International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 15, No. 1, March 2026, pp. 1~20
ISSN: 2089-4864, DOI: 10.11591/ijres.v15.i1.pp1-20

Energy-efficient reconfigurable architectures for Edge AI in healthcare IoT: trends, challenges, and future directions

Tole Sutikno[1], Aiman Zakwan Jidin[2], Lina Handayani[3]
[1]Master Program of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]Faculty of Electronics and Computer Technology and Engineering, Universiti Teknikal Malaysia Melaka, Malacca, Malaysia
[3]Faculty of Public Health, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Article Info

Article history:

Received Oct 13, 2025
Revised Jan 18, 2026
Accepted Jan 24, 2026

Keywords:

Edge artificial intelligence
Embedded systems
Energy-efficient computing
Health monitoring systems
Healthcare applications
Internet of things
Reconfigurable architectures

ABSTRACT

The integration of Edge artificial intelligence (AI) with internet of things (IoT) technologies is transforming healthcare applications, including wearable monitoring, telemedicine, and implantable medical devices, by enabling low-latency and intelligent data processing close to patients. However, stringent requirements on energy efficiency, reliability, real-time responsiveness, and data privacy continue to hinder scalable and long-term deployment in resource-constrained healthcare environments. Energy-efficient reconfigurable architectures—such as field-programmable gate arrays (FPGAs), coarse-grained reconfigurable arrays (CGRAs), and emerging memory-centric and heterogeneous platforms—have emerged as promising solutions to address these challenges by balancing flexibility, adaptability, and power efficiency. This review systematically examines recent advances in reconfigurable Edge AI architectures for healthcare IoT, highlighting key trends in hardware–software co-design, AI-assisted design automation, memory-centric optimization, and domain-specific overlays. It further identifies critical challenges, including energy–performance trade-offs, runtime reconfiguration overheads, security and privacy vulnerabilities, limited standardization, and reliability concerns in dynamic clinical settings. Finally, future research directions are outlined, emphasizing self-optimizing and context-aware architectures, secure and trustworthy reconfiguration mechanisms, unified frameworks for heterogeneous healthcare workloads, and sustainable, carbon-aware edge computing. Collectively, this review positions energy-efficient reconfigurable architectures as a foundational enabler for next-generation Edge AI in IoT-enabled healthcare systems.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Tole Sutikno
Master Program of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan
UAD 4th Campus, South Ring Road, Tamanan, Banguntapan, Bantul, Yogyakarta 55166, Indonesia
Email: tole@te.uad.ac.id

1. INTRODUCTION

In recent years, the convergence of the internet of things (IoT) and artificial intelligence (AI) at the network edge has gained significant momentum, driven by the growing demand for intelligent, real-time decision-making in resource-constrained environments. This paradigm shift is particularly impactful in IoT-enabled healthcare applications, where continuous monitoring, low-latency inference, data privacy, and long-term autonomous operation are essential [1]. Wearable devices, remote patient monitoring systems, and smart medical sensors increasingly rely on Edge AI to process sensitive health data locally, reducing latency and dependence on cloud infrastructure [2]–[4]. However, the ambition to deploy AI pervasively at the edge is

fundamentally constrained by energy efficiency, which remains a critical bottleneck for scalable and sustainable healthcare systems.

Reconfigurable architectures have emerged as a pivotal solution to address these challenges, offering a balance between flexibility, adaptability, and energy efficiency. Platforms such as field-programmable gate arrays (FPGAs), coarse-grained reconfigurable arrays (CGRAs), and emerging fabrics including embedded FPGAs and in-memory accelerators enable hardware customization tailored to application-specific workloads. This capability is particularly valuable in healthcare scenarios, where workloads vary widely—from biosignal processing and medical image analysis to activity recognition and anomaly detection—while operating under strict energy and reliability constraints. This review aims to explore state-of-the-art advances in energy-efficient reconfigurable architectures for Edge AI, examine key challenges, and outline future research directions, with a specific focus on IoT-enabled healthcare applications. Figure 1 illustrates a layered IoT architecture, highlighting the integration of edge intelligence with reconfigurable computing elements within healthcare-oriented systems [5], [6].

