Survey of Technologies, Applications, and Challenges. *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, 1–6. https://doi.org/10.1109/acet61898.2024.10730112
21. Singh, R., & Gill, S. S. (2023). Edge AI: A survey. *Internet of Things and Cyber-Physical Systems*, *3*, 71–92. https://doi.org/10.1016/j.iotcps.2023.02.004

22. Tuli, S., Tuli, S., Wander, G., Wander, P., Gill, S. S., Dustdar, S., Sakellariou, R., & Rana, O. (2019). Next generation technologies for smart healthcare: challenges, vision, model, trends and future directions. *Internet Technology Letters*, *3*(2). https://doi.org/10.1002/itl2.145

23. Yu, W., Liang, F., He, X., Hatcher, W. G., Lu, C., Lin, J., & Yang, X. (2017). A survey on the edge computing for the internet of things. *IEEE Access*, *6*, 6900–6919. https://doi.org/10.1109/access.2017.2778504

24. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge Intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, *107*(8), 1738–1762. https://doi.org/10.1109/jproc.2019.2918951