

EDGE AI: Leveraging Artificial Intelligence in Edge Computing

Ahmad Umar Labdo^{1*}, Bachcha Lal Pal², Asifa Faruk Shah²,
Fatima Ibrahim Madori³, Zainab Yakubu Azare⁴,
Abulfatahi Baita Ubawaru³

¹Department of Computer Applications, Mewar University Gangrar, Chittorgarh, Rajasthan India

²Assistant Professor, Department of Computer Applications, Mewar University Gangrar, Chittorgarh,
Rajasthan India

³Department of Computer Science, Northwest University, Kano, Nigeria

⁴Department of Software Engineering, Bayero University, Kano, Nigeria

Abstract: Edge Artificial Intelligence (Edge AI) combines edge computing and AI to process data in real time while lowering latency, improving privacy, and making use of available resources. In sectors including healthcare, smart manufacturing, and the Internet of Things (IoT), Edge AI improves operational efficiency by processing data closer to the source. Large Language Models (LLMs) have been optimized for edge devices, inference performance has been increased, and issues with model size, energy consumption, and unstable network settings have been addressed. An overview of Edge AI's current status is given in this paper, along with a discussion of its revolutionary effects and potential future research areas. Important advancements include frameworks for optimizing AI models according to edge computing constraints and architectural considerations for combining generative AI (GenAI) with edge-cloud computing. In vital applications like the Internet of Vehicles, where real-time processing is essential for efficiency and safety, Edge AI has shown notable advancements. Edge AI is anticipated to be crucial in developing intelligent systems and guaranteeing quick, safe, and effective decision-making as AI-enabled edge devices develop further. Future developments in intelligent and adaptable computing ecosystems will be shaped by the further convergence of edge computing and artificial intelligence.

Keywords: data, edge-cloud, GenAI, latency, optimization and real-time.

1. Introduction

According to numerous research papers, integrating artificial intelligence (AI) at the edge brings data processing closer to the data source, enabling real-time analytics and decision-making. This method greatly lowers latency, maximizes bandwidth utilization, improves privacy and security, and encourages responsive and customized actions in various industries, including smart manufacturing, healthcare, and IoT domains [1]. Using cutting-edge machine learning algorithms, Edge AI can extract actionable insights from large datasets, spurring innovation, improving operational efficiency, and changing conventional data platforms into flexible, adaptive ecosystems supporting real-time decision-making. Organizations can achieve effective implementation, enhanced privacy, security and latency by deploying AI capabilities at the edge. This will ultimately transform the future of edge computing and propel advances in intelligent decision-making that are seamlessly integrated into day-to-day operations.

Reduced latency, improved privacy and effective data processing are some of the advantages of Edge AI, a revolutionary solution that combines Artificial Intelligence (AI) with Internet of Things (IoT) devices [2]. Research has concentrated on optimizing Large Language Models (LLMs) for edge devices to improve inference throughput and resource efficiency, addressing issues like model size and self-attention mechanisms. Furthermore, collaborative edge AI inference architectures have been proposed by controlling quantization errors, utilizing AirComp techniques, and suppressing noise to maximize accuracy, demonstrating better optimization algorithms than baselines. The development of solutions like HawkVision for low-latency serving of vision DNNs with reduced latency and power consumption, outperforming current systems, was prompted by the trend towards modelless ML inference at the edge, which also presents challenges like low power consumption and volatile network environments [3]. Edge AI, which emphasizes privacy preservation, low latency, and resource optimization, has emerged as a result of the marriage of Edge Computing and AI. Its primary focus is on effective Edge ML technologies for real-time performance and advanced learning paradigms.

2. Edge-Ai Concepts

Artificial Intelligence (AI) is the study of computer science to develop machines that can carry out operations that normally call for human intelligence. Artificial Intelligence (AI) seeks to mimic or duplicate cognitive processes in humans so that machines can function independently and wisely. Some tasks include learning, reasoning, problem-solving, perception, language comprehension, and decision-making [9].

