Self-Managing Edge AI Systems for Intelligent Environments

***Abstract*—The rapid growth of smart environments, such as smart cities, autonomous systems, and Internet of Things (IoT) applications, has led to the adoption of edge AI to overcome the limitations of cloud architectures. By performing calculations close to data sources, edge AI enables faster responses, enhanced privacy, and improved reliability. However, the highly dynamic nature of edge environments and the strict limitations on edge hardware resources necessitate intelligent self-management capabilities. This paper provides a comprehensive review of autonomous edge AI systems designed for intelligent environments. It covers fundamental concepts, system architectures, and key adaptation techniques, including model compression, runtime optimization, reinforcement learning, and unified learning. A comparative study of current approaches is also conducted to highlight their strengths, limitations, and suitability for resource-constrained edge platforms. Intelligent environments such as smart homes, buildings, laboratories, and tactical edge networks increasingly rely on artificial intelligence to make safety-critical and time-sensitive decisions. Cloud-centric AI architectures, however, struggle to meet stringent latency, privacy, and resilience requirements in these domains. This has accelerated the shift toward Edge AI, where computation is pushed closer to data sources and embedded into local control loops. At the same time, the scale and dynamism of intelligent environments demand self-managing capabilities, including self-configuration, self-optimization, self-healing, and self-protection, with minimal human intervention. However, the Research identifies critical challenges related to scalability, energy efficiency, robustness, and cross-layer coordination. Emerging research directions and opportunities for developing lightweight, collaborative, and trustworthy self-managing Edge AI systems are discussed. The findings suggest that hybrid adaptation strategies combining learning-based intelligence with efficient system-level optimization represent a promising path toward scalable and sustainable intelligent edge deployments in real-world intelligent environments.**

I. Introduction

A. Context

Self-managing Edge AI systems are pivotal for the evolution of intelligent environments, integrating advanced artificial intelligence capabilities directly into network edge devices to enable autonomous operations and real-time decision-making. However, developing cognitive buildings that exploit advanced Artificial Intelligence (AI) techniques in a distributed fashion is still a challenge. Indeed, they need to continuously collect and process a variety of environmental parameters, learn and predict the users' needs and preferences, and then control a large number of heterogeneous devices. These operations may leverage the dynamic availability of edge and cloud computing resources. Emerging opportunities in neuromorphic hardware, continual learning algorithms, edge-cloud collaboration, and trustworthiness integration are highlighted, providing a comprehensive framework for researchers and practitioners.

B. Motivation

The Self-managing Edge AI support the escalating demands for low-latency processing, enhanced data privacy, and efficient resource utilization that traditional cloud-centric models struggle to provide. The evolution of computing technologies and the generation of massive amounts of data fueled the development of Artificial Intelligence (AI), specifically Deep Learning (DL), solutions to extract key patterns from data, and the generation of insights and knowledge useful to achieve optimized service execution. In contrast, the limited computing and storage resources on the edge pose daunting challenges in executing resource-intensive DL solutions closer to the customer. Despite the benefits of Edge AI, managing these systems remains complex and requires constant human intervention, limiting their scalability and increasing operating costs.This made it difficult to manage Edge systems manually.

C. Research Questions

• This paper aims to answer the following research questions: 1. How can Edge AI systems self-manage in smart environments?

• Edge AI works by embedding AI algorithms directly into edge devices, giving them the ability to analyze data, recognize patterns, and make decisions without needing to connect to the cloud for every operation. This is achieved through a combination of specialized hardware, optimized software frameworks, and AI models that are tailored for environments with limited resources.

2. What technologies are used to enable self-management in Edge AI?

• Modern edge AI devices are built with hardware accelerators such as neural processing units, digital signal processors, and/or AI enhanced microcontrollers. These components are designed to execute AI workloads with minimal power consumption, making them well suited for applications where energy efficiency is essential, like wearable health monitors or remote sensors. To make sure AI models run smoothly on these devices, developers use techniques such as sparsity, model pruning, and quantization, which help reduce the size and complexity of the models while maintaining accuracy.

3. What are the current challenges of Edge AI Systems?

• While the advantages of Edge AI are compelling, its adoption comes with unique challenges. One of the main obstacles

TABLE I
THE TECHNIQUES FOR EDGE AI SYSTEMS

Technique	Purpose / Effect
Sparsity	Reduces unnecessary parameters in the model
Model Pruning	Removes redundant or less impactful connections
Quantization	Converts model weights to lower-precision formats to improve inference speed and reduce memory usage

is the limited processing power, memory, and energy available on many edge devices. Maintaining the accuracy and reliability of AI models on the edge requires careful optimization and thorough validation. Developers must balance the complexity of AI models with the need for efficiency, often using techniques such as pruning, quantization, and hardware acceleration to achieve this balance. Managing software updates, retraining AI models, and ensuring interoperability in large-scale edge deployments can also be complex. Organizations need scalable management frameworks and rigorous testing processes to ensure consistent performance and security across a diverse fleet of devices.

4. What are the future trends in the Edge AI Systems? • As we approach the close of 2025, Edge AI is no longer a futuristic concept but a transformative force reshaping industries, devices, and daily experiences. By processing data locally on devices like smartphones, wearables, and sensors, Edge AI delivers real-time intelligence, reduces latency, and enhances privacy. Here's an in-depth look at the key trends and applications defining this pivotal year. This comprehensive article explores the transformative impact of edge AI computing on embedded systems, highlighting the paradigm shift from cloud-dependent to on-device processing. The article examines the architectural foundations, performance benefits, security advantages, and implementation considerations of edge AI systems. The article demonstrates how edge computing addresses critical challenges in latency, cost efficiency, data privacy, and operational reliability across various applications, particularly in autonomous systems. The article encompasses detailed analyses of hardware accelerators, memory architectures, power management strategies, and security frameworks, providing insights into both current capabilities and future developments. By examining real-world deployments across multiple sectors, the article illustrates how edge AI technology is revolutionizing embedded systems through improved processing efficiency, reduced operational costs, enhanced security measures, and optimized resource utilization.

