

Edge Artificial Intelligence: A Systematic Review of Evolution, Taxonomic Frameworks, and Future Horizons

Mohamad Abou Ali^{1,3,4} and Fadi Dornaika^{*1,2}

¹*University of the Basque Country*, ²*IKERBASQUE*, ³*Lebanese International University (LIU)*, ⁴*The International University of Beirut*,

mohamad.abouali01@liu.edu.lb, fadi.dornaika@ehu.eus

Abstract

Edge Artificial Intelligence (Edge AI) embeds intelligence directly into devices at the network edge, enabling real-time processing with improved privacy and reduced latency by processing data close to its source. This review systematically examines the evolution, current landscape, and future directions of Edge AI through a multi-dimensional taxonomy including deployment location, processing capabilities such as TinyML and federated learning, application domains, and hardware types. Following PRISMA guidelines, the analysis traces the field from early content delivery networks and fog computing to modern on-device intelligence. Core enabling technologies such as specialized hardware accelerators, optimized software, and communication protocols are explored. Challenges including resource limitations, security, model management, power consumption, and connectivity are critically assessed. Emerging opportunities in neuromorphic hardware, continual learning algorithms, edge-cloud collaboration, and trustworthiness integration are highlighted, providing a comprehensive framework for researchers and practitioners.

Keywords— Edge Artificial Intelligence, Systematic Review, Tiny Machine Learning (TinyML), Tiny Deep Learning (TinyDL), Tiny Reinforcement Learning (TinyRL), Federated Learning, Multi-Access Edge Computing (MEC), Hardware Accelerators, AI Privacy and Security

1 Introduction

1.1 The Imperative for Edge AI: A Paradigm Shift

The convergence of massive-scale Internet of Things (IoT) deployment and the critical need for real-time, intelligent decision-making has necessitated a fundamental evolution beyond traditional cloud-centric computing architectures. This transition is driven by the impracticalities of

*Corresponding author

cloud-dependent models—namely latency, bandwidth, privacy, and operational resilience—in applications ranging from autonomous vehicles to personalized healthcare. Edge Artificial Intelligence (Edge AI) emerges as the foundational response to these challenges, representing a paradigm that embeds computational intelligence directly into devices at the network periphery [1, 2]. By processing data locally, at or near its source, Edge AI enables unprecedented responsiveness, privacy preservation, and operational efficiency [3, 4].

1.2 Methodological Foundation and Analytical Framework

This review adopts a systematic methodology guided by PRISMA 2020 guidelines [5] to ensure a comprehensive, unbiased, and reproducible analysis of the Edge AI landscape. From an initial corpus of over 2,200 identified records, our rigorous screening process yielded 79 primary studies for in-depth qualitative synthesis, forming the analytical core of this review.

The cornerstone of our analysis is a novel multi-dimensional taxonomy (Figure 1) that provides an integrated framework for classifying and understanding Edge AI research [6]. This taxonomy synthesizes four critical dimensions: deployment location (D1), which spans from **Device Edge and Network Edge to Regional Edge/Multi-Access Edge Computing (MEC) and Cloud Edge** [7, 8]; processing capability (D2), encompassing **TinyML, TinyDL** [9, 10], **TinyRL** [11], and **federated learning** [12, 13] paradigms; application domain (D3), including **healthcare** [14], **industrial IoT** [15], **autonomous systems** [16], and **smart cities** [17]; and hardware architecture (D4), which covers **CPUs, ASICs, FPGAs, GPUs, and neuromorphic chips** [18, 19, 20, 21, 22]. This integrated framework enables the systematic identification of research gaps, technological trade-offs, and future opportunities across the entire Edge AI ecosystem.

1.3 Research Gaps and Contributions

Our systematic analysis reveals significant limitations within the current Edge AI literature, which this review addresses through its novel methodological approach. First, a notable historical fragmentation exists, as prior surveys [23, 24] lack comprehensive historical contextualization by failing to connect modern Edge AI developments to their technological origins in content delivery networks (CDNs) and fog computing. Second, current works exhibit isolated technological analysis; studies such as [25, 26] examine hardware, software, and application layers in isolation, thereby neglecting their critical interdependencies. Third, there is an incomplete challenge assessment across the literature, where existing reviews [27, 28] provide only partial coverage of the Edge AI challenge landscape by emphasizing either optimization or security in isolation. A systematic comparison of key Edge AI surveys in Table 1 further elucidates these research gaps.

1.4 Our Contributions

This review makes three significant contributions that advance the field of Edge AI.

First, it provides a novel historical synthesis by tracing the complete evolutionary trajectory from early distributed systems, such as content delivery networks (CDNs), to modern Edge AI paradigms, thereby establishing a critical historical continuity absent from previous surveys.

Second, it introduces a unified framework through its multi-dimensional taxonomy, which enables an integrated analysis across hardware, software, and application domains to reveal their essential interdependencies and inherent trade-offs.

Third, it offers a comprehensive challenge analysis that delivers complete coverage of the landscape, including technical constraints, deployment challenges, and fundamental performance

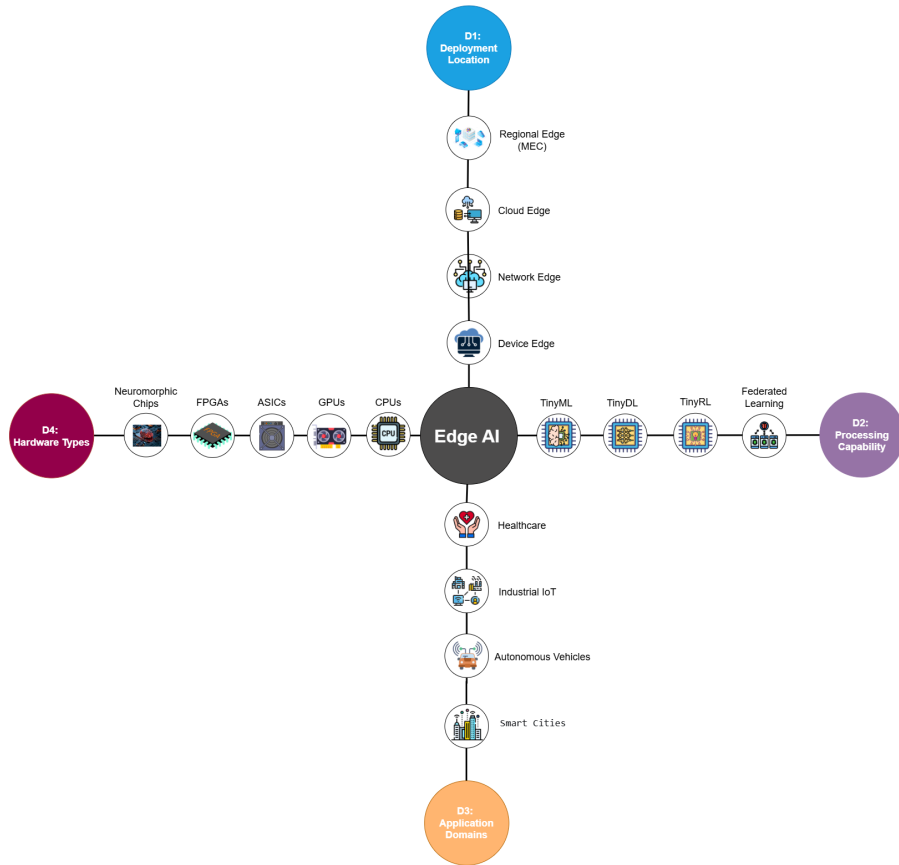


Figure 1: Multi-dimensional analytical framework for Edge AI systems, integrating deployment locations, processing capabilities, application domains, and hardware architectures.

Table 1: Comparison of Key Edge AI Surveys

Ref.	Focus	Strengths	Limitations
[23]	Architectures	<ul style="list-style-type: none"> • Broad coverage • HW/SW analysis 	<ul style="list-style-type: none"> • Shallow AI depth • No benchmarks
[29]	Algorithms	<ul style="list-style-type: none"> • Technical depth • Algorithm comparison 	<ul style="list-style-type: none"> • Dense presentation • Lacks tools
[24]	Lightweight AI	<ul style="list-style-type: none"> • Historical context • Application diversity 	<ul style="list-style-type: none"> • Edge computing bias • Weak metrics
[27]	Optimization	<ul style="list-style-type: none"> • Multi-layer taxonomy • Privacy focus 	<ul style="list-style-type: none"> • Conceptual • No benchmarks
[30]	Edge + LML	<ul style="list-style-type: none"> • Future insights • Trade-off analysis 	<ul style="list-style-type: none"> • No validation • Vague tools
[6]	Technologies	<ul style="list-style-type: none"> • Case studies • Real-time focus 	<ul style="list-style-type: none"> • Too brief • Weak analysis
[31]	Challenges	<ul style="list-style-type: none"> • Interdisciplinary • Agenda-setting 	<ul style="list-style-type: none"> • Abstract • No tools
[25]	Taxonomy	<ul style="list-style-type: none"> • Robust methodology • Collaboration focus 	<ul style="list-style-type: none"> • Surface-level • No toolchain
[26]	Optimization	<ul style="list-style-type: none"> • Model compression • HW-aware 	<ul style="list-style-type: none"> • Dense taxonomy • Narrow scope
[32]	On-Device AI	<ul style="list-style-type: none"> • Acceleration focus • Foundation models 	<ul style="list-style-type: none"> • Scalability gaps • Technical overload
[28]	Trustworthy AI	<ul style="list-style-type: none"> • Trust framework • XAI integration 	<ul style="list-style-type: none"> • Few examples • Tool gaps

trade-offs across the entire Edge AI stack. Together, these contributions provide a foundational and holistic perspective for future research.

1.5 Paper Organization

The remainder of this paper is structured as follows: Section 2 delineates the PRISMA-guided systematic methodology and multi-dimensional analytical framework. Section 3 traces the historical evolution of Edge AI from centralized cloud to distributed intelligence. Section 4 presents a taxonomic analysis of the contemporary Edge AI ecosystem. Section 5 examines the systemic challenges and fundamental trade-offs. Section 6 projects future research horizons and emerging paradigms. Finally, Section 7 concludes the review by synthesizing key insights and implications.

2 Research Methodology: A Systematic Multi-Dimensional Review

This review employs a Systematic Literature Review (SLR) methodology, conducted in strict adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [5]. This rigorous approach ensures a transparent, reproducible, and unbiased synthesis of the extant literature on Edge Artificial Intelligence (Edge AI). The process encompassed the formulation of research questions, a comprehensive search strategy, a multi-stage study selection process, systematic data extraction, and analysis through a novel analytical framework.

2.1 Research Questions

The review is guided by the following primary Research Questions (RQs), designed to comprehensively map the past, present, and future of Edge AI:

1. **RQ1:** What are the historical milestones and foundational technologies that have shaped the evolution of Edge AI?
2. **RQ2:** What constitutes the current state-of-the-art, including core technologies, architectural paradigms, and prevalent types of Edge AI (e.g., TinyML, TinyDL)?
3. **RQ3:** What are the significant application domains and their respective impacts?
4. **RQ4:** What are the predominant challenges and limitations inherent in Edge AI systems?
5. **RQ5:** What are the emerging opportunities and promising future research directions?

These questions provide a structured lens through which the vast body of literature is analyzed and synthesized.

2.2 Search Strategy

A systematic and multi-faceted search strategy was deployed to maximize the retrieval of relevant, high-quality academic literature.

2.2.1 Keywords and Search Strings

A comprehensive set of keywords was derived from the research questions and pilot searches to cover the breadth of the domain. The terms included: *"Edge AI"*, *"Edge Artificial Intelligence"*, *"Edge Computing"*, *"TinyML"*, *"Tiny Deep Learning"*, *"Federated Learning at Edge"*, *"On-Device AI"*, and *"Edge Intelligence"*, among others.

These keywords were combined using Boolean operators (AND, OR) to construct complex search queries tailored to each database's syntax. For example:

```
("Edge AI" OR "Edge Intelligence") AND ("survey" OR "review")
```

2.2.2 Data Sources and Search Period

The search was executed across eleven leading academic databases and publishers renowned for their coverage of computer science and engineering: IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect (Elsevier), arXiv, Wiley Online Library, MDPI, Taylor & Francis Online, Hindawi, Nature, and Science. An additional 36 records were identified through citation searching of relevant articles.