One component of Artificial Intelligence is Machine Learning (ML), a subset of AI focused on developing algorithms that allow computers to learn from and make predictions or decisions based on data. There are three types of AI learning procedures: Supervised Learning, which is learning from labeled data to make predictions; Unsupervised Learning, which is identifying patterns in unlabeled data through recognition of features and patterns; and Reinforcement Learning, which is learning through trial and error to achieve specific goals while rewarding and penalizing good and wrong results respectively [5].

Another component of AI is Deep Learning, a subset of machine learning that uses neural networks with many layers (deep neural networks) to model complex patterns in data. It is used in natural language processing, image and speech recognition. Natural Language Processing (NLP) in itself is a subset of AI, it is a field of AI focused on the interaction between computers and humans through natural language [5]. NLP is applied in chatbots, language translation and sentiment analysis. Computer Vision is another component of AI that enables machines to interpret and understand visual information from the world. It is used in facial recognition, object detection and autonomous driving. There are Expert Systems which are AI programs that simulate the decision-making ability of a human expert, they are applied in medical diagnosis and financial forecasting.

Edge computing is the practice of situating consolidated data centers close to the user (as far as networks go) to minimize latency, offer scalability, and use less energy. Edge computing is a paradigm that is becoming increasingly important in the changing digital landscape. It is particularly important for real-time, low-latency applications and Internet of Things (IoT) environments. Notwithstanding its benefits, edge computing is severely limited in terms of storage capacity and is prone to reliability problems because of its resource constraints and exposure to harsh environments [10].

The paper [9] described Edge Computing (EC), as a new computing paradigm, which has garnered a lot of attention as a solution to cloud computing issues. The EC model's main concept, to put it simply, is to shift computing, storage, and data processing tasks that were initially needed by the cloud to the network's edge, close to terminal devices. As a result, there is less strain on network bandwidth, data transmission costs are decreased, device response times are shortened, and decentralization is accomplished.

EdgeAI is defined by [9] as an area in edge computing that deals with the application and use of artificial intelligence (AI) algorithms on devices that are physically closer to data sources (such as sensors, cameras, and Internet of Things devices) than on cloud data centers. Keeping data local improves privacy and security, lowers latency, conserves bandwidth, and permits real-time data processing and decision-making.

3. Edge – Artificial Intelligence

Artificial intelligence (AI) and edge computing are combining to create a revolutionary synergy that is changing many industries. Field - Programmable Gate Array (FPGA) technology combined with edge computing systems improves real-time applications such as pavement defect detection in Advanced Driver-Assistance Systems (ADASs) [8]. To address resource constraints and privacy concerns regarding data, collaborative edge training makes use of trusted edge devices for long-term big AI model training at the edge. At the nexus of learning and edge computing, edge intelligence refines deep learning models such as convolutional neural networks for deployment on resource-constrained edge systems, opening the door to highly accurate and efficient real-time applications [10]. By automating and accelerating processes with highly optimized AI models that have been trained on datasets, edge AI—the combination of edge computing and AI—is seen as a potentially revolutionary paradigm shift that will enable real-time intelligent applications. AI, cloud computing, and edge computing together present a plethora of opportunities across a range of industries, including mobile gaming, the Internet of Things, and extended reality. This highlights the enormous potential for innovation and progress in communication technology.