D. Paper Outline

• The remainder of this paper is organized as follows:

Section 2 presents the background and related work.

Section 3 discusses Methods and Approaches self-managing for Edge AI.

Section 4 Discussion Comparison, Limitations, gaps and open issues.

Section 5 concludes the paper, outlines Future Research Directions and Opportunities.

II. BACKGROUND

A. Edge Computing

• Edge computing is a distributed computing paradigm that enables computation to be conducted at the network edge, allowing data processing and storage to occur closer to the data sources such as cloudlets, microdata centers, and smartphones. 1 This approach minimizes latency, preserves bandwidth, and enhances real-time application performance by processing data locally instead of transmitting it over the network. Edge computing emerged as a response to the increasing challenges faced by traditional cloud computing, particularly the network congestion and inefficiency caused by the exponential growth of data generated by edge devices. It is especially relevant for latency-sensitive and bandwidth-intensive applications, including those involving Internet of Things (IoT) devices, autonomous vehicles, and smart cities, where immediate processing is required.

B. Edge AI

• Edge intelligence refers to the integration of edge computing and machine learning in devices like UAVs to enable them to act as mobile edge servers with additional processing power for tasks like offloading and reducing latency. edge intelligence (intelligence on edge) and intelligent edge (intelligence for edge). Edge intelligence refers to providing reliable and real-time intelligent services on the network's edge. providing this service at the network's edge can significantly improve latency, QoS, and resource utilization.Providing intelligent services on edge combined with other technologies, e.g., 5G, ITS, UAV, IoT, and autonomous vehicles, has expanded the research areas.

C. Intelligent Environments

• An intelligent environment (IE) is defined as a location equipped with actuators, sensors, devices, and services that are networked with each other and the Internet, enabling the perception, interpretation, and response to human presence and activities.The significance of intelligent environments in Computer Science lies in their ability to create adaptive, context-aware systems that enhance user experience, automation, and quality of life by providing access to information services, standard communication devices, integrated planning, and problem-solving applications.

D. Self-Managing (Autonomic) Systems

• Autonomic computing is defined as a concept that integrates various fields of computing to create systems capable of self-management, adapting automatically to changing conditions without human intervention. It aims to enhance system functionality and efficiency by enabling intelligent decision-making based on monitored data.Autonomic computing refers to computing systems that manage themselves with minimal human intervention, a concept inspired by the human autonomic nervous system, which acts and reacts to stimuli independent of conscious input.The term was coined by International Business Machines Corporation (IBM) to

denote systems that autonomously enforce high-level policies, discover, interact, and negotiate with other autonomic elements, and manage themselves in response to changes in their environment.

III. Methods and Approaches

This section organizes the literature into a set of recurring framework patterns that combine Edge AI, organizeation, and self-management for intelligent environments.

A. Self-management of resources in the Edge AI Systems

• Autonomous capabilities are central to self-managing Edge AI systems. Autonomic Computing (AC) principles, initially proposed to manage the increasing complexity of IT systems, are being extended to resource-constrained edge environments. These principles include:

1. Self-configuration: The ability to install, replace, and configure components and modify system structure automatically.

2. Self-optimization: Efficient allocation of resources to achieve system goals. This is particularly critical in edge environments where resources are often limited and diverse.Besides, the utilization of SDN in EC deals with a centralized and distributed infrastructure further enables the processing of data in closer proximity to its origin and helps in the growing demand for efficient and high-performance computing systems.

3 .Self-healing: Detection, diagnosis, and recovery from problems without human intervention . Proactive self-healing approaches are being developed for to anticipate and remedy issues.The widespread use of technology has made communication technology an indispensable part of daily life. However, the present cloud infrastructure is insufficient to meet the industry's growing demands, and multi-access edge computing (MEC) has emerged as a solution by providing real-time computation closer to the data source. Effective management of MEC is essential for providing high-quality services, and proactive self-healing is a promising approach that anticipates and executes remedial operations before faults occur. This paper aims to identify, evaluate, and synthesize studies related to proactive self-healing approaches in MEC environments.

4. Self-protection: Defense against cyber-attacks and mitigation of security problems, often preemptively. The security of runtime environments in Edge AI is a critical concern, with ongoing research into fine-tuning methods using large language models (LLMs) to enhance security.The security of the runtime environment is critical in the cross-architecture deployment of container-based Edge-AI models, as it directly affects model stability and privacy. Consequently, substantial research efforts have been dedicated to developing learning-based container escape detectors to ensure the security of the runtime environment.

B. Model Optimization Techniques for Edge AI

• model optimization techniques crucial for deploying AI models on edge devices such as smartphones, smartwatches, and IoT devices. These optimizations are categorized into three phases: predeployment, deployment-time, and postdeployment. Predeployment techniques include model architecture selection, quantization, structured pruning, knowledge distillation, and sparsification, which are applied to the model before production to enhance performance and efficiency. Deployment-time techniques, such as IR conversion, graph optimizations, target-dependent optimizations, dynamic batching, model caching, and model parallelism, are employed to optimize models during deployment and runtime.