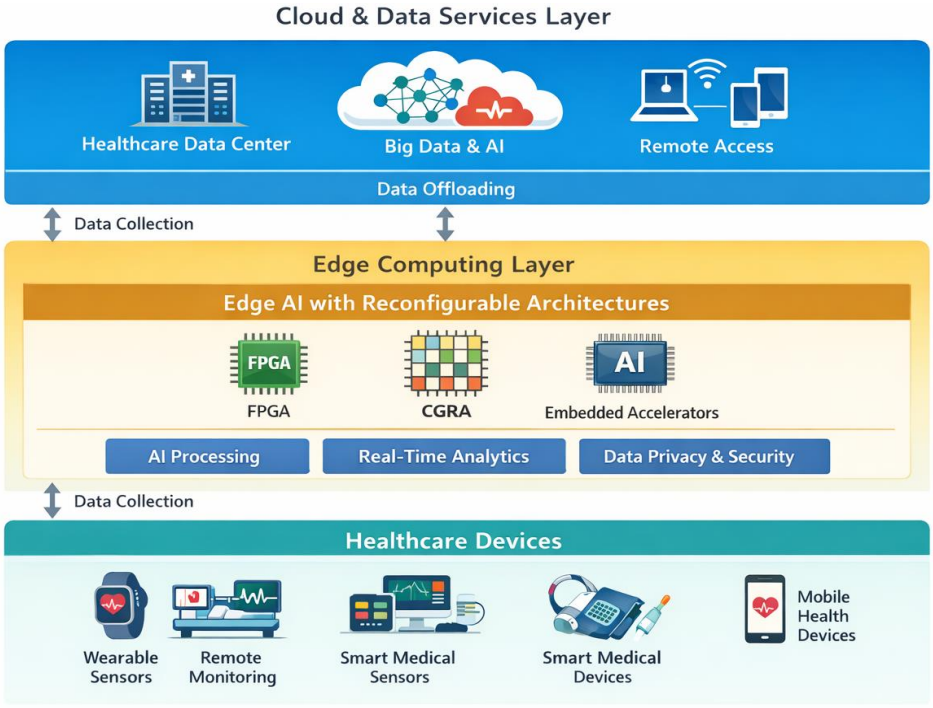

Figure 1. Layered architecture for IoT-enabled healthcare applications

The convergence of IoT and AI at the edge represents a fundamental shift in computing paradigms, driven by the need for low-latency processing, enhanced privacy, and efficient utilization of limited resources. As the number of connected devices continues to grow exponentially, especially in healthcare environments, centralized cloud-based processing becomes increasingly impractical due to latency, bandwidth, and security concerns. Edge AI addresses these limitations by enabling local intelligence; however, it simultaneously intensifies constraints on power consumption and computational efficiency. Reconfigurable architectures play a crucial role in this transformation by providing adaptable hardware platforms capable of meeting diverse and evolving workload requirements. Their flexibility allows designers to tailor computational resources to specific healthcare tasks, enabling efficient execution of AI models under stringent energy budgets. This review examines recent advancements in energy-efficient reconfigurable architectures while addressing persistent challenges such as security vulnerabilities, interoperability issues, and the lack of standardization across IoT ecosystems. Furthermore, it outlines future directions—including self-optimizing architectures and secure reconfiguration techniques—that are essential for enabling sustainable Edge AI deployments in healthcare and other critical application domains [7], [8].

The convergence of the IoT and AI at the edge signifies a pivotal shift in technological paradigms, driven by the demand for instantaneous data processing and efficient resource utilisation. As the volume of connected devices proliferates, energy efficiency emerges as a critical constraint, necessitating innovative

approaches to architecture design that balance performance and power consumption. Reconfigurable architectures, characterised by their adaptability and flexibility, serve as a catalyst for this transformation, allowing for tailored solutions that meet the diverse requirements of edge environments. This review elucidates cutting-edge advancements in energy-efficient reconfigurable architectures while addressing ongoing challenges, including security vulnerabilities and the lack of standardisation within IoT frameworks. By identifying future directions such as self-optimising architectures and secure reconfiguration methods, the work outlines a roadmap for sustainable advancements in Edge AI applications across sectors like smart cities and healthcare, encapsulated in alongside comprehensive analyses of AIs role in sustainable practices [7], [8] to inform ongoing research.