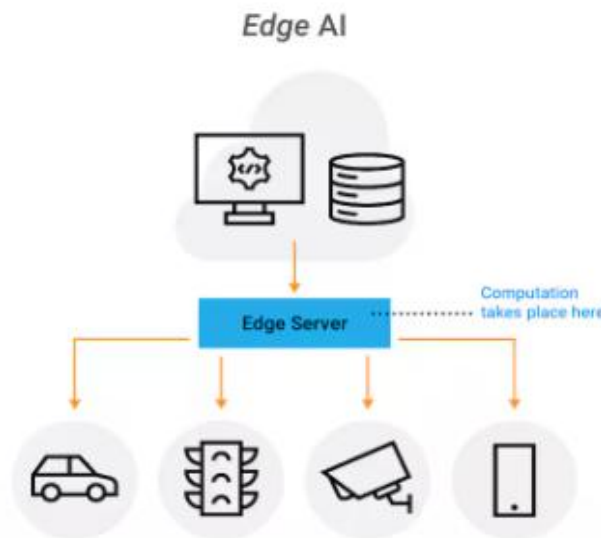


Figure 1: edgeAI architecture

Different parts and tiers are essential for effective communication and data processing in edge computing and Internet of Things systems. Sensors and readers, for example, gather data, which is then sent to edge servers near these devices to improve security and minimize communication overhead. To facilitate quicker deployment and better resource management, these edge servers communicate with cloud servers, which offer computational and storage services via the Internet. A variety of communication protocols, each with unique benefits and drawbacks based on the specifics of the application, have been developed to enable communication between these components [11]. A comprehensive approach to edge computing starts with integrating cloud, edge, and device architectures; this addresses problems like heterogeneity and inconsistent standards while promoting industry intelligence and development. Furthermore, it has been suggested that encrypted biometrics be used in edge computing architecture to create secure client-cloud server communication, protecting user privacy and thwarting security threats like replay and impersonation attacks [12].

3.1 Technical Foundations

The architecture of edge AI comprises devices, technologies and other components that make up the edge AI technology that is those components that enable the integration of AI into edge computing [13]. The architecture consists of three main components as follows:

These devices include: Sensors and Actuators, which collect real-time data from the environment (e.g., temperature sensors, cameras); Edge AI Hardware, which include processors (CPUs, GPUs, TPUs) and specialized AI chips designed for local data processing and inference; and Connectivity Modules, used to enable communication with edge servers and the cloud using protocols like Wi-Fi, Bluetooth, or cellular networks [11].

Edge servers are the components that handle data processing and storage. Local Data Processing performs computational tasks and analytics closer to the data source, reducing latency and bandwidth usage. Intermediate Storage is a temporary data storage for quick access and local data handling [6]. AI Model Deployment hosts and runs AI models for real-time inference, often optimized for low power consumption and high efficiency.

Centralized Data Storage stores large volumes of data aggregated from multiple edge devices for long-term analysis and historical insights. Advanced Analytics and AI Training provide the computational power and resources for complex data processing, machine learning model training, and large-scale analytics [14]. Management and Orchestration of centralized control and coordination of edge devices, including updates, security patches, and resource allocation.

3.2 Edge-AI Algorithms

These are algorithms that enable the integration of AI into edge computing, they include compression, lightweight and federated learning models [15].

Quantization reduces the precision of the numbers used in model calculations, significantly decreasing the model size and computational load without greatly affecting accuracy. Pruning eliminates unnecessary parameters (weights) from the model, reducing its complexity and enhancing inference speed on edge devices.

Knowledge Distillation involves training a smaller ‘student’ model to mimic the performance of a larger ‘teacher’ model, resulting in a more efficient model suitable for edge deployment [16].

MobileNets designed for mobile and edge applications, these models use depth-wise separable convolutions to reduce the number of parameters and computational cost. SqueezeNet is a model architecture that achieves AlexNet-level accuracy with significantly fewer parameters, making it ideal for resource-constrained environments.

Decentralized Training allows multiple edge devices to collaboratively train a model using their local data while keeping the data decentralized, enhancing privacy and reducing communication overhead [1]. Model Aggregation combines the locally trained models on a central server without sharing raw data, improving the overall model while preserving data privacy.