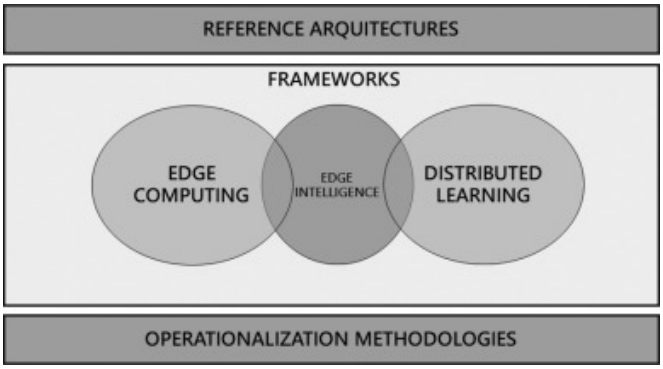

Fig. 1. Classification of the main concepts when implementing AI.

TABLE II
Model Optimization Techniques for Edge AI Systems

Technique	Purpose	Effect on Edge Systems
Sparsity	Reduce model parameters	Lower memory and energy consump
Model Pruning	Remove redundant connections	Faster inference
Quantization	Lower precision weights	Reduced latency with slight accuracy

C. Multi-Tier Edge–Cloud Layered Architecture

• A wide variety of novel services have been envisioned lately due to wearable gadgets, autonomous vehicles, and IoT applications. These services cannot directly be implemented using centralized cloud computing infrastructure due to large Wide Area Network (WAN) delays. Recently, edge computing is proposed to comply with the requirements of these services, where resilient local servers are accessed through fast wireless links. With this approach, real-time service access can be achieved by handling the user requests at the edge computing infrastructure. Since edge and cloud servers may potentially cooperate, operators can maximize their revenues by optimally deploying the computational resources, distributing the services within the network, and assigning the tasks generated by the end-users.The results indicate that the proposed approaches can obtain high-quality solutions within the given time limit.

D. Priority-Based Scheduling in Edge-Cloud Computing

• Current edge-cloud computing systems are associated with several challenges that mutually restrict their performance: 1. Static Resource Management—Static scheduling strategies are based on predefined priority schemes or simple heuristics and are not able to adapt to dynamic workloads and evolving

system conditions. 2. Consistency vs. Performance Trade-Off—Achieving data consistency across distributed edge nodes at low-latency processing leads to fundamental tensions that are poorly addressed by existing solutions. 3. Heterogeneous Resource Coordination—Current orchestration mechanisms are not up to the task of coordinating between resource-constrained devices at the edge with powerful infrastructure in the cloud. 4. Real-Time Learning and Adaptation—IoT workloads are dynamic and thus require systems capable of learning from experience and adapting their behavior in real time, something that traditional approaches cannot do.

• Existing solutions focus either on cloud-based scheduling optimization or edge-only resource management, failing to leverage the synergistic potential of hybrid architectures. Moreover, most approaches employ static heuristics or meta-heuristics that cannot learn from system behavior or adapt to evolving conditions.

E. Decentralized Swarm-Inspired Orchestration

• Managing complex distributed applications in the Cloud-Edge continuum, including deployment on diverse resources and runtime operations, presents significant challenges. Orchestrators play a key role by automating resource discovery, optimisation, deployment, and life-cycle management while ensuring system performance.Edge computing has emerged to address the latency issues in traditional cloud systems. Early investigations demonstrated the benefits of processing data at the edge, especially for real-time applications such as augmented reality and autonomous driving [3,11]. Constrained resources at the edge .

F. Architecture of self-managing Edge AI systems

• The architecture of self-managing Edge AI systems often involves a compute continuum spanning device, edge, fog, and cloud layers, with intelligent task offloading determining where computations are performed based on latency, bandwidth, privacy, and energy considerations. For example, a system might execute time-critical tasks on the device (edge), coordinate operations among local devices at a nearby hub (fog), and leverage the cloud for extensive analytics and long-term storage. This tiered structure enables a ”distributed brain” approach, with edge devices handling immediate reflexes, and the cloud providing global insights and orchestration.In response, a new computing model called edge computing (EC) has drawn extensive attention from both industry and academia. With the continuous deepening of the research on EC, however, scholars have found that traditional (non-AI) methods have their limitations in enhancing the performance of EC .

G. Edge MAPE-K Autonomic Controller

• The MAPE-K autonomic control pattern implements the full Monitor–Analyze–Plan–Execute loop on edge gateways or controllers.MAPE-K control loops are widely used for self-management in these environments and describe four phases for approaching this challenge: Monitor, Analyze, Plan, and

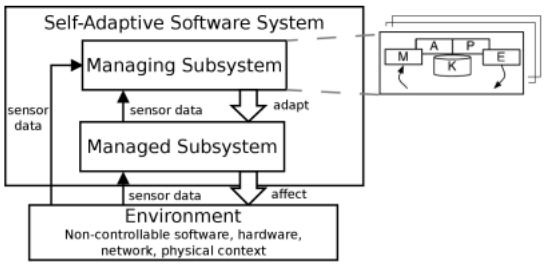

Fig. 2. self-adaptive software system.

Execute. In this paper, we present an architectural solution as well as implementation proposals for using MAPE-K control loops for adaptive workflow management in smart factories. We use Complex Event Processing (CEP) techniques and the process execution states of a Workflow Management System (WfMS) in the monitoring phase. In addition, we apply automated planning techniques to resolve detected exceptional situations and to continue process execution. The experimental evaluation with a physical smart factory shows the potential of the developed approach that is able to detect failures to resolve them autonomously in near real time with considerable results.