Energy efficiency is a defining challenge at the intersection of IoT and Edge AI, particularly in healthcare applications that require continuous operation over extended periods [9]. Wearable and implantable medical devices, remote health monitoring systems, and mobile diagnostic tools often rely on limited battery capacity or energy harvesting, making power consumption a primary design constraint. The bottleneck arises not only from the computational demands of AI workloads but also from the inflexibility of conventional hardware platforms, which are poorly suited to adapt dynamically to varying performance and energy requirements. Reconfigurable architectures—most notably FPGAs and CGRAs—offer a promising pathway to overcome these limitations by enabling fine-grained control over resource allocation, parallelism, and power management. Through techniques such as dynamic partial reconfiguration and workload-aware customization, these platforms can achieve significant energy savings while maintaining performance. Nevertheless, challenges remain, including the absence of standardized development frameworks, increased design complexity, and security risks associated with dynamic reconfiguration. Addressing these issues is essential to unlock the full potential of energy-efficient Edge AI in IoT-enabled healthcare systems [3].

As Edge AI applications continue to evolve, the flexibility and adaptability of reconfigurable architectures become increasingly critical. These systems enable dynamic adaptation to changing workloads and operating conditions, which is particularly important in healthcare environments characterized by heterogeneous data types and variable computational demands. Reconfigurable platforms support advanced optimization techniques such as algorithm–hardware co-design, dynamic voltage and frequency scaling, and partial reconfiguration, allowing systems to balance energy consumption and performance in real time. Moreover, emerging approaches such as memory-centric and in-memory computing significantly reduce data movement, a major source of energy overhead in edge systems. AI-driven design automation further enhances the efficiency of mapping, scheduling, and optimizing AI workloads on reconfigurable hardware, reducing development time while improving energy efficiency. Together, these capabilities position reconfigurable architectures as a foundational technology for next-generation Edge AI systems in healthcare [1], [3], [10].

This review provides a comprehensive analysis of energy-efficient reconfigurable architectures for Edge AI in IoT-enabled healthcare applications, focusing on trends, challenges, and future research directions. The main contributions are as follows:

- Trend analysis: identification of key developments, including the transition toward heterogeneous and domain-specific reconfigurable platforms, and the growing adoption of green AI practices in healthcare-oriented edge systems.
- Challenges: in-depth discussion of persistent issues such as energy–performance trade-offs, lack of standardization, runtime reconfiguration overheads, and security and privacy concerns in healthcare deployments.
- Future directions: exploration of emerging research opportunities, including self-optimizing and context-aware architectures, secure and trustworthy reconfiguration mechanisms, and sustainable, carbon-aware edge computing frameworks.

By aligning technological advancements with the stringent requirements of healthcare applications and sustainability goals, this review establishes reconfigurable architectures as a key enabler of scalable, energy-efficient, and trustworthy Edge AI systems for IoT-enabled healthcare environments [11], [12].

2. FOUNDATIONS OF EDGE AI AND IOT SYSTEMS

The rapid evolution of Edge AI and IoT systems has intensified the interaction between emerging intelligent technologies and conventional computing paradigms. At the core of this evolution lies the need to process data closer to its source, enabling low-latency decision-making while operating under strict energy and resource constraints. These requirements are particularly pronounced in IoT-enabled healthcare applications, where continuous sensing, real-time inference, data privacy, and long-term autonomous operation are essential. Wearable health devices, remote patient monitoring systems, and smart medical sensors demand architectures that are both computationally capable and energy efficient.