The initial search yielded 2,220 records. After removing 520 duplicates, a total of 1,700 records were screened by title and abstract. The search period was demarcated from **January 1, 2000, to June 30, 2025**. This timeframe was selected to capture the foundational work in edge computing and distributed systems, the emergence of key enabling technologies, and the most recent advancements in Edge AI, ensuring a complete historical contextualization.

2.3 Study Selection and Eligibility Criteria

The study selection process followed the PRISMA 2020 protocol, as detailed in the flow diagram (Figure 2).

The initial pool of 1,700 records was rigorously screened against formal inclusion and exclusion criteria (Table 2) to ensure both relevance and academic rigor. The title and abstract screening phase resulted in the exclusion of 1,490 records. The full text of the remaining 210 reports was sought for retrieval, of which 10 were not accessible. Consequently, 200 reports were thoroughly assessed for eligibility.

Of these, 121 reports were excluded for specific reasons: 76 were off-topic or lacked a primary focus on Edge AI, 25 presented no novel technical contribution, 15 were not peer-reviewed, and 5 were published in a language other than English.

This meticulous process yielded a final corpus of **79 primary studies** deemed suitable for qualitative synthesis.

2.4 Data Extraction and Synthesis

Data from the 79 included studies were extracted into a standardized template. The extracted fields included: bibliographic information (authors, title, year, source), key contributions and findings, identified challenges, and proposed future directions.

2.5 Analytical Framework: A Multi-Dimensional Taxonomy

To enable a structured and insightful synthesis that moves beyond a descriptive summary, the extracted data was analyzed through a novel multi-dimensional analytical framework, as visualized in Figure 1. This framework categorizes each contribution across four interdependent dimensions.

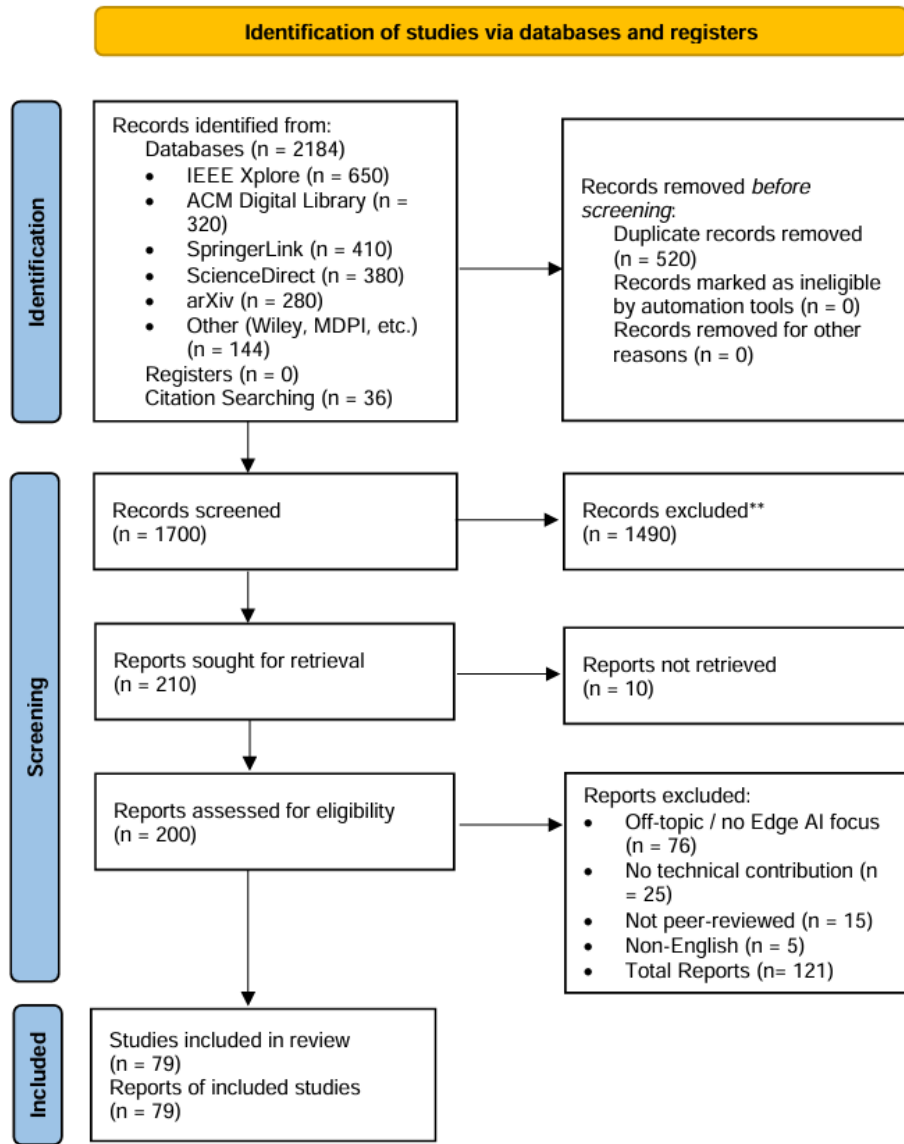


Figure 2: PRISMA 2020 flow diagram of the systematic literature identification, screening, and inclusion process.

Table 2: Study Inclusion and Exclusion Criteria

Criterion	Description
Inclusion	
Publication Type	Peer-reviewed journal articles, conference proceedings, and comprehensive survey papers.
Language	English.
Topic	Primary focus on Edge AI technologies, architectures, applications, challenges, or futures.
Time Frame	January 2000 – June 2025.
Exclusion	
Publication Type	Short papers (<4 pages), posters, abstracts, editorials, books, theses, and non-academic sources (e.g., blogs, whitepapers).
Topic	Focus solely on cloud computing or traditional data centers without an explicit edge component.
Accessibility	Full text not retrievable.

The first dimension, *deployment location (D1)*, concerns the physical or logical placement of intelligence, which ranges from the Device Edge — encompassing microcontrollers and sensors—to the Network Edge, including gateways and MEC servers, and finally the Cloud Edge.

The second dimension, *processing capability (D2)*, identifies the computational paradigm employed, spanning from ultra-constrained TinyML to more capable TinyDL and collaborative Federated Learning.

The third dimension, *application domain (D3)*, classifies the sector or use-case addressed, such as Healthcare, Industrial IoT, Autonomous Vehicles, or Smart Cities. Finally, the fourth dimension, *hardware type (D4)*, specifies the underlying processing substrate, which includes ASICs, FPGAs, GPUs, and Neuromorphic Chips.

This integrated framework facilitates a nuanced analysis of trade-offs, synergies, and research gaps across the entire ecosystem. It thereby allows for the identification of under-explored intersections, such as federated learning algorithms optimized for neuromorphic hardware. Consequently, the synthesis presented in subsequent sections is structured to critically examine the literature through this cohesive and original lens.

3 Historical Evolution: From Cloud to the Intelligent Edge

3.1 From Cloud to Edge: A Paradigm Shift

The evolution of computing paradigms represents a continuous quest for optimal efficiency, responsiveness, and scalability, progressively distributing resources closer to the point of demand. This trajectory began with centralized mainframes, evolved through client-server architectures, and culminated in the cloud computing era, which revolutionized data processing through on-demand access to immense, shared computational resources [33]. However, the explosive proliferation of connected devices and the emergence of applications requiring real-time processing of massive data volumes at the network periphery exposed critical limitations of the cloud-centric

model. Intrinsic bottlenecks related to latency, bandwidth consumption, and data privacy became fundamentally incompatible with the requirements of autonomous systems, real-time analytics, and privacy-sensitive applications [34, 35].

This analysis, guided by our methodological framework, reveals that the shift to edge-based processing was not a single event but a series of innovations, each addressing specific dimensions of the cloud limitation problem. As illustrated in Figure 3, this progression established the foundational layers that constitute the modern Edge AI landscape, directly informing the deployment location (D1) and processing capability (D2) dimensions of our taxonomy.

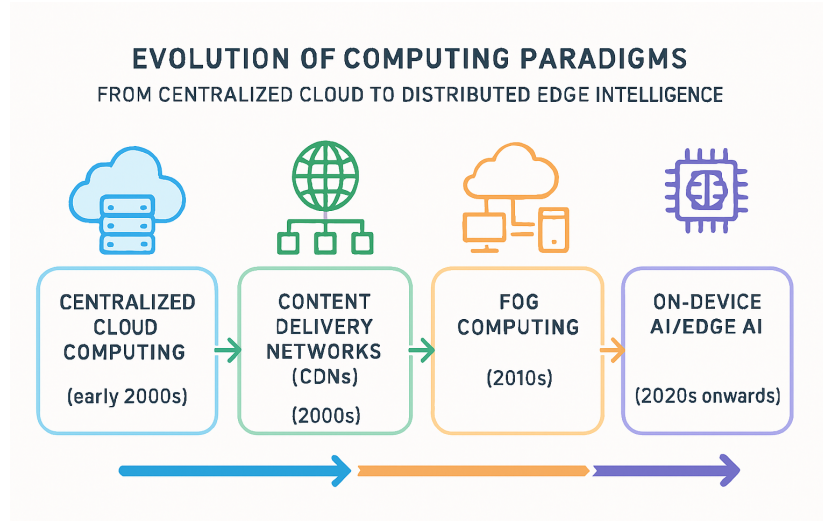


Figure 3: Evolution from centralized cloud computing to distributed Edge AI, mapping key technological milestones to the dimensions of our analytical framework.

Edge AI, therefore, represents the culmination of this evolutionary trajectory—a deliberate paradigm shift from reliance on distant cloud data centers to the strategic distribution of intelligence adjacent to data sources. This architectural decentralization minimizes long-distance data transmission, thereby critically reducing latency, conserving bandwidth, and enhancing the privacy of sensitive information through localized processing [36, 37]. The driving force behind this shift is the ubiquitous deployment of smart devices and IoT endpoints that generate continuous data streams in environments where immediate, autonomous decision-making is paramount, such as in industrial automation, autonomous vehicles, and remote healthcare monitoring [38, 39].

3.2 Foundational Technologies: The Pillars of Distributed Intelligence

The emergence of Edge AI is intrinsically linked to and built upon several foundational technologies that established the core principles of distributed processing. Our systematic review identifies three pivotal technologies that sequentially paved the way for modern Edge AI, each contributing a critical piece to the architectural puzzle.

3.2.1 Content Delivery Networks (CDNs): The Precursor to Edge Locality

Content Delivery Networks (CDNs) constituted the earliest widespread form of distributed computing, designed primarily to optimize web content delivery by geographically dispersing cached content closer to end-users [40]. While their initial focus was content replication rather than computation, CDNs introduced the seminal concept of leveraging network proximity to enhance performance and reduce congestion. This demonstrated the fundamental benefits of localized resource allocation and established the core architectural principle of bringing computation closer to the consumer, laying the essential conceptual groundwork for more sophisticated edge processing paradigms [41]. CDNs represent the initial instantiation of what would become the **Deployment Location (D1)** dimension in our taxonomy.

3.2.2 Fog Computing: Bridging the Cloud-Edge Divide

Fog computing emerged as a strategic extension of cloud computing, explicitly designed to bridge the conceptual and architectural gap between centralized cloud resources and edge devices [42]. By extending cloud services to the network edge, fog computing enabled computation, storage, and networking capabilities to be performed in closer proximity to data sources. Fog nodes—often implemented on routers, switches, or dedicated servers—functioned as intelligent intermediaries between edge devices and the cloud, providing crucial localized processing and reducing the dependency on continuous, high-bandwidth cloud connectivity [8, 43]. This architecture introduced a hierarchical computing model, a concept central to modern Edge AI, which explicitly defined different processing tiers at varying distances from the data source. Fog computing directly informs the hierarchical nature of both the **Deployment Location (D1)** and **Processing Capability (D2)** dimensions in our taxonomy.

3.2.3 Mobile Edge Computing (MEC): The Ultralow-Latency Enabler

Mobile Edge Computing (MEC), standardized as Multi-access Edge Computing, advanced the distribution paradigm by integrating cloud computing capabilities directly within the radio access network (RAN) infrastructure [7]. By positioning computation and storage resources at the base station level, MEC brings unprecedented proximity to mobile users and devices, offering ultralow latency and high bandwidth essential for advanced applications. This proximity is critical for latency-sensitive use cases such as augmented reality, virtual reality, and real-time video analytics, where millisecond-scale delays significantly impact user experience and system performance [44, 45]. MEC platforms enable application deployment at cellular base stations or access points, facilitating immediate data processing and responses, a capability that defines the high-performance end of the **Processing Capability (D2)** dimension for mobile scenarios.