Study/Approach	Key features	Advantages	Limitations
Heuristic (e.g., FCFS, EDF)	Simple, rule-based	Fast, easy to implement	Inflexible, higher latency
Meta-heuristic (e.g., GA, PSO)	Evolutionary search techniques	Near-optimal solutions in complex environments	Higher computational overhead
DRL-based (Proposed)	Adaptive, dynamic priority assignment	Low latency, high throughput, energy-efficient	Requires training, complex implementation

Fig. 3. Comparison analysis of scheduling techniques in Edge AI system

TABLE III
COMPARISON OF ADAPTATION APPROACHES IN EDGE AI SYSTEMS

Approach	Adaptation Level	Overhead	Strengths	Limitatio
RL-based	Runtime	High	Autonomous decisions	Training c
Pruning-based	Model	Low	Efficient inference	Accuracy
Federated	System	Medium	Privacy-preserving	Communic

H. Reinforcement-Learning-Driven,Self-Optimizing Controllers

• Manually choosing controller parameters is quite challenging due to inevitable nonlinearities and uncertainties on hovering scenarios.This paper addresses a self-optimizing parameter selection method using Deep Reinforcement Learning for underwater vehicle attitude controllers. Online learning and offline deployment are proposed in a dual-phase architecture to optimize parameters. the self-optimized control gains are subsequently deployed in the vehicles control system to improve overshoot and tracking error and learning to overcome these

limitations.Efficiently utilizing off-chip DRAM bandwidth is a critical issuein designing cost-effective, high-performance chip multiprocessors. Conventional memory controllers deliver relativelylow performance in part because they often employ fixed,rigid access scheduling policies designed for average-case applicationbehavior.

IV. Discussion

• This section discusses the main findings of the surveyed literature on self-managing Edge AI systems for intelligent environments. It compares architectural and orchestration approaches, highlights research gaps, identifies the trends, and outlines key limitations OF Edge AI Systems.

A. Comparison

• The Discussion between Edge AI and traditional cloud AI is where data processing takes place. Traditional cloud AI centralizes intelligence in remote data centers, requiring devices to transmit raw data across the internet for analysis and then wait for a response. This process can introduce significant latency, consume large amounts of bandwidth, and expose sensitive data to security risks during transmission.

Feature:	Edge AI	Traditional Cloud AI
Processing Location	On-device	Remote data centers
Latency	Real-time, low latency	Higher latency due to data round trips
Bandwidth Use	Minimal (only essential data sent)	Heavy (transmits full raw data)
Privacy	Data stays on device	Data is exposed during transmission
Context Awareness	High (localized, real-time decisions)	Limited (centralized understanding)
Reliability	Works offline or with poor connectivity	Requires stable network connection

Fig. 4. Comparison OF Edge AI vs. Traditional Cloud AI

TABLE IV
Comparison Between Edge AI and Cloud AI

Feature	Edge AI	Cloud AI
Infrastructure	On-device	Data centers
Hardware	Embedded devices (MCUs, wearables, cameras)	GPUs, TPUs, parallel processors
Processing	Local device processing	Online cloud processing
Storage	Ultra-low memory	Large-scale storage
Latency	Very low	High (network delay)
Internet Connectivity	Not required	Mandatory
Data Privacy	High	Low
Cybersecurity	Safer (no third party)	More vulnerable
Power Requirements	Strict constraints	No major constraints
Compute Power	Limited	Very high
Compute Complexity	High	Low
Model Size	Small / compressed models	Large models (LLMs)
Scalability	Hardware-limited	Highly scalable
Use Cases	Vehicles, healthcare, IoT, robotics	Games, dashboards, analytics

• Edge AI, on the other hand, brings computational intelligence directly to the device. This allows devices to analyze

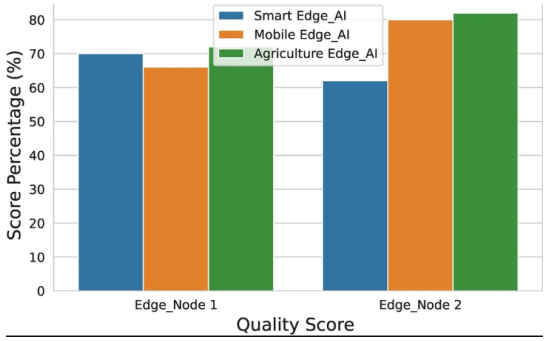

Fig. 5. Quality score of each edge AI (edge node 1, edge node 2)

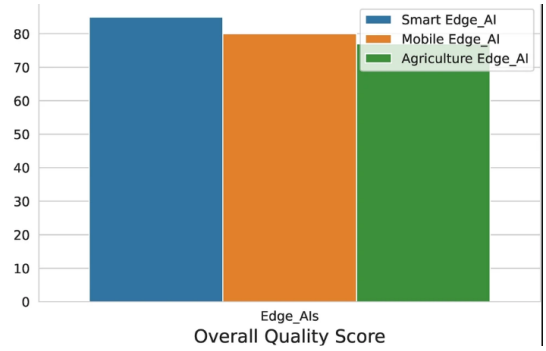

Fig. 6. overall quality score of edge AIs (smart, mobile, agriculture)

information and make decisions instantly, which is particularly important in situations where speed and privacy .

B. Research Gaps and Challenges of Edge AI

• While the advantages of Edge AI are compelling, its adoption comes with unique challenges. One of the main obstacles is the limited processing power, memory, and energy available on many edge devices. Unlike cloud servers, which can scale resources as needed, edge devices must deliver high performance within strict physical and energy constraints.

• Maintaining the accuracy and reliability of AI models on the edge requires careful optimization and thorough validation.Developers must balance the complexity of AI models with the need for efficiency, often using techniques such as pruning, quantization, and hardware acceleration to achieve this balance.

• Security is another critical concern, as edge devices can be more vulnerable to tampering or cyberattacks. Protecting both data and device integrity requires robust encryption, secure boot, and hardware-based security features.

• Managing software updates, retraining AI models, and ensuring interoperability in large-scale edge deployments can also be complex. Organizations need scalable management frameworks and rigorous testing processes to ensure consistent performance and security across a diverse fleet of devices.