3.3 Hardware for Edge-AI

Several hardware devices are employed by edge computing to effectively integrate it with AI [17], these include:

- a. Central Processing Units (CPUs) provide versatile processing power for various tasks, including AI inference, but may not be as efficient as specialized hardware for intensive AI workloads.
- b. Graphics Processing Units (GPUs) excel at parallel processing, making them suitable for running complex neural network inferences. They are widely used in edge AI for tasks requiring high computational power.
- c. Tensor Processing Units (TPUs) are designed specifically for AI workloads. They offer high performance and efficiency for running machine learning models, particularly in inference tasks.
- d. Field-Programmable Gate Arrays (FPGAs) allow for hardware customization and optimization for specific AI tasks, providing a balance between performance and energy efficiency.
- e. Application-Specific Integrated Circuits (ASICs) are custom-built for specific applications, offering the highest performance and energy efficiency for dedicated AI tasks, such as Google's Edge TPU.

4. Reviewed Works

The notion of Edge AI emerged from the combination of Artificial Intelligence (AI) and the Internet of Things (IoT). This allows data processing to occur directly on IoT devices, improving responsiveness and efficiency. It focuses on important benefits like lower latency, bandwidth conservation, improved privacy, and quicker decision-making, especially in fields like urban planning and healthcare. This article discusses issues, such as IoT device resource limitations, and offers hybrid models to balance cloud and on-device processing. The study highlights Edge AI's transformative potential and provides examples of its use in smart city traffic management and real-time health monitoring. However, it also makes a case for more research to improve AI algorithms and increase device range [4].

A thorough summary of the incorporation of edge computing and AI in Internet of Things applications can be found in [2]. The objective of the review was to identify current research gaps and areas that warrant further investigation. With an emphasis on how these approaches can improve edge-based applications, the paper provides a qualitative comparison of various AI techniques and their functions at the network edge. Smart agriculture, smart environment, smart grid, smart healthcare, smart industry, smart education, smart transportation, and security and privacy are the eight application areas into which the authors divided their analysis. The article addresses some issues with AI and edge computing, including the need for effective resource orchestration and synchronization, resource limitations, and computational complexity. Topics like redundant, irrelevant, and missing data as a result of network outages are covered, highlighting the necessity of edge data management techniques. The use of AI techniques for data preprocessing, such as handling spatial and temporal redundancy was mentioned by the authors.

The paper [5] explores sustainable AI processing at the edge, focusing on machine learning acceleration while considering environmental impacts like embodied energy and carbon emissions. It compares various convolutional neural network (CNN) acceleration methods, including processing-in-memory (PIM), mobile GPUs, and FPGAs, highlighting trade-offs in energy consumption. Results indicate that Racetrack memory PIM can recover its embodied energy in about one year, while mobile GPUs, despite being more sustainable under high activity, have higher embodied energy. The study emphasizes the importance of evaluating both embodied and operational energy to assess the sustainability of edge computing systems.

In a different study [1], the task execution problem in edge computing is formulated as a multi-class classification problem, and task offloading is optimized through the use of an AI-powered fast task decision method. Although specific datasets are not detailed in the contexts provided, the study uses extensive simulations to assess the proposed framework. When compared to traditional algorithms, the suggested method

shows appreciable gains in decision-making speed and accuracy for offloading tasks, guaranteeing a superior user experience.

[6]studied the integration of generative artificial intelligence (GenAI) with edge-cloud computing. According to the study, GenAI is a subset of artificial intelligence that produces content that resembles that produced by humans. It highlights how quickly GenAI systems are developing and the difficulties that follow in handling the enormous volumes of data produced on the Internet. It talks about how GenAI services are dependent on conventional cloud computing, which causes a lot of latency because of data transmission and a high volume of queries. Applications using GenAI may perform worse as a result. According to scientists, edge-cloud computing can help by lowering latency and offering sufficient processing power. Collaboration between cloud resources and edge devices is emphasized as a workable approach to efficiently scale GenAI systems. The study demonstrates the technological difficulties that arise when scaling solutions through edge-cloud collaborative systems using two example GenAI applications. The need for creative methods to maximize performance is emphasized in this section.