3.3 The Rise of On-Device AI: The Ultimate Realization of Edge Intelligence

The convergence of these foundational technologies, coupled with breakthroughs in hardware miniaturization and energy-efficient AI algorithms, has catalyzed the most significant evolution: the rise of On-Device AI. This paradigm refers to the execution of sophisticated AI models directly on end-user devices—such as smartphones, wearables, sensors, and microcontrollers—often independent of continuous cloud or fog connectivity [46, 47].

On-Device AI represents the ultimate expression of edge intelligence, offering transformative advantages in privacy, latency, and operational autonomy. By ensuring data never leaves the device, it fundamentally mitigates privacy and security risks associated with data transmission.

The elimination of network latency ensures genuine real-time inference, and devices maintain intelligent functionality even in offline or intermittently connected environments [48, 49]. This shift has precipitated remarkable innovation in highly optimized, compact AI models, spurring not only TinyML but also the more capable Tiny Deep Learning (TinyDL) and Tiny Reinforcement Learning (TinyRL) paradigms, alongside the development of specialized hardware accelerators tailored for severely resource-constrained environments.

This historical analysis, structured through our methodological framework, demonstrates that the evolution of Edge AI has been a process of continuous refinement across the dimensions of deployment, processing, and hardware. The following section will delve into the current landscape of Edge AI, utilizing our taxonomy to provide a structured analysis of the technologies and architectures that have emerged from this evolutionary journey.

4 A Taxonomic Analysis of the Contemporary Edge AI Ecosystem

The contemporary state of Edge AI is defined by a complex, synergistic ecosystem of specialized hardware, optimized software stacks, and efficient communication protocols, all orchestrated to enable intelligent processing under stringent resource constraints. This section employs the multi-dimensional taxonomy introduced in Figure 1 to provide a structured analysis of this landscape. We deconstruct the ecosystem into its core technological pillars and then examine the dominant paradigms (TinyML, TinyDL, Federated Learning) that emerge from the interplay of these pillars, concluding with a analysis of their application across critical domains.

4.1 The Core Technology Stack: Hardware, Software, and Communication

The effective deployment of Edge AI is predicated on the co-design of hardware, software, and communication layers. This tripartite foundation directly maps to the **Hardware Type (D4)** and **Processing Capability (D2)** dimensions of our taxonomy, enabling the diverse **Deployment Locations (D1)** discussed in Section 3.

The Edge AI technology stack (Figure 4) is not a collection of isolated components but a tightly integrated hierarchy. breakthroughs in hardware accelerators (e.g., Google’s Edge TPU) create the substrate for efficient inference, which software frameworks (e.g., PyTorch Mobile) leverage through advanced optimization techniques, while communication protocols (e.g., 5G) enable the orchestration of intelligence across the edge-to-cloud continuum.

4.1.1 Hardware Accelerators for Edge AI: The Physical Substrate (D4)

The computational exigencies of contemporary artificial intelligence models, particularly deep neural networks, substantially surpass the capabilities of conventional general-purpose processors in resource-constrained edge environments. This discrepancy has precipitated the development of specialized hardware accelerators, each embodying distinct architectural trade-offs within the **Hardware Type (D4)** dimension of our taxonomic framework. These co-processors are fundamentally engineered to maximize computational efficiency, quantified as trillions of operations per second per watt (TOPS/W) [18, 19].

Four principal architectural paradigms dominate this landscape. **Application Specific Integrated Circuits (ASICs)**, epitomized by Google’s Edge TPU, represent the zenith of performance and energy efficiency for fixed-function, high-volume inference workloads; however, this

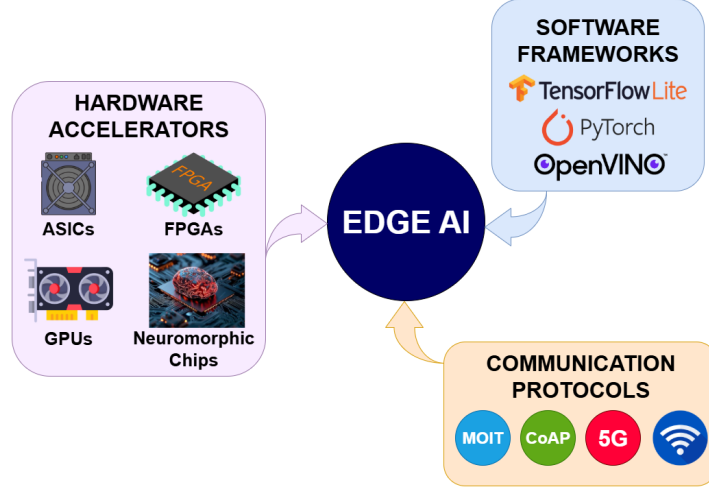


Figure 4: The Edge AI technology stack, illustrating the synergistic relationship between hardware accelerators, software frameworks, and communication protocols that enable intelligent processing across the deployment continuum.

optimization comes at the expense of architectural flexibility due to their hardened circuitry. **Field-Programmable Gate Arrays (FPGAs)**, such as the Xilinx Versal series, occupy a middle ground by offering reconfigurable logic fabrics that balance computational efficiency with post-deployment adaptability, making them particularly suitable for evolving algorithmic requirements and prototyping applications [20]. **Graphics Processing Units (GPUs)**, including power-optimized variants like the NVIDIA Jetson platform, leverage massive parallel processing capabilities to accelerate complex computational workloads such as real-time video analytics, though this often necessitates accepting higher power envelopes [21]. Finally, **neuromorphic computing platforms**, exemplified by Intel’s Loihi architecture, constitute a paradigm-shifting approach that emulates biological neural networks through event-driven, asynchronous processing, thereby offering transformative potential for ultra-low-power operation on sparse, temporal data streams [22].

This heterogeneous ecosystem of accelerator architectures demonstrates that no single solution optimally addresses all edge computing constraints, thereby necessitating careful architectural co-design across the hardware-software stack to meet specific application requirements within the Edge AI domain.

Selecting the appropriate accelerator (Table 3) is a critical system-level decision dictated by the constraints of the **Application Domain (D3)** and the target **Deployment Location (D1)**. ASICs deliver unmatched efficiency for static workloads at the network edge, while FPGAs provide crucial adaptability for industrial settings. Neuromorphic chips, though nascent, offer a disruptive potential for next-generation always-on sensing applications at the extreme device edge.

Table 3: Performance characteristics and trade-offs of dominant Edge AI hardware platforms [18, 20, 22], cataloged by the Hardware Type (D4) dimension.

Type (D4)	Example	Pros	Cons	Optimal Deployment (D1) & Use Case
ASICs	Google Edge TPU	High TOPS/W, low latency	Fixed architecture, costly design	Network/Cloud Edge; Fixed, high-volume inference
FPGAs	Xilinx Versal	Reconfigurable, energy-efficient	High development complexity	Network Edge; Evolving models, prototyping
GPUs	NVIDIA Jetson	High parallelism, flexible	Power-hungry, expensive	Device/Network Edge; Video analytics, training
Neuromorphic	Intel Loihi	Event-driven, ultra-low power	Niche programming model	Device Edge; Sensor fusion, sparse data

4.1.2 Edge AI Frameworks and Runtimes: The Software Abstraction Layer

To effectively harness the computational capabilities of heterogeneous edge hardware, a sophisticated suite of software frameworks has emerged, specializing in model optimization, quantization, and efficient inference execution. These frameworks constitute the critical software abstraction layer that operationalizes the **Processing Capability (D2)** dimension for any given **Hardware Type (D4)** within our taxonomy [50, 51].

Several prominent frameworks exemplify this technological stratum. *TensorFlow Lite* and *PyTorch Mobile* provide lightweight runtimes derived from their respective mainstream machine learning ecosystems, specifically engineered for deployment on mobile and embedded devices with stringent requirements for low latency and minimal binary footprint. The *OpenVINO Toolkit* represents a vendor-specific optimization suite that enhances deep learning model performance across Intel’s heterogeneous hardware portfolio, including CPUs, GPUs, FPGAs, and Vision Processing Units (VPUs), thereby maximizing inference efficiency on dedicated silicon architectures. In contrast, *ONNX Runtime* embodies a cross-platform inference paradigm that promotes framework interoperability by enabling models trained within one ecosystem (e.g., PyTorch) to be deployed through hardware-optimized runtimes from alternative toolchains, effectively mitigating vendor lock-in concerns.

Collectively, these frameworks provide crucial abstraction of underlying hardware complexity, enabling developers to concentrate on algorithmic innovation and model design while the software stack manages the intricate challenges of executing computational graphs efficiently across diverse accelerator architectures. This architectural separation between hardware capabilities and software implementation fundamentally enables the practical deployment of advanced AI models within the constrained environments characteristic of edge computing scenarios.

4.1.3 Communication Protocols for Edge AI: The Connectivity Fabric

Efficient and reliable data exchange is the connective tissue of distributed Edge AI systems, enabling collaborations between devices at different **Deployment Locations (D1)**. The choice of protocol is dictated by the constraints of the **Application Domain (D3)**, particularly its latency and bandwidth requirements [52, 53].

- **MQTT & CoAP:** Lightweight messaging protocols designed for constrained devices and unreliable networks, ideal for telemetry data transmission in IoT scenarios.
- **5G and Beyond:** The ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) capabilities of 5G are transformative enablers, facilitating real-time collaboration between edge devices and servers for advanced applications like autonomous driving and augmented reality [54].

4.2 Paradigms of Processing: From TinyML to Federated Learning

Edge AI is not monolithic but comprises a spectrum of paradigms tailored to specific resource constraints and processing requirements. These paradigms represent different points along the **Processing Capability (D2)** dimension, as visualized in Figure 5.

The strategic selection of a paradigm involves navigating a complex trade-off space, as systematically compared in Table 4. This decision is primarily driven by the constraints of the **Hardware Type (D4)** and the requirements of the **Application Domain (D3)**.

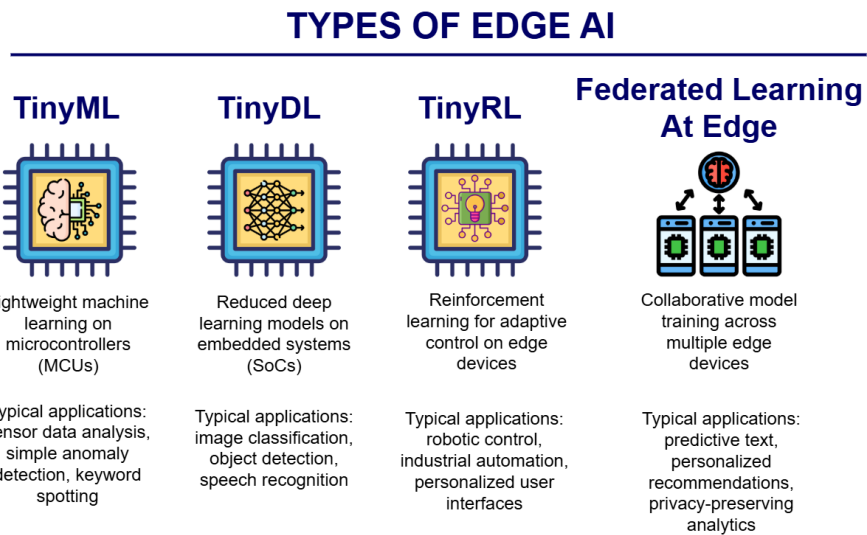


Figure 5: The spectrum of Edge AI paradigms, classified by Processing Capability (D2), ranging from ultra-constrained TinyML to collaborative Federated Learning.

Table 4: Comparative Analysis of Edge AI Paradigms along the Processing Capability (D2) dimension.

Feature	TinyML	TinyDL	TinyRL	Federated Learning at the Edge
Target Hardware (D4)	Microcontrollers	Embedded SoCs	Embedded SoCs	Heterogeneous Device Fleets
Compute Resources	KBs of RAM, μ W–mW	MBs of RAM, mW–W	MBs of RAM, mW–W	Aggregated resources
Typical Applications (D3)	Anomaly detection	Object detection	Robotic control,	Privacy-preserving analytics
Key Advantage	Ultra-low power	Efficient deep learning	Adaptive decision-making	Privacy, leverage distributed data
Key Challenge	Extreme constraints	Model optimization	Sample efficiency, policy compression	Communication overhead, heterogeneity

4.2.1 TinyML: Intelligence at the Extremes (D2)

TinyML operates at the most constrained end of the **Processing Capability (D2)** spectrum, deploying highly optimized models onto microcontrollers with kilobytes of memory and microwatt power budgets [55, 10]. Its value proposition is enabling always-on, always-sensing capabilities for applications like industrial monitoring and wearable health devices, where data privacy and power autonomy are paramount [56, 57]. The core innovation lies in extreme model compression and quantization techniques that strip neural networks down to their bare essentials without sacrificing critical functionality.