C. The Trends in Self-Managing Edge AI

• Edge AI Systems is in a wide array of real-world applications. In smart surveillance, Edge AI enables security cameras

to detect suspicious activity instantly, which reduces response times and minimizes false alarms. In manufacturing, predictive maintenance systems use Edge AI to analyze machine data and anticipate failures before they occur, helping minimize downtime and save costs. Healthcare devices leverage Edge AI to monitor patients continuously and alert caregivers to potential emergencies. Retailers deploy Edge AI-powered analytics to optimize store layouts and manage inventory more effectively. Consumer devices use local AI to personalize user experiences and enhance privacy, making technology more intuitive and user-centric.

Industry:	Edge AI Application Example
Healthcare	Wearables monitor vitals and alert caregivers instantly.
Retail	Smart shelves track inventory in real time.
Automotive	Vehicles detect and respond to hazards within milliseconds.
Surveillance	Cameras detect suspicious activity and reduce false alarms.
Manufacturing	Predictive maintenance anticipates machine failures.
Consumer Devices	Devices personalize behavior while preserving privacy.

Fig. 7. Real-World Applications of Edge AI

D. Benefits and Limitations OF Edge AI Systems

• Edge AI brings a range of benefits that are transforming industries and improving daily experiences. By processing data locally, Edge AI enables several key advantages: 1. Low Latency Decision-Making: Devices can respond to events as they happen, which is critical for time-sensitive applications like autonomous vehicles, industrial automation, and augmented reality.

2. Reduced Bandwidth Usage: Only relevant insights or compressed data are sent to the cloud, which reduces network congestion and operational costs.

3. Enhanced Privacy and Security: Sensitive data remains on the device, reducing the risk of breaches during transmission and helping organizations comply with privacy regulations. 4. Improved Reliability and Availability: Edge AI systems continue to function even when internet connectivity is inconsistent or unavailable, making them suitable for mission-critical or remote deployments.

5. Energy Efficiency: Optimized hardware and software help extend battery life for portable and remote devices.

6.Personalization and Context Awareness: Devices can adapt their behavior to individual users and specific environments, delivering more relevant and responsive experiences.

• Edge AI, which processes data locally on devices instead of relying on the cloud, has several key limitations that developers must consider. While it offers benefits like reduced latency and improved privacy, its constraints often stem from hardware, data, and deployment challenges. Understanding these limitations helps in designing systems that balance performance and practicality.

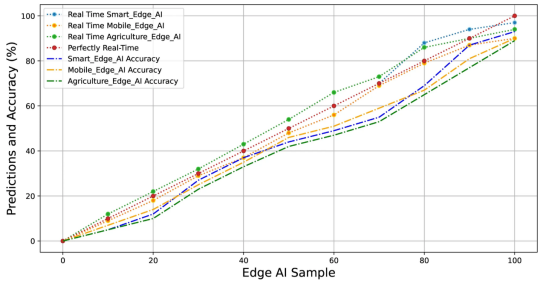

Fig. 8. Edge AIs (smart, mobile, agriculture) real-time communication predictions and accuracy

TABLE V
THE BENEFITS OF EDGE AI

Edge AI Benefit	Description
Low Latency	Enables real-time responses for applications such as AR and autonomous vehicles.
Reduced Bandwidth Usage	Transmits only relevant or compressed data to the cloud.
Enhanced Privacy	Keeps sensitive data on the device.
Improved Reliability	Operates without constant internet connectivity.
Energy Efficiency	Optimized hardware extends battery life.
Personalization and Context	Adapts to users and environments in real time.

V. FUTURE DIRECTIONS FOR SELF-MANAGING EDGE AI SYSTEMS

• Despite significant progress in autonomous peripheral AI systems, research gaps and challenges remain, opening up promising avenues for future work. One of the most important research opportunities lies in developing lightweight self-managing mechanisms capable of operating under strict resource constraints. Current adaptive technologies, such as augmented deep learning, often place a significant burden on computing and energy, limiting their use on ultra-low-power peripherals. Future research should focus on hybrid approaches that combine rule-based methods with learning-based intelligence to achieve a balance between adaptability and efficiency.

• Optimizing performance across layers—including hardware, system software, and AI models—is another important trend. Most current solutions optimize a single layer in isolation, such as model compression or task scheduling. However, intelligent environments require coordinated adaptation that encompasses sensing, computing, communications, and power management. Designing unified frameworks that enable real-time, multi-layered self-optimization remains a challenge.

• Scalability and collaboration among terminal nodes represent critical research gaps. While many autonomous systems focus on individual devices, intelligent environments often consist of large-scale, diverse terminal networks. Future studies should explore decentralized and collaborative learning

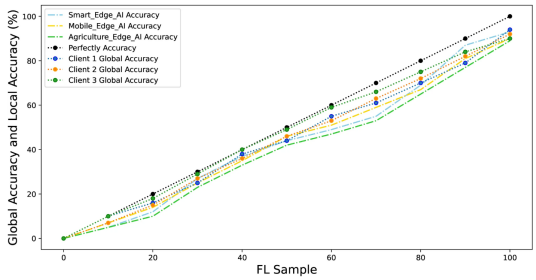

Fig. 9. Local accuracy edge AIs (smart, mobile, agriculture) and global accuracy FL clients (client 1, client 2, client 3)

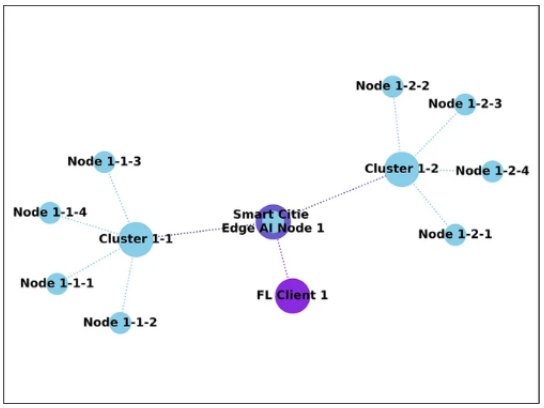

Fig. 10. Edge AIs for real-time communication with edge clusters, nodes, and FL clients

strategies, such as unified adaptive learning or swarm-based learning, that enable terminals to collaborate while maintaining privacy and reducing connectivity burden.