A study describes a thorough method for incorporating edge AI into 6G networks in a research paper. The authors start by outlining their vision for reliable and scalable edge AI systems that combine decentralized machine learning models with wireless communication technologies. The application of distributed optimization methods to edge AI model training is highlighted in this research. This entails reducing a global loss function, which is essential for attaining scalability and reliability in edge AI systems, based on decentralized data produced by several intelligent devices [7]. The authors provide new, service-driven design concepts for resource allocation, with an emphasis on multi-objective optimization techniques and network slicing optimization. The goal of this methodology is to improve edge AI systems' resource management efficiency. We present a comprehensive end-to-end (E2E) architecture that includes edge AI management, data governance, and network infrastructure, among other elements. This architecture's low latency and energy efficiency are guaranteed to accommodate a wide range of applications.

A thorough analysis of the approaches and successes associated with AI model optimization for edge computing was provided by [8]. The study carried out an extensive survey to investigate different optimization strategies that might be used with edge AI models. This required going over the body of research and classifying various optimization techniques according to how well they worked and how well they fit into edge situations. Model compression, regularization strategies, and hyper-parameter optimization are just a few of the categories the study uses to group optimization techniques. The authors include many tools for adjusting hyper-parameters, including Bayesian optimization techniques like Spearmint and Scikit-learn for grid and random search. These resources are necessary to quickly and effectively identify the best solutions. The research highlights the significance of taking into account edge device limits, such as processing power, memory, and network bandwidth, even though it does not name specific datasets. Given this, it would seem that experiment datasets should consider the limitations common in edge contexts.

4.1 Comparative Analysis

Table 1: Summary of Reviewed Works

S/N	Title	Methodology	Dataset	Result
1	Edge AI and IoT: Direct integration for on-the-device data processing	Integrating and evaluating models (TinyML& TensorFlow Lite)	100 IoT devicesProcessing time, accuracy and energy consumption	Enhanced real-time processing and improved accuracy
2	Sustainable AI Processing at the Edge	Comparative analysis, life cycle assessment (LCA), activity ratios and sleep ratios	Synthetic Dataset Energy consumption and carbon emissions	The paper provides sustainable insights, operational Efficiency and holistic energy evaluation
3	An Overview on Generative AI at Scale with Edge-Cloud Computing	Thorough review and application Visualization	Synthetic Dataset User-generated content, user interactions and preferences	The paper outlines a structured methodology, discusses the importance of data management, and highlights significant achievements in providing a comprehensive overview
4	A Survey on Optimization Techniques for Edge Artificial	Hyper-parameter optimization, model compression, regularization	Synthetic DatasetProcessing power, memory and network bandwidth	The development of a framework for AI model optimization that addresses the unique challenges posed

	Intelligence (AI)	techniques, Scikit-learn, and Bayesian optimization		by edge computing
5	Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications	Distributed optimization techniques and holistic end-to-end (E2E) architecture	Synthetic Dataset Data availability, quality, and regulations compliance (GDPR)	The development of scalable edge AI systems, a holistic architecture for integration, and the identification of application scenarios highlight edge AI's transformative potential in 6G networks
6	AI-Empowered Fast Task Execution Decision for Delay-Sensitive IoT Applications in Edge Computing Networks	AI-empowered fast task execution method and simulations	Lyapunov Optimization based Dynamic Computation (LODCO) algorithm Edge server capacity, packet execution deadline, energy consumption and channel condition	The proposed framework improves decision-making speed and accuracy and provides a comprehensive framework

5. Applications of AI in Edge Computing

Applications of edge AI span numerous industries, such as autonomous cars, which need to react quickly for safety and navigation, smart cities and homes, where technology controls energy use and security, and healthcare, which uses wearable technology to provide real-time diagnostics and monitoring [9].