4.2.2 TinyDL: Embedded Deep Learning (D2)

TinyDL occupies the middle ground, enabling more complex deep learning tasks (e.g., image classification, object detection) on embedded systems like Raspberry Pi or NVIDIA Jetson platforms [9, 58]. These devices possess marginally more resources (MBs of RAM, watt-level power), allowing for the execution of deeper networks. The focus shifts from extreme compression to sophisticated optimization techniques like pruning, knowledge distillation, and neural architecture search to balance accuracy, latency, and power consumption for applications in smart cameras, drones, and robotics [59, 60].

4.2.3 TinyRL: Reinforcement Learning at the Edge (D2)

Tiny Reinforcement Learning (TinyRL) represents the cutting edge of on-device learning, enabling autonomous decision-making and control policies directly on resource-constrained hardware [11]. Unlike the supervised learning focus of TinyML and TinyDL, TinyRL algorithms learn optimal behaviors through interaction with their environment, making them ideal for applications in robotics, industrial control, and network management where systems must adapt in real-time without cloud dependency [61]. The primary challenge lies in compressing the high sample complexity and memory requirements of traditional RL into the extreme constraints of the device edge, often leveraging techniques like policy distillation and efficient experience replay.

4.2.4 Federated Learning: Collaborative Intelligence

Federated Learning (FL) is a unique paradigm that transcends a single **Deployment Location (D1)**. It is a privacy-preserving distributed training method that leverages the collective computational power of a heterogeneous fleet of edge devices (**D4**) [62, 13]. Instead of centralizing raw data, FL trains models locally on each device and aggregates only model updates. This makes it particularly suited for **Application Domains (D3)** with stringent privacy concerns, such as healthcare and finance, though it introduces challenges in communication efficiency and handling device heterogeneity [63, 64].

4.2.5 Other Variants: Deployment Location (D1) Specialization

Beyond the core computational paradigms previously examined, the Edge AI ecosystem exhibits specialized architectural implementations distinguished primarily by their positioning within the **Deployment Location (D1)** dimension. Two particularly significant specializations merit explicit consideration.

First, **Edge AI on Gateways/Servers** (including Regional Edge/MEC nodes) represents an approach that deploys more capable models (TinyDL, TinyRL) on robust appliances, enabling

sophisticated processing like real-time multi-sensor fusion in smart manufacturing and urban infrastructure [65, 66].

Second, **Edge AI in Mobile Devices** constitutes a distinct paradigm that harnesses dedicated neural processing units (NPUs) integrated within smartphones to facilitate advanced on-device applications including augmented reality interfaces and computational photography enhancements [67, 12].

These deployment-specific implementations demonstrate that the physical and logical placement of computational resources serves as a fundamental architectural determinant that directly governs system capabilities, operational constraints, and performance characteristics across diverse Edge AI applications. The strategic selection of deployment location thus represents a critical design consideration that profoundly influences both the technical feasibility and practical efficacy of Edge AI solutions within their intended operational environments.

Table 5: Deployment Tiers of Edge AI, synthesizing Hardware Type (D4), resource constraints, and typical Application Domains (D3).

Category	Device Examples (D4)	Examples	Memory	Power	Typical Use Cases (D3)
TinyML	Micro-controllers		KBs	μ W–mW	Keyword spotting, vibration monitoring
TinyDL	Embedded SoCs		MBs	mW–W	Real-time object detection, drones
Edge Servers	Gateways, MEC nodes		GBs	W–kW	Smart city analytics, multi-sensor fusion

The stratification of the Edge AI landscape into distinct deployment tiers (Table 5) is a direct consequence of the trade-offs analyzed through our taxonomic framework. The choice of tier dictates the feasible paradigms and ultimately the applications that can be successfully deployed.

4.3 Application Domains: Transformative Impact Across Industries

Edge AI is fundamentally transforming industries by enabling intelligent, real-time decision-making at the data source. Its value proposition—local processing, reduced latency, and enhanced privacy—makes it indispensable across diverse sectors. The following analysis, structured by the **Application Domain (D3)** dimension, highlights the most impactful use cases, with their benefits summarized in Table 6.

4.3.1 Smart Homes and Cities

In smart homes, Edge AI enhances privacy and responsiveness. Devices like smart speakers, security cameras, and thermostats can process voice commands, detect intruders, or optimize energy consumption locally without sending sensitive data to the cloud [68, 69]. This on-device processing ensures immediate responses and reduces concerns about data breaches. For instance, a smart doorbell with Edge AI can identify known visitors or detect suspicious activity in real-time, sending alerts only when necessary [70].

In smart cities, Edge AI plays a crucial role in managing urban infrastructure and services. Applications include intelligent traffic management systems that optimize signal timings based

Table 6: Edge AI Applications categorized by Application Domain (D3), highlighting domain-specific benefits.

Domain (D3)	Application Examples	Key Benefits of Edge AI
Smart Cities	Traffic management, public safety	Real-time processing, low-latency decision-making
Industrial IoT	Predictive maintenance, quality control	On-device analytics, minimized downtime
Autonomous Vehicles	Object detection, navigation	Ultra-low latency, enhanced safety, independence
Healthcare	Remote patient monitoring, diagnostics	Strong data privacy, real-time insights
Retail	Inventory tracking, customer analytics	Real-time insights, improved user experience

on real-time traffic flow detected by edge cameras [17], smart streetlights that adjust illumination based on pedestrian and vehicle presence [71], and waste management systems that optimize collection routes using AI-powered sensors on bins [72]. These deployments leverage Edge AI to improve efficiency, sustainability, and public safety.

4.3.2 Industrial IoT (IIoT)

Edge AI is a cornerstone of Industry 4.0, enabling predictive maintenance, quality control, and operational optimization in manufacturing and industrial settings. By deploying AI models directly on factory equipment, sensors can monitor machine health, detect anomalies, and predict potential failures before they occur, minimizing downtime and reducing maintenance costs [73, 15]. For example, vibration sensors with embedded AI can analyze machine vibrations to identify early signs of wear and tear, triggering alerts for proactive maintenance [74].

Furthermore, Edge AI facilitates real-time quality inspection on production lines, where cameras with embedded AI can identify defects in products with high accuracy and speed, ensuring consistent product quality [75]. This localized processing is critical in environments where network latency to the cloud could lead to significant production delays or errors.

4.3.3 Autonomous Vehicles

Autonomous vehicles are perhaps one of the most demanding applications for Edge AI, requiring ultra-low latency and highly reliable real-time decision-making. Self-driving cars must process vast amounts of sensor data (from cameras, LiDAR, radar, etc.) instantaneously to perceive their environment, predict the behavior of other road users, and make critical navigation decisions [76, 77]. Cloud-based processing for such tasks is impractical due to the inherent latency, which could lead to catastrophic delays. Edge AI enables these vehicles to operate autonomously and safely by performing complex AI computations directly on board [16]. This includes object detection, lane keeping assistance, pedestrian recognition, and real-time path planning, all executed at the edge to ensure immediate responses to dynamic road conditions [78].

4.3.4 Healthcare

In healthcare, Edge AI offers transformative potential, particularly in remote patient monitoring, diagnostics, and personalized medicine. Wearable health devices equipped with Edge AI can continuously monitor vital signs, detect anomalies, and alert patients or healthcare providers to critical changes, all while preserving patient privacy by processing sensitive data on-device [79, 14]. For instance, an Edge AI-powered ECG device can detect arrhythmias in real-time, providing immediate feedback to the user or triggering emergency services [80].

Edge AI also supports intelligent medical imaging analysis at the point of care, allowing for faster preliminary diagnoses in remote clinics or emergency settings without relying on cloud connectivity [81]. This can significantly reduce the time to diagnosis and improve patient outcomes, especially in underserved areas.

4.3.5 Retail

Edge AI is revolutionizing the retail sector by enhancing customer experience, optimizing store operations, and improving inventory management. In smart retail environments, Edge AI-powered cameras can analyze customer traffic patterns, optimize product placement, and detect shoplifting in real-time, without transmitting continuous video feeds to the cloud [82, 83]. This not only improves security but also provides valuable insights into customer behavior while maintaining privacy.

Furthermore, Edge AI can facilitate automated checkout systems, smart shelves that monitor inventory levels, and personalized advertising displays that adapt to customer demographics or preferences, all processed locally to ensure immediate and relevant interactions [84, 85]. These applications demonstrate how Edge AI can create more efficient, secure, and customer-centric retail experiences.

5 Systemic Challenges and Fundamental Trade-Offs

Despite its transformative potential, the widespread deployment of Edge AI is contingent upon overcoming significant and interconnected challenges. These limitations are not merely technical hurdles but fundamental trade-offs inherent to the distributed, resource-constrained, and heterogeneous nature of edge environments. As illustrated in Figure 6, these five core challenges—resource constraints, data privacy, model management, power consumption, and connectivity—and their interconnections span the entire stack, from hardware and algorithms to security and systems management. This section analyzes these barriers through the lens of our multi-dimensional taxonomy, examining how constraints in one dimension (e.g., Hardware Type - D4) precipitate challenges in another (e.g., Processing Capability - D2 or Deployment - D1).

5.1 Resource Constraints: The Fundamental Trade-Off

The most fundamental challenge in Edge AI stems from the severe resource constraints intrinsic to devices at the **Device Edge (D1)**, particularly those running **TinyML (D2)**. Furthermore, emerging paradigms like TinyRL introduce additional complexity, as their training-inference cycles and memory requirements for experience replay present novel optimization hurdles beyond those of static inference or even federated learning. These constraints—encompassing computational power, memory (RAM and storage), and energy budgets—define the design space and force a continuous trade-off between model accuracy, latency, inference speed, and power consumption [86, 87].

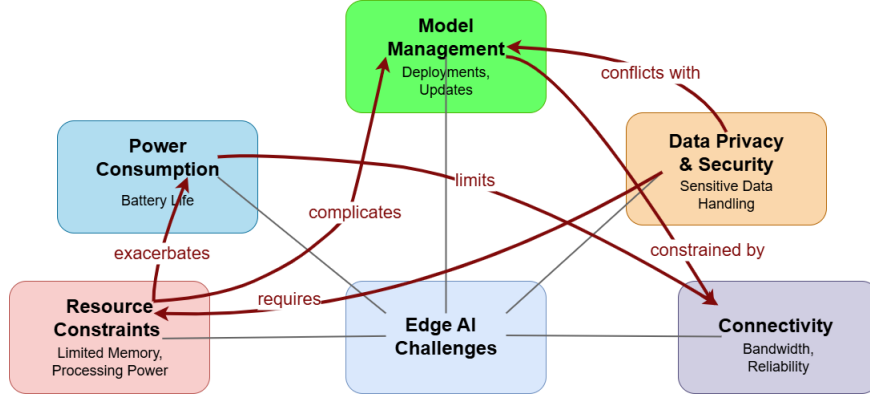


Figure 6: Systemic challenges in Edge AI, depicting five core challenges (resource constraints, data privacy & security, model management, power consumption, and connectivity) and their causal interrelationships (e.g., "exacerbates", "requires") across the technology stack.

While techniques like model compression (pruning, quantization), efficient neural architectures (e.g., MobileNets), and specialized **hardware accelerators (D4)** are crucial mitigations, they are not panaceas. For instance, aggressive quantization can achieve 3–5× energy savings but often at the cost of non-negligible accuracy loss, especially for complex models [88, 89]. Consequently, developing robust AI capabilities for devices with kilobyte-scale memory and milliwatt power budgets remains a formidable engineering challenge that dictates the feasible **Application Domains (D3)** for a given hardware class.

5.2 Data Privacy and Security: The Expanded Attack Surface

While Edge AI inherently enhances privacy by processing data locally, reducing its exposure over networks, it simultaneously introduces a new set of security vulnerabilities by distributing the attack surface. Physically exposed **edge devices (D1)** are susceptible to tampering, theft, and side-channel attacks, threatening the integrity of both the AI models and the sensitive data they process [90, 91].