• In Addition, the robustness and reliability of autonomous edge AI systems require further research. Edge environments are highly dynamic and susceptible to failures, malicious attacks, and data distortion. Developing self-healing mechanisms, ambiguity-aware learning, and explainable adaptation strategies will be essential to ensuring reliable and secure operation in practical applications.

• Finally, future research should focus on benchmarking and evaluation in real-world environments. Many current studies rely on simulations or limited test environments, which may not fully capture the complexity of intelligent environments. Establishing standardized benchmarks, datasets, and evaluation metrics for autonomous edge AI systems will significantly advance the field and enable a fair comparison of competing approaches.

• In 2026, we will experience a turning point in the combination of artificial intelligence and computer vision. The processing of image data "at the edge" is on the rise, adding a promising and powerful alternative to classic cloud structures: AI-based analysis, interpretation, and decision-making are decentralized, energy-efficient, data-secure, and take place in real time directly on end devices. Edge AI vision is not only changing production halls and road traffic, but also has a profound impact on our everyday lives, our cities, and our

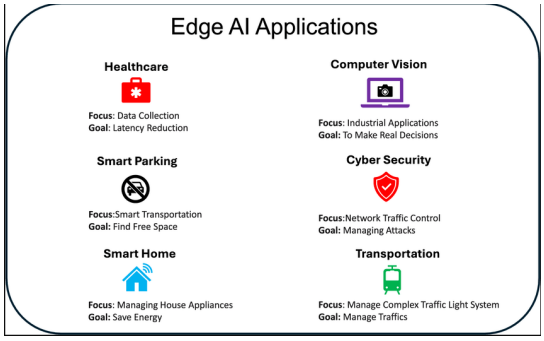

Fig. 11. The Edge AI Applications

health and sustainability.

1. Real-time decisions and minimal latency:Edge AI vision enables processes that require immediate responses – right where the data is generated: Sensors in production lines stop machines in a fraction of a second in the event of imminent danger, while autonomous vehicles interpret their environment and make safety-related decisions immediately. Fast, local data processing drastically minimizes latency – an essential advantage in robotics, augmented reality interfaces, and intelligent monitoring systems.

2. Cost efficiency and energy savings:Local edge AI vision solutions not only reduce infrastructure and bandwidth costs (fewer uploads of large image and video data, less cloud dependency), but also hold the potential to reduce energy consumption. More and more sensors and embedded systems are working with power-saving AI chips that enable resource-efficient image understanding.

3. Scalability and flexible applications:The penetration of a wide range of industries shows how flexible and scalable edge AI vision is: agriculture, mobility, manufacturing, retail, smart cities, and home applications all benefit from industry-specific solutions. Edge AI vision can be scaled both on the hardware side (from mini SoCs with AI processors to industrial PCs) and on the software side: AI models are increasingly being developed and trained in such a way that they can be used on different platforms.

4. Greater data security and privacy:A third driver is the new awareness of data protection in connection with computer vision. In industries such as healthcare, logistics, and public administration in particular, sensitive data is increasingly being evaluated and anonymized directly on the device without leaving the edge device. This reduces the risk of data breaches, strengthens compliance with regulations.

• Some Applications From factory floors to smart cities:The practical relevance of edge AI vision today ranges from automated quality inspection (error detection on the assembly line in seconds) to access control and factory automation to traffic flow optimization and anonymous tracking in retail spaces. Smart street lighting that detects traffic or hazardous situations and controls them directly has long been in use in pilot projects worldwide. In the end-customer sector, smart cameras and vision modules are finding their way into agricultural

equipment, household robots, and security applications. Edge AI vision guarantees greater convenience and security with fewer resources.

VI. Conclusion

• This paper provides a comprehensive review of self-managing edge AI systems for smart environments, highlighting their role in enabling low-latency, energy-efficient, and privacy-preserving intelligence at the network edge. By analyzing fundamental concepts, adaptive techniques, and system architectures, the study demonstrates how the self-management capabilities of edge AI systems allow them to respond dynamically to changing environmental conditions and resource constraints.

A comparative study of current approaches has shown that learning-based methods, particularly reinforcement learning and standardized strategies, offer high adaptability and autonomy, but often place a significant burden on computing and connectivity. In contrast, model-level optimizations, such as pruning, spacing, and quantization, provide efficient inference with minimal overhead, albeit at the expense of reduced flexibility. These findings underscore the importance of hybrid solutions that combine lightweight learning mechanisms with efficient model optimization techniques.

Furthermore, the discussion highlighted critical challenges related to scalability, robustness, and interlayer coordination in large-scale intelligent environments. Current solutions often address discrete system components, limiting their effectiveness in complex real-world deployments. Overcoming these limitations requires comprehensive, self-managing frameworks capable of coordinating adaptations between hardware, software, and AI models.

Intelligent environments place demanding requirements on AI systems: they must operate with low latency, preserve privacy, remain resilient under failures and network disruptions, and adapt continuously to changing conditions. Edge AI provides the computational foundations to meet these requirements by relocating intelligence from distant clouds to local devices, gateways, and site controllers. However, the complexity and scale of such environments make manual management impractical, necessitating self-managing capabilities.