- a. **Smart Cities, Homes and Devices:** Traffic Management, Real-time traffic monitoring and control to reduce congestion and improve safety. Public Safety, surveillance systems use AI to detect and respond to incidents in real time. Edge AI enables real-time processing and decision-making in smart home devices, improving responsiveness and efficiency. For example, smart thermostats can adjust temperatures based on real-time data without needing to communicate with a cloud server.
- b. **Healthcare:** Remote Patient Monitoring, continuous monitoring of patient vitals using wearable devices to provide real-time health insights and alerts. Diagnostic Support, AI-powered diagnostic tools that analyze medical images and data at the point of care. In healthcare, Edge AI can be used in wearable devices to monitor patient vitals and provide real-time health analytics, potentially alerting users to critical conditions before they become severe.
- c. **Industrial IoT:** Predictive Maintenance, analyzing data from industrial equipment to predict failures and schedule maintenance, reducing downtime. Quality Control, automated inspection systems that detect defects in manufacturing processes. Edge AI enhances security systems by processing video feeds locally to detect anomalies or potential threats in real-time, reducing the latency and bandwidth needed for cloud-based analysis. Edge AI can also be applied in industrial settings to monitor machinery, predict maintenance needs, and optimize operations through real-time data analysis.
- d. **Agriculture:** Precision Farming, using sensors and AI to monitor crop conditions and optimize irrigation and fertilization. Livestock Monitoring, tracking the health and activity of livestock to improve management and productivity.
- e. **Retail:** Customer Behavior Analysis, using in-store sensors to analyze customer movements and preferences to optimize store layouts and inventory. Automated Checkout, AI systems that enable cashier-less checkout experiences. In retail, Edge AI enables personalized customer experiences through real-time data processing on in-store devices, such as smart mirrors and interactive kiosks.
- f. **Autonomous Vehicles:** Edge AI is crucial for the operation of autonomous vehicles, as it allows for real-time processing of data from sensors and cameras, facilitating immediate decisions such as obstacle avoidance and navigation.

6. Conclusion

To give researchers in the field of edgeAI a roadmap for future work, [2] effectively outlines open obstacles and future research directions in the integration of AI and edge computing. In crucial applications, like the Internet of Vehicles, where prompt and precise data processing is necessary to prevent accidents, the integration of AI with edge computing has been demonstrated to enhance user experience [17].

To improve the integration of GenAI with edge-cloud computing, researchers [6] offer architectural considerations for training and deploying GenAI systems at scale as well as future research directions. This study addresses both opportunities and problems in this quickly developing subject by offering a useful overview of the present status of generative AI and its potential for scalability through edge-cloud computing. The proper integration of AI capabilities into the network edge, as highlighted in [7], improves the overall efficacy and efficiency of 6G networks.

Regarding the issues of latency, energy consumption, and privacy problems related to classical AI systems, this integration is considered a major accomplishment. The establishment of an all-encompassing framework for AI model optimization that takes into account the particular difficulties presented by edge computing is one of the major accomplishments noted in [8]. The research effectively identifies and addresses some issues, such as hardware and energy limits, related to integrating AI capabilities into edge devices.

Edge AI enables the processing of data and decision-making on local devices, marking a breakthrough in the field of artificial intelligence. There are many advantages to switching from cloud-based AI to edge computing, such as lower latency, better privacy, and increased bandwidth efficiency [1]. Edge AI is ideal for applications that need to respond quickly, like industrial automation, smart homes, autonomous cars, and healthcare monitoring, because it allows real-time data processing.

The future of edge AI holds promising potential as improvements in software and hardware keep enhancing edge device capabilities. We may anticipate more cutting-edge uses of AI in a variety of sectors as the technology develops, which will promote its continued integration into daily life. In addition to improving user experiences, edge AI's continued development and application will help intelligent systems become more widely used, improving technology's responsiveness, security, and efficiency.