The distributed nature of Edge AI complicates centralized security management. Ensuring trust across a heterogeneous fleet of devices—from microcontrollers to edge servers—requires robust mechanisms like secure boot, trusted execution environments (TEEs), and homomorphic encryption. Furthermore, paradigms like **Federated Learning (D2)** are being adopted as a privacy-preserving training method, but they themselves introduce new challenges related to the security of aggregated model updates [92, 93]. A comprehensive, standardized security framework for these heterogeneous ecosystems is still nascent, representing a critical gap for sensitive **Application Domains (D3)** like healthcare and industrial control.

5.3 Model Management and Deployment: The Operational Bottleneck

The lifecycle management of AI models across vast, geographically dispersed edge deployments presents a profound operational challenge. The core of this problem is the extreme **hardware and software heterogeneity (D4)** across the edge continuum. Deploying, updating, and maintaining models on millions of devices, each with potentially different capabilities, requires sophisticated orchestration platforms that are both robust and secure [94, 95].

Over-the-air (OTA) updates must be efficient and fault-tolerant, especially for devices with intermittent connectivity. Strategies such as A/B partitioning and rollback capabilities are essential to ensure reliability. Furthermore, maintaining model consistency and version control across this diverse fleet, while simultaneously debugging issues in the field, adds significant complexity to the DevOps cycle for Edge AI, potentially stalling the deployment of new applications [96].

5.4 Power Consumption: The Energy Efficiency Frontier

Energy efficiency represents arguably the paramount challenge for battery-operated or energy-harvesting devices operating at the **Device Edge (D1)**. The fundamental tension between the objective of continuous, always-on sensing and inference capabilities and the requirement for multi-year operational lifespans necessitates innovative solutions. Although specialized low-power **accelerators (D4)** provide partial mitigation, a comprehensive, system-wide approach to power management remains essential, requiring optimization across every system component from sensors to communication modules [97, 98].

Current research converges on three synergistic strategies to advance the energy-efficiency frontier. First, **hardware-level optimization** employs quantized models—utilizing 8-bit integers rather than floating-point representations—to achieve 3–5× reductions in memory bandwidth and computational energy consumption. This approach is complemented by dedicated low-power cores that handle simple always-on tasks at microwatt power levels, thereby maintaining main processors in deep sleep states for extended durations [88, 97]. Second, **algorithmic efficiency** techniques, including pruning and sparsity exploitation, eliminate redundant operations to reduce energy consumption by 30–60%. Event-triggered inference further enhances efficiency by radically reducing the duty cycle, processing data exclusively in response to meaningful sensor events such as detected motion [89, 98]. Third, **system-wide power gating** implements techniques including dynamic voltage and frequency scaling (DVFS) to adjust computational resources in real-time, while selective peripheral disabling powers down unused sensors and radios between inference cycles [99, 100].

Significant challenges persist, particularly regarding the minimization of accuracy penalties associated with extreme quantization and the development of standardized power management interfaces capable of operating across heterogeneous **hardware platforms (D4)**. The most promising trajectory forward appears to lie in hybrid approaches that combine hardware-software co-design with adaptive, learning-based algorithms [98, 99].

5.5 Connectivity: The Reliability Bottleneck

The efficacy of distributed Edge AI systems is fundamentally dependent on reliable and sufficient network connectivity, encompassing both bandwidth and reliability [53]. Unlike cloud computing, the edge continuum often operates in environments with unstable or low-bandwidth connections, such as rural areas, moving vehicles, or dense industrial settings. This variability directly complicates model management (e.g., failed OTA updates) and constrains the feasibility of collaborative paradigms like Federated Learning, which require frequent model update

exchanges. Furthermore, strategies to overcome poor connectivity, such as data buffering or more complex compression algorithms, can exacerbate power consumption challenges, creating a critical trade-off between operational reliability and energy efficiency.

5.6 Interoperability and Standardization: The Integration Challenge

The fragmented nature of the Edge AI ecosystem presents a critical barrier to widespread adoption. The proliferation of proprietary hardware accelerators (D4), diverse software frameworks, and incompatible communication protocols frequently results in vendor lock-in, increased development complexity, and substantial challenges in integrating disparate components into cohesive, scalable systems [101, 102].

This absence of common standards significantly impedes innovation and elevates development costs. While industry consortia and open-source initiatives are actively working to establish common APIs—such as the Open Neural Network Exchange (ONNX)—standardized data formats, and consistent deployment methodologies, the remarkable diversity of the edge landscape renders widespread adoption a long-term objective [103, 104]. Achieving genuine interoperability remains crucial for unlocking the full potential of Edge AI technologies and enabling seamless collaboration between devices operating across different **Deployment Locations (D1)**, as required by many advanced **Applications (D3)**.

5.7 Synthesis of Challenges

The challenges facing Edge AI are not isolated but are deeply interconnected. The drive for greater energy efficiency (**Power**) can exacerbate security vulnerabilities. Unreliable **Connectivity** complicates model management and can constrain processing capabilities. The resource constraints of a chosen **Hardware (D4)** platform directly limit the complexity of models that can be deployed, affecting the achievable **Processing Capability (D2)** and thus the feasible **Applications (D3)**. Navigating this complex web of trade-offs is the central task for Edge AI system architects, requiring a holistic approach informed by the multi-dimensional perspective provided by our taxonomic framework.

6 Future Research Horizons and Emerging Paradigms

The evolution of Edge AI is accelerating, driven by the imperative to overcome current limitations and unlock new frontiers of decentralized intelligence. The future landscape will be characterized by unprecedented hardware specialization, algorithms capable of autonomous adaptation, and the seamless, trust-aware orchestration of intelligence across the edge-to-cloud continuum. This section projects these future trajectories, framing them as natural progressions within the multi-dimensional taxonomy that structures this review. These emerging opportunities, summarized in Figure 7, promise to address the challenges outlined in Section 5 and radically expand the application horizon of Edge AI.

6.1 Next-Generation Hardware: Redefining the Performance-Energy Frontier (D4)

The relentless pursuit of enhanced energy efficiency and computational density is driving a paradigm shift within the **Hardware Type (D4)** dimension, heralding a transition from incremental improvements to fundamentally novel computing architectures. Future Edge AI hardware

FUTURE DIRECTIONS AND OPPORTUNITIES IN EDGE AI

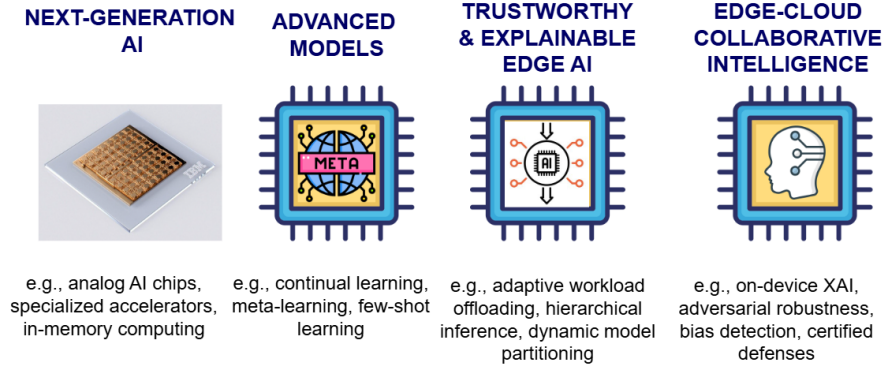


Figure 7: Emerging research vectors in Edge AI, spanning next-generation hardware paradigms, advanced algorithmic capabilities, and systemic architectures for collaborative intelligence.

development will be characterized by four transformative trajectories that collectively redefine the performance-energy frontier.

First, the emergence of **hyper-specialized accelerators** marks a departure from general-purpose AI chips toward processors meticulously optimized for specific model architectures—including transformers and graph neural networks—and specialized data modalities such as sparse and event-based data. This architectural evolution promises order-of-magnitude efficiency gains for niche Application Domains (D3), particularly in real-time sensor fusion and edge-based natural language understanding [105, 106].

Second, **in-memory and near-memory computing (IMC/NMC)** architectures represent a fundamental rethinking of computational paradigms by executing operations directly within or adjacent to memory cells through emerging technologies including memristors and magnetoresistive random-access memory (MRAM). This approach substantially mitigates the von Neumann bottleneck, dramatically reducing data movement energy and enabling unprecedented ultra-low-power, high-throughput inference capabilities essential for data-intensive applications at the Device Edge (D1) [107, 108].

Third, **post-digital computing paradigms** are transitioning from theoretical constructs to practical implementations, with analog AI and optical computing offering transformative potential. Analog compute-in-memory techniques, leveraging fundamental physical principles including Ohm’s and Kirchhoff’s laws for matrix operations, demonstrate potential for 10–100× energy reduction in always-on processing scenarios, particularly for sparse data applications in industrial monitoring and environmental sensing [109, 110].

Finally, the development of **self-powered intelligent systems** embodies the convergence of advanced energy harvesting technologies—spanning solar, kinetic, and radio frequency harvesting—with ultra-low-power AI processors. This integration enables perpetually operational, maintenance-free devices that fundamentally expand viable Deployment Locations (D1) to include remote, inaccessible, and hazardous environments, thereby transcending conventional power constraints in intelligent sensing applications [111, 112].

Collectively, these innovations represent not merely incremental advances but rather a fun-

damental reimagining of computational approaches that will enable previously impossible Edge AI applications while addressing critical energy constraints.

6.2 Advanced Algorithms: Towards Adaptive and Resource-Aware Intelligence (D2)

Algorithmic research is poised to fundamentally transform Edge AI systems by imbuing them with greater autonomy, efficiency, and contextual understanding, thereby substantially advancing the capabilities within the **Processing Capability (D2)** dimension. This evolution will manifest through several critical research directions that collectively address the unique constraints and opportunities of edge computing environments.

A primary focus will be the development of **continual and lifelong learning systems**, which will progressively replace static, pre-deployed models with architectures capable of incremental learning from continuous data streams. Such systems will enable models to adapt to concept drift and personalize to local environments without suffering catastrophic forgetting, thereby drastically reducing the need for costly retraining and redeployment cycles [113, 114].

Concurrently, **meta-learning and few-shot learning** approaches will emerge as crucial enablers for agile deployment in diverse and dynamic scenarios where labeled data is inherently scarce. These techniques will facilitate "plug-and-play" intelligence, allowing pre-optimized models to rapidly specialize for new sensors or environmental conditions at the edge [115, 116].

The deployment of **Tiny Reinforcement Learning (TinyRL) agents** directly on edge devices will represent another significant advancement, enabling autonomous, real-time decision-making and control loops. This capability proves particularly transformative for Application Domains (D3) such as robotics, industrial automation, and network management, where systems must learn and react to complex environments without the latency of cloud dependency [11, 61].

Finally, as Edge AI penetrates increasingly critical applications, research will focus on developing **explainable AI (XAI)** techniques specifically designed for **resource-constrained environments**. These methods will provide interpretable rationales for model decisions without overwhelming the limited computational and memory resources of edge hardware, thereby addressing growing demands for transparency and accountability in autonomous systems [117, 118].

Together, these algorithmic advancements will create a new generation of adaptive, resource-aware intelligence systems capable of operating effectively within the stringent constraints of edge computing environments while maintaining sophisticated learning and decision-making capabilities.

6.3 Edge-Cloud Collaborative Intelligence: The Emergence of a Cognitive Continuum

The conventional rigid dichotomy between edge and cloud computing is evolving toward a fluid, hierarchical intelligence continuum, representing the maturation of the **Deployment Location (D1)** dimension into a dynamic, integrated system. This architectural shift will be characterized by several defining features that collectively enable more efficient and responsive intelligent systems.

Future architectures will be defined by **adaptive hierarchical systems** that dynamically distribute intelligence across computational tiers. In this paradigm, TinyML models will perform initial filtering and time-critical inference at the Device Edge (D1), while more complex analytical tasks—such as multi-sensor fusion and long-term trend analysis—will be offloaded to Network Edge servers. The cloud will increasingly specialize in large-scale model training and

global aggregation, thereby establishing a seamless flow of computation and data across the entire continuum [119, 120].

Complementing this architectural evolution, **intelligent dynamic offloading** mechanisms will employ AI-powered controllers to make real-time decisions about workload placement based on a comprehensive assessment of network conditions, computational load, energy availability, and application requirements. This approach will systematically optimize the complex trade-offs between latency, bandwidth, accuracy, and power consumption across the entire system [121, 122].

Furthermore, **enhanced federated learning frameworks** will advance beyond simple update averaging to incorporate stronger privacy guarantees—including differential privacy and homomorphic encryption—along with robust aggregation algorithms capable of handling extreme non-IID data and device heterogeneity. These frameworks will also integrate mechanisms for detecting and mitigating malicious participants, thereby creating a truly scalable and secure solution for privacy-sensitive domains [123, 124].