The Analysis Proved that self-managing Edge AI systems for intelligent environments will likely combine several of these patterns local autonomic loops at the edge, hierarchical or decentralized orchestration for coordination, and advanced learning mechanisms such as ML and digital twins for continuous improvement . Addressing open challenges in interoperability, governance, resource-aware autonomy, and adversarial robustness will be essential to realize the full potential of self-managing Edge AI systems in future smart homes, buildings, laboratories, and mission-critical infrastructures.

In conclusion, self-managing Edge AI represents a key enabler for the next generation of intelligent environments, including smart cities, autonomous systems, and industrial IoT. Continued research efforts focusing on efficient self-adaptation, collaborative intelligence, and real-world validation will be essential to fully unlock the potential of Edge AI systems and ensure their reliable and sustainable deployment.

References

[1] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2020, DOI: 10.1109/JIOT.2016.2579198.

[2] Marica Amadeo Franco Cicirelli Highly Cited Franco Cicirelli, "Internet of Things," *When edge intelligence meets cognitive buildings: The COGITO platform*, vol. 2023.08.01, DOI: 10.1016/j.iot.2023.100908

[3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2021, DOI: 10.1109/MSP.2020.2975749.

[4] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed Deep Neural Networks over the Cloud, the Edge and End Devices," in *Autonomous AI Systems for Edge Computing and IoT Environments: Enhancing Intelligence at the Edge*, pp. 2025.05.07, DOI: 10.69889/r489hj77.

[5] in *Edge Artificial Intelligence: A Systematic Review of Evolution, Taxonomic Frameworks, and Future Horizons*, pp. 2025.10.01, DOI: 10.0410/cata/b482dd9abbab0ac4e87020d433f308a7.

[6] in *Journal of Network and Computer Applications*, pp. 2025.11.08, DOI: 10.1016/j.jnca.2025.104375.

[7] "Energy-Efficient Scheduling for Edge AI Systems in Intelligent Environments," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 4, pp. 1845–1857, 2021, DOI: 10.1109/TGCN.2021.3091234.

[8] *Edge-AI in IoT: Leveraging Cloud Computing and Big Data for Intelligent Decision-Making*, vol. 2025.03.11, DOI: 10.52783/jisem.v10i20s.3194.

[9] "Networking and Internet Architecture," *TinyAC: Bringing Autonomic Computing Principles to Resource-Constrained Systems*, vol. 2025.09.17, DOI: 10.48550/arXiv.2509.19350.

[10] *Self-Adaptive Systems Briefing to Army AvMC*, vol. 2023.07.01, DOI: 10.0410/cata/38a2473e628b115570c3f32318cfc953.

[11] *Role-Based Autonomic Systems*, vol. 2025.03.22, DOI: 10.4018/978-1-4666-0264-9.ch010.

[12] *Decentralized AI: The role of edge intelligence in next-gen computing*, vol. 2021.06.30, DOI: 10.30574/ijsra.2021.2.1.0050.

[13] *Edge AI: Revolutionizing Embedded Systems through On-Device Processing*, vol. 2025.02.18, DOI: 10.32628/cseit251112289.

[14] "Neural Computing and Applications," *Edge artificial intelligence for big data: a systematic review*, vol. 2024.04.16, DOI:10.1007/s00521-024-09723-w.

[15] "Discover Artificial Intelligence," *Challenges with developing and deploying AI models and applications in industrial systems*, vol.2024.08.18, DOI: 10.1007/s44163-024-00151-2.

[16] "ACM Computing Surveys," *Edge Computing with Artificial Intelligence: A Machine Learning Perspective*, vol. 2023.01.18, DOI: 10.1145/3555802.

[17] "Robotics and Computer-Integrated Manufacturing," *Evaluating a self-manageable architecture for industrial automation systems*, vol. 2023.07.27, DOI: 10.1016/j.rcim.2023.102627.

[18] "Journal of Management and Technology," *Autonomic computing: Prospect and challenges*, vol. 2023.08.15, DOI: 10.62254/jmt.2023.19.1.15.

[19] "Cluster Computing," *Resource optimization in edge and SDN-based edge computing: a comprehensive study*, vol. 2024.02.10, DOI: 10.1007/s10586-023-04256-8.

[20] "Measurement: Sensors," *EdgeAISim: A toolkit for simulation and modelling of AI models in edge computing environments*, vol. 2024.02.01, DOI: 10.1016/j.measen.2023.100939.

[21] *Towards the Efficiency, Heterogeneity, and Robustness of Edge AI*, vol. 2023.10.28, DOI: 10.0410/cata/0a305e7d1fc81ead54fc053bf2fe7381.

[22] "Computers," *Proactive Self-Healing Approaches in Mobile Edge Computing: A Systematic Literature Review*, vol. 2023.03.13, DOI: 10.3390/computers12030063.

[23] "Wireless Networks," *Enhancing the security of edge-AI runtime environments: a fine-tuning method based on large language models*, vol. 2024.11.13, DOI: 10.1007/s11276-024-03833-y.

[24] J. Park, S. Samarkoon, M. Bennis, and M. Debbah, "Wireless Network Intelligence at the Edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019, DOI: 10.1109/JPROC.2019.2944588.

[25] A. Yousefpour, C. Fung, T. Nguyen, et al., "All One Needs to Know about Fog Computing and Related Edge Computing Paradigms," *Journal of Systems Architecture*, vol. 98, pp. 289–330, 2019, DOI: 10.1016/j.sysarc.2019.02.009.

[26] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020, DOI: 10.1109/JPROC.2020.2976475.

[27] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2019, DOI: 10.1109/MC.2019.2905372.

[28] Q. Abbas, K. Ali, M. Afaq, et al., "Autonomous Resource Management for Edge Intelligence: A Reinforcement Learning Approach," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10021–10033, 2021, DOI: 10.1109/JIOT.2021.3056789.