7. References

- [1]. S. Tuli *et al.*, "AI augmented Edge and Fog computing: Trends and challenges," *J. Netw. Comput. Appl.*, vol. 216, p. 103648, Jul. 2023, doi: 10.1016/j.jnca.2023.103648.
- [2]. K. S. Ali, "Edge AI and IoT: Direct integration for on-the-device data processing," *Adv. Eng. Innov.*, vol. 5, no. 1, pp. 20–23, Dec. 2023, doi: 10.54254/2977-3903/5/2023040.
- [3]. C. Lao, J. Gao, G. Ananthanarayanan, A. Akella, and M. Yu, "HawkVision: Low-Latency Modeless Edge AI Serving," May 29, 2024, *arXiv*: arXiv:2405.19213. Accessed: Aug. 08, 2024. [Online]. Available: <http://arxiv.org/abs/2405.19213>
- [4]. S. Ollivier *et al.*, "Sustainable AI Processing at the Edge," Jul. 04, 2022, *arXiv*: arXiv:2207.01209. Accessed: Oct. 04, 2024. [Online]. Available: <http://arxiv.org/abs/2207.01209>
- [5]. L. Paripati, J. Shah, N. Prasad, N. Narukulla, and V. R. Hajari, "Edge Computing for AI and ML: Enhancing Performance and Privacy in Data Analysis," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 12, no. 2, 2023.
- [6]. Y.-C. Wang, J. Xue, C. Wei, and C.-C. J. Kuo, "An Overview on Generative AI at Scale with Edge-Cloud Computing," Oct. 31, 2023. doi: 10.36227/techrxiv.23272271.
- [7]. K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022, doi: 10.1109/JSAC.2021.3126076.
- [8]. C. Surianarayanan, J. J. Lawrence, P. R. Chelliah, E. Prakash, and C. Hewage, "A Survey on Optimization Techniques for Edge Artificial Intelligence (AI)," *Sensors*, vol. 23, no. 3, p. 1279, Jan. 2023, doi: 10.3390/s23031279.
- [9]. A. Haldorai and S. Anandakumar, "Motivation, Definition, Application and the Future of Edge Artificial Intelligence," *J. Comput. Nat. Sci.*, pp. 77–87, Jul. 2022, doi: 10.53759/181X/JCNS202202011.
- [10]. H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, "Edge Computing with Artificial Intelligence: A Machine Learning Perspective," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023, doi: 10.1145/3555802.
- [11]. P. Kang and A. Somtham, "An Evaluation of Modern Accelerator-Based Edge Devices for Object Detection Applications," *Mathematics*, vol. 10, no. 22, p. 4299, Nov. 2022, doi: 10.3390/math10224299.
- [12]. W. Li, H. Hacid, E. Almazrouei, and M. Debbah, "A Comprehensive Review and a Taxonomy of Edge Machine Learning: Requirements, Paradigms, and Techniques," *AI*, vol. 4, no. 3, pp. 729–786, Sep. 2023, doi: 10.3390/ai4030039.

- [13]. S. Dayal, “AI on Edge: A Mass Accessible Tool for Decision Support Systems,” in *Decision Support Systems (DSS) and Tools [Working Title]*, IntechOpen, 2024. doi: 10.5772/intechopen.1003945.
- [14]. F. Saeik *et al.*, “Task offloading in Edge and Cloud Computing: A survey on mathematical, artificial intelligence and control theory solutions,” *Comput. Netw.*, vol. 195, p. 108177, Aug. 2021, doi: 10.1016/j.comnet.2021.108177.
- [15]. A. Acheampong, Y. Zhang, X. Xu, and D. Appiah Kumah, “A Review of the Current Task Offloading Algorithms, Strategies and Approach in Edge Computing Systems,” *Comput. Model. Eng. Sci.*, vol. 134, no. 1, pp. 35–88, 2023, doi: 10.32604/cmes.2022.021394.
- [16]. C. Hao, J. Dotzel, J. Xiong, L. Benini, Z. Zhang, and D. Chen, “Enabling Design Methodologies and Future Trends for Edge AI: Specialization and Codesign,” *IEEE Des. Test*, vol. 38, no. 4, pp. 7–26, Aug. 2021, doi: 10.1109/MDAT.2021.3069952.
- [17]. T. Nguyen, H. Nguyen, and T. Nguyen Gia, “Exploring the integration of edge computing and blockchain IoT: Principles, architectures, security, and applications,” *J. Netw. Comput. Appl.*, vol. 226, p. 103884, Jun. 2024, doi: 10.1016/j.jnca.2024.103884.