This collaborative paradigm, illustrated in Figure 8, envisions a future where embedded TinyDL models in smart cameras perform real-time object detection, edge servers aggregate multiple feeds for sophisticated crowd analytics, and cloud-based federated learning systems securely improve global models—all functioning as a single, cohesive intelligent system that dynamically adapts to changing conditions and requirements.

6.4 Trustworthy and Explainable Edge AI: The Foundation for Adoption

For Edge AI systems to achieve deployment in truly critical applications, trust must be elevated to a first-class design constraint, systematically integrated across both the **Processing Capability (D2)** and **Application Domain (D3)** dimensions. This imperative will drive several critical research thrusts that collectively address the unique challenges of trustworthy computing in resource-constrained environments.

A primary research focus will center on developing **robustness against adversarial attacks** through defense mechanisms specifically designed for the computational constraints of edge hardware. This includes advancing efficient adversarial training techniques, input purification methods, and runtime detection algorithms that can secure Edge AI systems against evolving threat landscapes without compromising operational efficiency [125, 126].

Equally important will be addressing the unique challenges of **bias detection and mitigation** in models trained on decentralized, non-IID edge data. Research must develop specialized techniques for detecting, quantifying, and mitigating bias directly on edge devices or during federated learning processes, ensuring equitable and ethical AI outcomes across diverse deployment scenarios [127, 128].

Finally, achieving practical **on-device explainability** will require methods capable of generating concise, meaningful explanations for model decisions directly on edge hardware, utilizing only a fraction of the resources required for inference itself. This capability will prove paramount for establishing user trust, ensuring regulatory compliance, and enabling effective developer debugging in field deployments [129, 130].

Together, these research directions will establish the foundational trustworthiness necessary for Edge AI systems to expand into increasingly critical applications while maintaining the efficiency requirements inherent to edge computing environments.

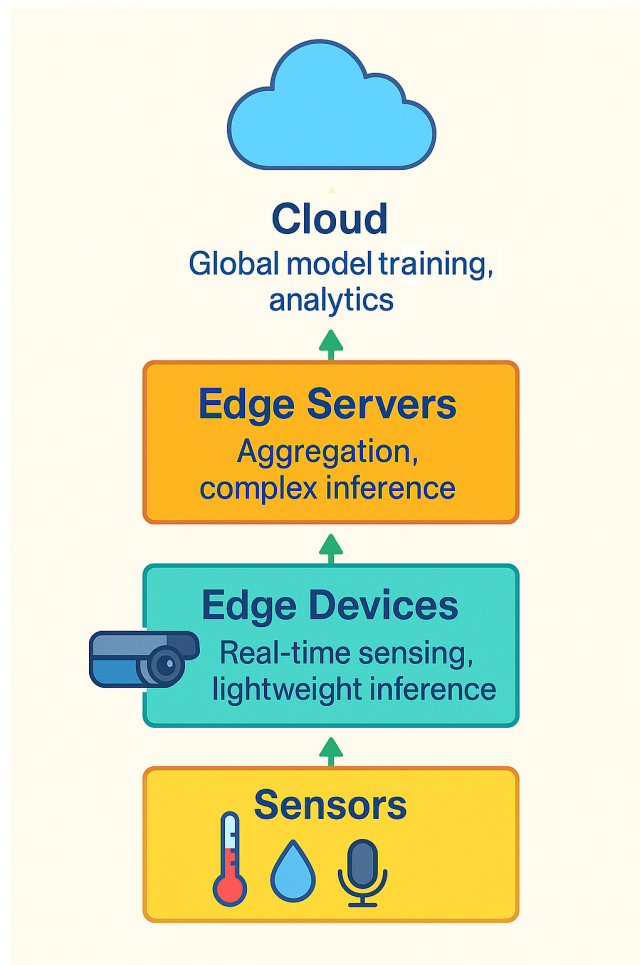


Figure 8: The future cognitive continuum: A dynamic, hierarchical architecture that intelligently partitions and orchestrates AI workloads across the device-edge-cloud spectrum based on real-time constraints and requirements.

6.5 The Role of 6G and Advanced Networking: The Connective Tissue

Next-generation wireless networks, particularly 6G, will transcend their conventional role as mere data conduits to become the intelligent connective tissue that unifies the edge computing continuum, thereby directly enabling transformative **Application Domains (D3)**. This evolution will manifest through several foundational capabilities that collectively redefine the relationship between networking and distributed intelligence.

6G networks are anticipated to provide **sub-millisecond latency and holographic-type communication capabilities**, which will fundamentally unlock applications requiring unprecedented responsiveness. These advancements will enable collaborative swarm robotics, autonomous vehicle platooning, and immersive extended reality (XR) experiences where haptic feedback and real-time interaction demand near-instantaneous communication [131, 132].

Furthermore, the **deep integration of artificial intelligence and communication** technologies will transform the network infrastructure itself into an inherently intelligent system. Through the embedding of AI capabilities directly within the radio access network (RAN), future networks will dynamically allocate resources, predict network states, and optimize performance parameters specifically tailored to Edge AI workload requirements [133, 134].

Additionally, the emergence of **sensing-as-a-service** capabilities will represent a paradigm shift in network functionality. 6G infrastructure is projected to incorporate integrated sensing technologies, effectively transforming the network into a distributed sensor array that can provide rich perceptual context to Edge AI devices, thereby significantly enhancing their situational awareness and operational capabilities across diverse environments.

Collectively, these advancements will establish next-generation networks as active enablers of Edge AI systems rather than passive communication channels, creating a symbiotic relationship between networking capabilities and distributed intelligence that will drive innovation across the entire computing continuum.

6.6 Synthesis: Towards a Cognitive Edge

The convergence of these directions points to a future of a "Cognitive Edge"—an intelligent, self-adapting, and trustworthy fabric of distributed computation. This evolution will be characterized by a shift from deploying static models on isolated devices to orchestrating dynamic intelligence across a continuum of heterogeneous resources. The multi-dimensional taxonomy presented in this review provides the essential framework for navigating this complex and exciting future, outlining the inter-dependencies between hardware, algorithms, deployment strategies, and applications that will define the next decade of Edge AI innovation.

7 Conclusion

Edge Artificial Intelligence represents a paradigm shift in computational architecture, fundamentally redefining how intelligent systems are designed, deployed, and integrated into the physical world. This comprehensive review has systematically charted the evolution, current state, and future trajectory of Edge AI, offering a holistic analysis through a novel multi-dimensional taxonomy that integrates deployment location, processing capability, application domain, and hardware type.

Our analysis reveals that the journey from cloud-centric computing to distributed edge intelligence has been neither accidental nor instantaneous. It is the result of a coherent evolution through foundational technologies—CDNs, fog computing, and mobile edge comput-

ing—each solving critical limitations of its predecessor and collectively paving the way for modern paradigms like TinyML, TinyDL, TinyRL, and federated learning. This historical contextualization, often neglected in prior surveys, provides essential perspective for understanding current developments and future directions.

Through our systematic methodology and taxonomic framework, we have demonstrated that the contemporary Edge AI landscape is characterized by sophisticated hardware-software co-design across a spectrum of resource constraints. We have further shown how these technological capabilities enable transformative applications across diverse sectors—from healthcare and industrial automation to autonomous systems and smart cities—each with unique requirements for latency, privacy, and autonomy.

However, this review also identifies significant challenges that constrain widespread adoption. Resource constraints, security vulnerabilities, model management complexities, and power consumption limitations present formidable hurdles that require continued innovation. These challenges are not isolated but interconnected, demanding holistic solutions that address trade-offs across hardware, software, and deployment architectures.

Looking forward, we project that Edge AI’s future lies in several key directions: (1) next-generation hardware paradigms that redefine energy-performance trade-offs through in-memory computing, analog AI, and specialized accelerators; (2) advanced algorithms capable of continuous adaptation, few-shot learning, and explainable decision-making within resource constraints; (3) seamless edge-cloud collaborative intelligence that dynamically distributes workloads across the computational continuum; and (4) the integration of trustworthiness and explainability as fundamental design principles rather than afterthoughts.

The realization of Edge AI’s full potential will require unprecedented interdisciplinary collaboration across hardware engineering, computer systems, algorithm design, and application domains. As 5G/6G networks mature and AI workloads become increasingly pervasive, the principles and architectures discussed in this review will become central to next-generation intelligent systems.

Ultimately, Edge AI is poised to create a future where artificial intelligence becomes truly ubiquitous—embedded not just in devices but woven into the very fabric of our environment, enabling responsive, intelligent, and autonomous systems that operate seamlessly within our physical world while respecting the constraints of resources, privacy, and energy. This survey provides a comprehensive foundation for researchers, practitioners, and policymakers to navigate and contribute to this rapidly evolving field.

A Complete Edge AI Reference Taxonomy

Table 7: Systematic classification of seminal Edge AI literature (2017–2025) organized by the review’s taxonomic categories: hardware accelerators, TinyML/TinyDL/TinyRL paradigms, federated learning, edge systems (Fog/MEC/CDN), application domains, and emerging challenges.

Year	Category	Subcategory	Reference	Key Contribution
2020	Hardware	ASIC	[19]	Google Edge TPU architecture analysis
2021	Hardware	GPU	[21]	NVIDIA Jetson performance profiling
2022	Hardware	FPGA	[20]	Real-time SRAM-based acceleration
2023	Hardware	Neuromorphic	[22]	Intel Loihi2 edge deployment
2024	Hardware	CiM	[107]	Memristor-based compute-in-memory
2024	Hardware	Survey	[18]	Comparative analysis of 32 accelerators
2025	Hardware	Analog	[109]	Mythic analog AI chip case study
2017	TinyML	DL	[59]	First just-in-time DL compilation
2020	TinyML	Tools	[56]	TensorFlow Lite Micro framework
2022	TinyML	MCU	[55]	ARM Cortex-M4 optimizations
2023	TinyML	Vision	[60]	Neural architecture search for MCUs
2024	TinyML	Survey	[10]	State-of-the-art techniques review
2023	TinyDL	Survey	[9]	From TinyML to TinyDL: A comprehensive survey
2023	TinyDL	Architecture	[58]	Analysis of deep learning on microcontrollers
2024	TinyDL	Optimization	[60]	Hardware-aware NAS for TinyDL models
2024	TinyRL	Algorithms	[11]	Design principles for RL on edge devices
2023	TinyRL	Applications	[61]	Diffusion-based RL for edge-generated content
2019	Federated	Foundational	[12]	First edge FL framework
2021	Federated	Privacy	[63]	Differential privacy enhancements
2022	Federated	Survey	[13]	Analysis of 58 deployments
2023	Federated	Robustness	[92]	Adversarial attack defenses

Continued on next page

Year	Category	Subcategory	Reference	Key Contribution
2018	Edge	Fog	[8]	Fog computing architecture
2020	Edge	MEC	[7]	5G MEC standardization
2021	Edge	CDN	[40]	AI-enhanced content delivery
2022	Edge	Survey	[42]	10-year evolution analysis
2024	Edge	MEC	[44]	MEC for video streaming and VR
2020	Apps	Healthcare	[14]	Wearable ECG monitoring
2021	Apps	Automotive	[16]	Real-time object detection
2022	Apps	Industrial IoT	[15]	Resource-efficient Edge AI for predictive maintenance
2022	Apps	Industry	[15]	Predictive maintenance systems
2019	Challenges	Privacy	[37]	Edge data protection framework
2021	Challenges	Power	[97]	Energy harvesting techniques
2022	Challenges	Security	[90]	Attack vector taxonomy
2022	Future	Continual	[113]	Lifelong learning algorithms
2023	Future	Bio	[111]	Neuromorphic edge systems
2025	Future	6G	[132]	AI-optimized RAN architectures