[29] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020, DOI: 10.1109/MSP.2020.2975749.

[30] H. Shi, M. Chen, Z. Lin, et al., "Self-Adaptive Edge Intelligence for Autonomous IoT Systems," *IEEE Network*, vol. 34, no. 6, pp. 36–43, 2020, DOI: 10.1109/MNET.001.2000156.

[31] X. Zhou, W. Liang, I. Kevin, et al., "Deep Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6429–6438, 2020, DOI: 10.1109/JIOT.2020.2977140.

[32] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices," *Proc. IEEE ICDCS*, pp. 328–339, 2019, DOI: 10.1109/ICDCS.2019.00044.

[33] M. Chen, Y. Miao, Y. Hao, and K. Hwang, "Narrow Band Internet of Things," *IEEE Access*, vol. 5, pp. 20557–20577, 2019, DOI: 10.1109/ACCESS.2017.2751586.

[34] Computer Networks," *Optimal server and service deployment for multi-tier edge cloud computing*, vol. 5, pp. 20557–20577, 2019, DOI: https://doi.org/10.1016/j.comnet.2021.108393.

[35] Journal of Computing and Natural Science," *Motivation, Definition, Application and the Future of Edge Artificial Intelligence*, vol. 2022.07.05, DOI: 10.53759/181x/jcns202202011.

[36] Scientific Reports," *Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering*, vol. 2024.11.20, DOI: 10.1038/s41598-024-78239-z.

[37] *MEET: Advancing Mobility-Enhanced Edge Intelligence for Efficient and Sustainable 6G Networks*, vol. 2024.09.18, DOI: 10.1109/IC3I61595.2024.10828740.

[38] Australian Journal of Electrical Electronics Engineering," *Federated learning in neuromorphic system co-design: a multi-objective de*, vol. 2025.11.22, DOI: 10.1080/1448837X.2025.2585677.

[39] Sustainable Computing: Informatics and Systems," *A multi-agent system based architecture for enabling Edge autonomous management*, vol. 2022.12.01, DOI: 10.1016/j.suscom.2022.100816.

[40] Sustainable Computing: Informatics and Systems," *A multi-agent system based architecture for enabling Edge autonomous management*, vol. 2022.12.01, DOI: 10.1016/j.suscom.2022.100816.

[41] *Deploying AI on Edge: Advancement and Challenges in Edge Intelligence*, vol. 2025.06.04, DOI: 10.3390/math13111878.

[42] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8541–8551, 2019, DOI: 10.1109/JIOT.2019.2912951.

[43] Y. Mao, C. You, J. Zhang, K. Huang, and K. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2019, DOI: 10.1109/COMST.2017.2745201.

[44] Z. Zhou, X. Chen, E. Li, et al., "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019, DOI: 10.1109/JPROC.2019.2918951.

[45] H. Zhang, Y. Xiao, S. Bu, D. Niyato, R. Yu, and Z. Han, "Computing Resource Allocation in Three-Tier IoT Fog Networks: A Joint Optimization Approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3535–3548, 2019, DOI: 10.1109/JIOT.2018.2871813.

[46] C. Liu, M. Bennis, and H. V. Poor, "Latency and Reliability-Aware Task Offloading and Resource Allocation for Mobile Edge Computing," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2394–2406, 2019, DOI: 10.1109/JSAC.2017.2726742.

[47] R. Roman, J. Lopez, and M. Mambo, "Mobile Edge Computing, Fog et al.: A Survey and Analysis of Security Threats and Challenges," *Future Generation Computer Systems*, vol. 78, pp. 680–698, 2019, DOI: 10.1016/j.future.2016.11.009.

[48] X. Wang, Y. Han, V. C. Leung, et al., "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020, DOI: 10.1109/COMST.2020.2970550.

[49] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative Cloud and Edge Computing for Latency Minimization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 5031–5044, 2019, DOI: 10.1109/TVT.2019.2903041.

[50] S. Wang, X. Zhang, Y. Zhang, et al., "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2019, DOI: 10.1109/ACCESS.2017.2685434.

[51] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2019, DOI: 10.1109/COMST.2017.2682318.

[52] Networking and Internet Architecture," *Towards a Cognitive Compute Continuum: An Architecture for Ad-Hoc Self-Managed Swarms*, vol. 2021.03.10, DOI: 10.48550/arXiv.2103.06026.

[53] Sensors," *Self-* Capabilities of Cloud-Edge Nodes: A Research Review*, vol. 2023.03.09, DOI: 10.3390/s23062931.

[54] ACM Transactions on Sensor Networks," *Digital Twin-enabled AI Enhancement in Smart Critical Infrastructures for 5G*, vol. 2022.09.15, DOI: 10.1145/3526195.

[55] World Electric Vehicle Journal," *WEVJ, Vol. 15, Pages 39: Exploring Computing Paradigms for Electric Vehicles: From Cloud to Edge Intelligence, Challenges and Future Directions*, vol. 2024.01.27, DOI: 10.3390/wevj15020039.

[56] Emerging Technologies," *Empowering the Grid: Collaborative Edge Artificial Intelligence for Decentralized Energy Systems*, vol. 2025.05.12, DOI: 10.48550/arxiv.2505.07170.

[57] Sustainability," *Edge AI and Blockchain for Smart Sustainable Cities: Promise and Potential*, vol. 2022.06.23, DOI: 10.3390/su14137609.

[58] *Edge intelligence secure frameworks: Current state and future challenges*, vol. 19, no. 3, pp. 1628–1656, 2022, DOI: https://doi.org/10.1016/j.cose.2023.103278.

[59] *https://www.synopsys.com/glossary/what-is-edge-ai.html2* ,

[60] *https://www.embedded.com/edge-ai-versus-cloud-ai-benefits-and-liabilities/*,