References

- [1] S. Vasuki. Edge AI: A comprehensive survey of technologies, applications, and challenges. In *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pages 1–6, 2024.
- [2] T. Sipola, J. Alatalo, T. Kokkonen, and M. Rantonen. Artificial intelligence in the IoT era: A review of edge AI hardware and software. In *2022 31st Conference of Open Innovations Association (FRUCT)*, pages 320–331, 2022.
- [3] A. Karras, A. Giannaros, C. Karras, K. C. Giotopoulos, D. Tsolis, and S. Sioutas. Edge artificial intelligence in large-scale IoT systems, applications, and big data infrastructures. In *2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–8, 2023.
- [4] M. Sibanda, E. Bhero, and J. Agee. AI edge processing - a review of distributed embedded systems. In *2023 31st Southern African Universities Power Engineering Conference (SAUPEC)*, pages 1–6, 2023.
- [5] Matthew J Page et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews*, 10(1):1–11, 2021.
- [6] Vikram Shankar. Edge ai: A comprehensive survey of technologies, applications, and challenges. In *Proceedings of the 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pages 1–6, 2024.
- [7] S. Ahmed, H. Khalid, M. Hamza, and D. Farhat. Mobile edge computing. *arXiv [cs.DC]*, 2024.
- [8] H. Gupta and A. K. Bharti. Fog computing & IoT: Overview, architecture and applications. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(5):30–34, May 2018.
- [9] S. Somvanshi et al. From tiny machine learning to tiny deep learning: A survey. *arXiv [cs.LG]*, 2025.
- [10] J. Lin, L. Zhu, W.-M. Chen, W.-C. Wang, and S. Han. Tiny machine learning: Progress and futures [feature]. *IEEE Circuits Syst. Mag.*, 23(3):8–34, 2023.
- [11] G. Wu, D. Zhang, Z. Miao, W. Bao, and J. Cao. How to design reinforcement learning methods for the edge: An integrated approach toward intelligent decision making. *Electronics (Basel)*, 13(7):1281, 2024.
- [12] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Netw.*, 33(5):156–165, 2019.
- [13] H. G. Abreha, M. Hayajneh, and M. A. Serhani. Federated learning in edge computing: A systematic survey. *Sensors (Basel)*, 22(2):450, 2022.
- [14] A. Rocha et al. Edge AI for internet of medical things: A literature review. *Comput. Electr. Eng.*, 116(109202):109202, 2024.
- [15] V. Artiushenko, S. Lang, C. Lerez, T. Reggelin, and M. Hackert-Oschätzchen. Resource-efficient edge AI solution for predictive maintenance. *Procedia Comput. Sci.*, 232:348–357, 2024.
- [16] J. Xie, X. Zhou, and L. Cheng. Edge computing for real-time decision making in autonomous driving: Review of challenges, solutions, and future trends. *Int. J. Adv. Comput. Sci. Appl.*, 15(7), 2024.

- [17] G. P. Sharma. Real-time traffic management using IoT sensors and edge computing in smart cities. *International Journal of Trend in Research and Development*, 11(6):96–100, Dec 2024.
- [18] S. Alam, C. Yakopcic, Q. Wu, M. Barnell, S. Khan, and T. M. Taha. Survey of deep learning accelerators for edge and emerging computing. *Electronics (Basel)*, 13(15):2988, 2024.
- [19] B. Liang. Design of ASIC accelerators for AI applications. In *International Conference on Electrical Engineering and Intelligent Control (EEIC 2024)*, pages 147–154, 2024.
- [20] H. Liu et al. A high-performance accelerator for real-time super-resolution on edge FPGAs. *ACM Trans. Des. Automat. Electron. Syst.*, 29(3):1–25, 2024.
- [21] H. Bouzidi, H. Ouarnoughi, S. Niar, and A. A. E. Cadi. Performance modeling of computer vision-based CNN on edge GPUs. *ACM Trans. Embed. Comput. Syst.*, 21(5):1–33, 2022.
- [22] R. S. Das. Emerging neuromorphic computing for edge AI application: A systematic literature review. *Journal of Technological Innovations*, 5(1):1–8, 2024.
- [23] Jorge Mendez et al. Edge intelligence: Concepts, architectures, applications, and future directions. *ACM Transactions on Embedded Computing Systems (TECS)*, 21(5):1–36, 2022.
- [24] Rajdeep Singh and Sukhpal Singh Gill. Edge ai: A survey. *IoT and Cyber-Physical Systems*, 3:1–20, 2023.
- [25] Sukhpal Singh Gill et al. Edge ai: A taxonomy, systematic review and future directions. *Cluster Computing*, 28(1):1–25, 2025.
- [26] Xiaolong Wang and Wenlong Jia. Optimizing edge ai: A survey on data, model, and system strategies. *Qeios*, 2025.
- [27] Chellammal Surianarayanan et al. A survey on optimization techniques for edge ai. *Sensors*, 23(3):1279, 2023.
- [28] Xiaolong Wang et al. A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability. *IEEE Communications Surveys & Tutorials*, 27(3):2102–2138, 2025.
- [29] Wei Su et al. Ai on the edge: A comprehensive review. *Artificial Intelligence Review*, 55(8):6125–6184, 2022.
- [30] Kyle Hoffpauir et al. A survey on edge intelligence and lightweight ml. *ACM Journal of Data and Information Quality (JDIQ)*, 15(2), 2023.
- [31] Tobias Meuser et al. Revisiting edge ai: Opportunities and challenges. *IEEE Internet Computing*, 28(4):45–53, 2024.
- [32] Xiaolong Wang et al. Empowering edge intelligence: A comprehensive survey on on-device ai models. *ACM Computing Surveys*, 57(9):1–39, 2025.
- [33] J. H. Kiswani, S. M. Dascalu, and A. F. C. Jr Harris. Cloud computing and its applications: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(1):3–24, Mar 2021.
- [34] B. R. Cherukuri. Edge computing vs. cloud computing: A comparative analysis for real-time AI applications. *IEEE Access*, 6(5):1–17, Oct 2024.
- [35] B. Charyyev, E. Arslan, and M. H. Gunes. Latency comparison of cloud datacenters and edge servers. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–6, 2021.

- [36] M G Avram. Advantages and challenges of adopting cloud computing from an enterprise perspective. *Procedia Technol.*, 12:529–534, 2014.
- [37] Panjun Sun. Security and privacy protection in cloud computing: Discussions and challenges. *J. Netw. Comput. Appl.*, 160(102642):102642, 2020.
- [38] A. Choudhary. Internet of things: a comprehensive overview, architectures, applications, simulation tools, challenges and future directions. *Discover Internet of Things*, 4(1):1–41, Dec 2024.
- [39] J. Wang, M. K. Lim, C. Wang, and M.-L. Tseng. The evolution of the internet of things (IoT) over the past 20 years. *Computers & Industrial Engineering*, 155:107174, May 2021.
- [40] B. Zolfaghari et al. Content delivery networks: State of the art, trends, and future roadmap. *ACM Comput. Surv.*, 53(2):1–34, 2021.
- [41] Vagmi and R. K. Gupta. Content delivery networks in the modern age: Analyzing trends, overcoming challenges, and pioneering developments. In *Smart Innovation, Systems and Technologies*, pages 793–806. Springer Nature Singapore, Singapore, 2024.
- [42] S. N. Srirama. A decade of research in fog computing: Relevance, challenges, and future directions. *arXiv [cs.DC]*, 2023.
- [43] S. H. Han and V. Naik. A review on fog computing: Architecture, fog with IoT, algorithms and research challenges. *ICT Express*, 7(2):162–176, 2021.
- [44] M. A. Khan et al. A survey on mobile edge computing for video streaming: Opportunities and challenges. *arXiv [cs.MM]*, 2022.
- [45] Z. Li et al. Optimizing mobile edge computing for virtual reality rendering via UAVs: A multi-agent deep reinforcement learning approach. *IEEE Internet Things J.*, pages 1–1, 2025.
- [46] J. J. Moon et al. A new frontier of AI: On-device AI training and personalization. *arXiv [cs.LG]*, 2022.
- [47] X. Wang et al. Empowering edge intelligence: A comprehensive survey on on-device AI models. *ACM Comput. Surv.*, 57(9):1–39, 2025.
- [48] S. Ubale. On-device AI models: Advancing privacy-first machine learning for mobile applications. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, 11(1):61–68, 2025.
- [49] D. Xu et al. Fast on-device LLM inference with NPUs. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pages 445–462, 2025.
- [50] H. Rexha and S. Lafond. Data collection and utilization framework for edge AI applications. *arXiv [cs.LG]*, 2021.
- [51] R. Dagli and S. Eken. Deploying a smart queuing system on edge with intel OpenVINO toolkit. *Soft Comput.*, 25(15):10103–10115, 2021.
- [52] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief. Communication-efficient edge AI: Algorithms and systems. *arXiv [cs.IT]*, 2020.
- [53] C. Mwase, Y. Jin, T. Westerlund, H. Tenhunen, and Z. Zou. Communication-efficient distributed AI strategies for the IoT edge. *Future Gener. Comput. Syst.*, 131:292–308, 2022.
- [54] E. Kartsakli et al. AI-powered edge computing evolution for beyond 5g communication networks. In *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 478–483, 2023.

- [55] R. Immonen and T. Hämäläinen. Tiny machine learning for resource-constrained micro-controllers. *J. Sens.*, 2022:1–11, 2022.
- [56] P. P. Ray. A review on TinyML: State-of-the-art and prospects. *J. King Saud Univ. - Comput. Inf. Sci.*, 34(4):1595–1623, 2022.
- [57] H. Han and J. Siebert. TinyML: A systematic review and synthesis of existing research. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, 2022.
- [58] M. Roveri. Is tiny deep learning the new deep learning? In *Computational Intelligence and Data Analytics*, pages 23–39. Springer Nature Singapore, Singapore, 2023.
- [59] B. Darvish Rouhani, A. Mirhoseini, and F. Koushanfar. TinyDL: Just-in-time deep learning solution for constrained embedded systems. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017.
- [60] A. Burrello, M. Risso, B. A. Motetti, E. Macii, L. Benini, and D. J. Pagliari. Enhancing neural architecture search with multiple hardware constraints for deep learning model deployment on tiny IoT devices. *IEEE Trans. Emerg. Top. Comput.*, 12(3):780–794, 2024.
- [61] H. Du et al. Diffusion-based reinforcement learning for edge-enabled AI-generated content services. *arXiv [cs.NI]*, 2023.
- [62] S. G. Thomas and P. K. Myakala. Beyond the cloud: Federated learning and edge AI for the next decade. *J. Comput. Commun.*, 13(02):37–50, 2025.
- [63] A. Brecko, E. Kajati, J. Koziorek, and I. Zolotova. Federated learning for edge computing: A survey. *Appl. Sci. (Basel)*, 12(18):9124, 2022.
- [64] A. Hemmati, H. M. Arzanagh, and A. M. Rahmani. Fundamentals of edge AI and federated learning. In *Model Optimization Methods for Efficient and Edge AI*, pages 1–23. Wiley, 2025.
- [65] T. Wang, J. Guo, B. Zhang, G. Yang, and D. Li. Deploying AI on edge: Advancement and challenges in edge intelligence. *Mathematics*, 13(11):1878, 2025.
- [66] L. H. Nguyen, K. D. Tran, X. Zeng, and K. P. Tran. Human-centered edge artificial intelligence for smart factory applications in industry 5.0: A review and perspective. In *Artificial Intelligence for Safety and Reliability Engineering*, pages 79–100. Springer Nature Switzerland, Cham, 2024.
- [67] M. Hirsch, C. Mateos, and T. A. Majchrzak. Exploring smartphone-based edge AI inferences using real testbeds. *Sensors (Basel)*, 25(9), 2025.
- [68] P. Thakur, S. Goel, and E. Puthooran. Edge AI enabled IoT framework for secure smart home infrastructure. *Procedia Comput. Sci.*, 235:3369–3378, 2024.
- [69] A. M. Sheikh, M. R. Islam, M. H. Habaebi, S. A. Zabidi, A. R. Bin Najeeb, and A. Kabani. A survey on edge computing (EC) security challenges: Classification, threats, and mitigation strategies. *Future Internet*, 17(4):175, 2025.
- [70] V. Patel, S. Kanani, T. Pathak, P. Patel, M. I. Ali, and J. Breslin. A demonstration of smart doorbell design using federated deep learning. *arXiv [cs.DC]*, 2020.
- [71] A. Omar et al. Smart city: Recent advances in intelligent street lighting systems based on IoT. *J. Sens.*, 2022:1–10, 2022.
- [72] A. Choubey, S. Mishra, R. Misra, A. K. Pandey, and D. Pandey. Smart e-waste management: a revolutionary incentive-driven IoT solution with LPWAN and edge-AI integration for environmental sustainability. *Environ. Monit. Assess.*, 196(8):720, 2024.

- [73] J. I. Argungu et al. A survey of edge computing approaches in smart factory. *Nternational J. Adv. Res. Comput. Commun. Eng.*, 12(9), 2023.
- [74] D. Ji, J. Y. Kim, H. W. Kim, and Y. Park. MEMS vibration sensor-based edge AI for machinery fault prediction: feasibility study using a petrochemical plant process simulation facility. *J. Incl. Phenom. Macrocycl. Chem.*, 105(3–4):249–259, 2025.
- [75] S. Mandapaka, C. Diaz, H. Irisson, A. Akundi, V. Lopez, and D. Timmer. Application of automated quality control in smart factories - a deep learning-based approach. In *2023 IEEE International Systems Conference (SysCon)*, pages 1–8, 2023.
- [76] M. Rahmati. Edge AI-powered real-time decision-making for autonomous vehicles in adverse weather conditions. *arXiv [cs.RO]*, 2025.
- [77] S. Manivannan, V. Muralidharan, S. Kumar, B. Sundarambal, Kirubakaran, and C. Selvaganesan. Edge intelligence in autonomous vehicle navigation. In *2025 International Conference on Data Science and Business Systems (ICDSBS)*, pages 1–7, 2025.
- [78] I. Ahmed, M. Ahmad, M. U. R. Siddiqi, A. Chehri, and G. Jeon. Toward AI-powered edge intelligence for object detection in self-driving cars: Enhancing IoV efficiency and safety. *IEEE Internet Things J.*, 12(11):16990–16997, 2025.
- [79] A. Sathiya, D. Angel, M. Iswarya, R. Poonkodi, K. M. Angelo, and N. Priyadharshini. IoT enabled healthcare framework using edge AI and advanced wearable sensors for real time health monitoring. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pages 384–392, 2025.
- [80] Z. Huang et al. Efficient edge-AI models for robust ECG abnormality detection on resource-constrained hardware. *J. Cardiovasc. Transl. Res.*, 17(4):879–892, 2024.
- [81] Y. Xu, T. M. Khan, Y. Song, and E. Meijering. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artif. Intell. Rev.*, 58(3), 2025.
- [82] A. Biswas, A. Jain, and Mohana. Survey on edge computing–key technology in retail industry. In *Computer Networks and Inventive Communication Technologies*, pages 97–106. Springer Nature Singapore, Singapore, 2021.
- [83] N. Rashvand, G. A. Noghre, A. D. Pazho, S. Yao, and H. Tabkhi. Exploring pose-based anomaly detection for retail security: A real-world shoplifting dataset and benchmark. *arXiv [cs.CV]*, 2025.
- [84] A. Savit and A. Damor. Revolutionizing retail stores with computer vision and edge AI: A novel shelf management system. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 69–74, 2023.
- [85] R. Islam, M. S. Arafat, M. A. M. Jony, S. M. S. Rafi, M. S. Jalil, and F. Hossen. Hyper-personalization with AI and edge computing: The future of customer experience in e-commerce and retail. *Advanced International Journal of Multidisciplinary Research*, 2(6):1–18, Dec 2024.
- [86] R. Dong, Y. Mao, and J. Zhang. Resource-constrained edge AI with early exit prediction. *J. Commun. Inf. Netw.*, 7(2):122–134, 2022.
- [87] Z. Jia et al. The importance of resource awareness in artificial intelligence for healthcare. *Nat. Mach. Intell.*, 5(7):687–698, 2023.
- [88] E. J. Husom et al. Sustainable LLM inference for edge AI: Evaluating quantized LLMs for energy efficiency, output accuracy, and inference latency. *arXiv [cs.CY]*, 2025.
- [89] A. C. Muhoza, E. Bergeret, C. Brdys, and F. Gary. Power consumption reduction for IoT devices thanks to edge-AI: Application to human activity recognition. *Internet Things (Amst.)*, 24(100930):100930, 2023.

- [90] A. Shafee, S. R. Hasan, and T. A. Awaad. Privacy and security vulnerabilities in edge intelligence: An analysis and countermeasures. *Comput. Electr. Eng.*, 123(110146):110146, 2025.
- [91] A. Shafee, T. A. Awaad, and A. Moro. A survey of edge computing privacy and security threats and their countermeasures. In *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 484–489, 2024.
- [92] E. Villar-Rodriguez, M. A. Pérez, A. I. Torre-Bastida, C. R. Senderos, and J. López-de Armentia. Edge intelligence secure frameworks: Current state and future challenges. *Comput. Secur.*, 130(103278):103278, 2023.
- [93] R. Jayanth, N. Gupta, and V. Prasanna. Benchmarking edge AI platforms for high-performance ML inference. *arXiv [cs.AI]*, 2024.
- [94] S. Choudhary, Vijitha, D. D. Bhavani, Bhuvaneswari, M. Tiwari, and Subburam. Edge AI deploying artificial intelligence models on edge devices for real-time analytics. *ITM Web Conf.*, 76:01009, 2025.
- [95] J. V. Anchitaalagammai, S. Kavitha, R. Buurvidha, T. S. Santhiya, M. D. Roopa, and S. S. Sankari. Edge artificial intelligence for real-time decision making using NVIDIA jetson orin, google coral edge TPU and 6g for privacy and scalability. In *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*, pages 150–155, 2025.
- [96] A. L. Baresi and D. F. Mendonca. Towards a serverless platform for edge computing. In *2019 IEEE International Conference on Fog Computing (ICFC)*, pages 1–10, 2019.
- [97] S. Soro. TinyML for ubiquitous edge AI. *arXiv [cs.LG]*, 2021.
- [98] M. Lee, X. She, B. Chakraborty, S. Dash, B. Mudassar, and S. Mukhopadhyay. Reliable edge intelligence in unreliable environment. In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 896–901, 2021.
- [99] Y. Meng, Z. Xudong, Z. Jianwen, X. Xinxin, W. Changling, and W. Fang. A ultra-low power system design method of AI edge computation. In *2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1–5, 2023.
- [100] D. Katare, M. Zhou, Y. Chen, M. Janssen, and A. Y. Ding. Energy-aware vision model partitioning for edge AI. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 671–678, 2025.
- [101] D. M. K. Dave and B. K. Mittapally. Data integration and interoperability in IOT: Challenges, strategies and future direction. *International Journal of Computer Engineering and Technology*, 15(1):45–60, Feb 2024.
- [102] A. Stanko, O. Duda, A. Mykytyshyn, O. Totosko, and R. Koroliuk. Artificial intelligence of things (AIoT): Integration challenges, and security issues. *Bioinformatics and Applied Information Technologies (BAIT'2024)*, 3842:1–14, 2024.
- [103] 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.*, 40(1):5–36, 2022.
- [104] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc. IEEE Inst. Electr. Electron. Eng.*, 107(8):1738–1762, 2019.
- [105] W. Li and M. Liewig. A survey of AI accelerators for edge environment. In *Trends and Innovations in Information Systems and Technologies*, pages 35–44. Springer International Publishing, Cham, 2020.

- [106] J. Haris, R. Saha, W. Hu, and J. Cano. SECDA-LLM: Designing efficient LLM accelerators for edge devices. *arXiv [cs.AR]*, 2024.
- [107] W.-S. Khwa et al. A mixed-precision memristor and SRAM compute-in-memory AI processor. *Nature*, 639(8055):617–623, 2025.
- [108] T.-H. Wen et al. Fusion of memristor and digital compute-in-memory processing for energy-efficient edge computing. *Science*, 384(6693):325–332, 2024.
- [109] D. Fick. Analog compute-in-memory for AI edge inference. In *2022 International Electron Devices Meeting (IEDM)*, pages 21.8.1–21.8.4, 2022.
- [110] D. Pile. Optical computing and artificial intelligence. *Nat. Photonics*, 2024.
- [111] A. R. Trivedi, J. Kung, and J. H. Ko. Architectures for self-powered edge intelligence. In *Handbook of Computer Architecture*, pages 89–125. Springer Nature Singapore, Singapore, 2025.
- [112] M. Ben Ammar, I. Ben Dhaou, D. El Houssaini, S. Sahnoun, A. Fakhfakh, and O. Kannon. Requirements for energy-harvesting-driven edge devices using task-offloading approaches. *Electronics (Basel)*, 11(3):383, 2022.
- [113] A. Soltoggio et al. A collective AI via lifelong learning and sharing at the edge. *Nat. Mach. Intell.*, 6(3):251–264, 2024.
- [114] H. Wang, S. Lin, and J. Zhang. *Continual and reinforcement learning for edge AI: Framework, foundation, and algorithm design*. Springer Nature Switzerland, Cham, 2025.
- [115] Y. Chen, H. Yu, Q. Guo, S. Zhao, and T. Taleb. Dynamic edge AI service management and adaptation via off-policy meta-reinforcement learning and digital twin. In *ICC 2024 - IEEE International Conference on Communications*, volume 80, pages 867–872, 2024.
- [116] J. Lu. Few-shot learning on edge devices using CLIP: A resource-efficient approach for image classification. *Information Technology and Control*, 53(3):833–845, Jun 2024.
- [117] A. E. Hassanien, D. Gupta, A. K. Singh, and A. Garg, editors. *Explainable edge AI: A futuristic computing perspective*. Springer International Publishing, Cham, Switzerland, 1st edition, 2023.
- [118] H. T. T. Nguyen, L. P. T. Nguyen, and H. Cao. XEdgeAI: A human-centered industrial inspection framework with data-centric explainable edge AI approach. *Inf. Fusion*, 116(102782):102782, 2025.
- [119] N. S. Dhakad, Y. Malhotra, S. K. Vishvakarma, and K. Roy. SHA-CNN: Scalable hierarchical aware convolutional neural network for edge AI. *arXiv [cs.NE]*, 2024.
- [120] F. Firouzi, B. Farahani, and A. Marinšek. The convergence and interplay of edge, fog, and cloud in the AI-driven internet of things (IoT). *Inf. Syst.*, 107(101840):101840, 2022.
- [121] S. Kumari, M. Rakshith, C. K. S. Sibi, and T. Sayyed. Leveraging artificial intelligence for dynamic workload management in edge and cloud environments. *International Journal for Research Trends and Innovation*, 9(7):309–316, Jul 2024.
- [122] Y. Li, S. Cheng, H. Zhang, and J. Liu. Dynamic adaptive workload offloading strategy in mobile edge computing networks. *Comput. Netw.*, 233(109878):109878, 2023.
- [123] J. Rane, O. Kaya, S. K. Mallick, and N. L. Rane. Federated learning for edge artificial intelligence: Enhancing security, robustness, privacy, personalization, and blockchain integration in IoT. In *Future Research Opportunities for Artificial Intelligence in Industry 4.0 and 5.0*. Deep Science Publishing, 2024.

- [124] K. Wang, Q. He, F. Chen, H. Jin, and Y. Yang. FedEdge: Accelerating edge-assisted federated learning. In *Proceedings of the ACM Web Conference 2023*, pages 2895–2904, 2023.
- [125] D. Zhong, B. Li, X. Chen, and C. Liu. EdgeShield: A universal and efficient edge computing framework for robust AI. *arXiv [cs.CR]*, 2024.
- [126] Z. Yi, Y. Qian, M. Chen, S. A. Alqahtani, and M. S. Hossain. Defending edge computing based metaverse AI against adversarial attacks. *Ad Hoc Netw.*, 150(103263):103263, 2023.
- [127] W. Hutiri. *Design patterns for detecting and mitigating bias in edge AI*. PhD thesis, Delft University of Technology, Delft, Netherlands, 2023.
- [128] D. Katare, N. Kourtellis, S. Park, D. Perino, M. Janssen, and A. Y. Ding. Bias detection and generalization in AI algorithms on edge for autonomous driving. In *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*, pages 342–348, 2022.
- [129] B. Xu et al. Towards explainability for AI-based edge wireless signal automatic modulation classification. *J. Cloud Comput. Adv. Syst. Appl.*, 13(1), 2024.
- [130] J. Xu et al. On-device language models: A comprehensive review. *arXiv [cs.CL]*, 2024.
- [131] P. Zhang and H. Zhu. The fusion of edge computing and artificial intelligence in 5g communication. In *Proceedings of the 2024 3rd International Conference on Frontiers of Artificial Intelligence and Machine Learning*, pages 106–112, 2024.
- [132] Y. Zhou and X. Chen. Edge intelligence: Edge computing for 5g and the internet of things. *Future Internet*, 17(3):101, 2025.
- [133] P. Zhang, D. Wen, G. Zhu, Q. Chen, K. Han, and Y. Shi. Collaborative edge AI inference over cloud-RAN. *IEEE Trans. Commun.*, 72(9):5641–5656, 2024.
- [134] R. Barker. Advancements in mobile edge computing and open RAN: Leveraging artificial intelligence and machine learning for wireless systems. *arXiv [cs.NI]*, 2025.