A fundamental enabler of this paradigm is the integration of heterogeneous and reconfigurable architectures, which provide a balance between flexibility, adaptability, and power efficiency. Unlike fixed-function hardware, reconfigurable platforms can dynamically tailor computational resources to application-specific workloads, thereby reducing latency and optimizing energy consumption—an essential capability for battery-powered or energy-harvesting healthcare devices. Conventional platforms such as CPUs and GPUs often fail to meet these stringent performance–energy requirements at the edge, underscoring the need for adaptive and energy-aware solutions. Figure 2 provides an overview of the foundational relationship between healthcare IoT workloads, system constraints, and computing architectures in Edge AI environments [13]. This section introduces the foundational concepts of Edge AI and IoT systems, examines the limitations of traditional hardware, and motivates the growing adoption of reconfigurable architectures as a key building block for sustainable Edge AI in healthcare environments [13].

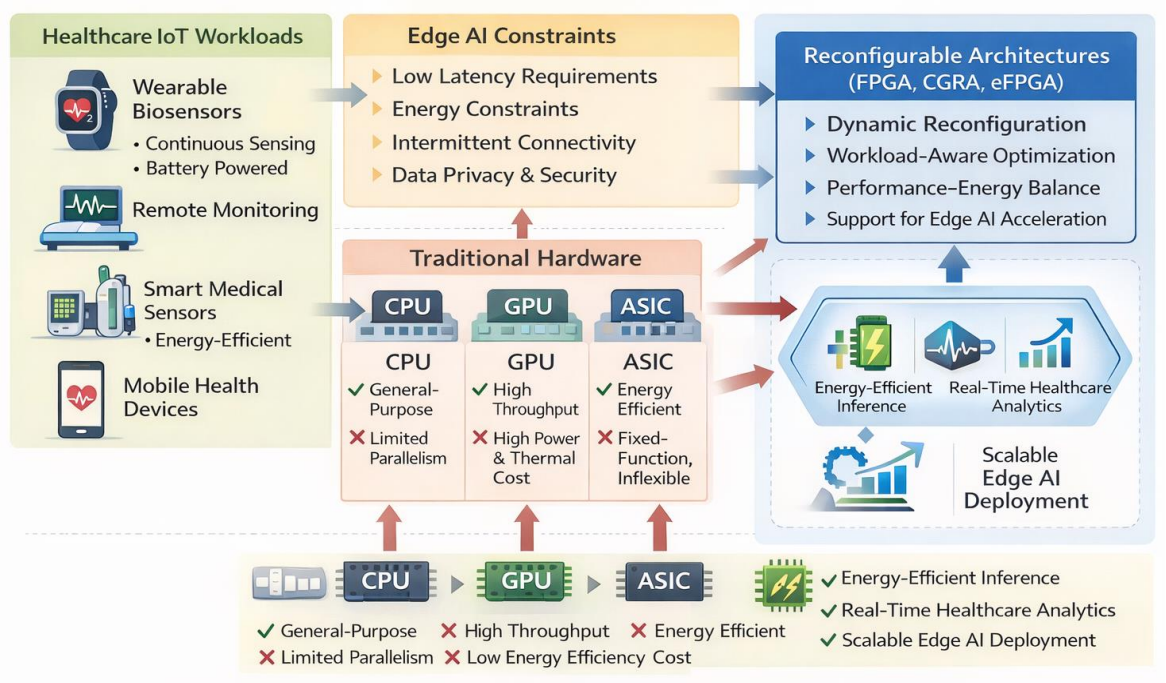

Figure 2. Foundations of Edge AI and IoT systems for healthcare applications

2.1. Defining Edge AI in IoT: workloads, latency, connectivity, and energy constraints

Edge AI in IoT systems refers to the deployment of AI inference and, in some cases, learning capabilities directly on edge devices or near-data sources. This approach minimizes reliance on cloud infrastructure, enabling low-latency processing, improved data privacy, and reduced communication overhead. In healthcare applications, such capabilities are critical for tasks such as real-time biosignal analysis, anomaly detection, activity recognition, and early warning systems, where delays or data loss can have serious consequences.

Edge AI workloads are characterized by high heterogeneity, ranging from lightweight signal processing to compute-intensive deep learning inference, all operating under constrained energy budgets and intermittent connectivity. These challenges are exacerbated by the need for reliable operation in diverse deployment scenarios, including mobile, wearable, and remote healthcare settings. Conventional hardware platforms often struggle to maintain an optimal balance between performance and energy efficiency across such diverse workloads. Reconfigurable architectures address this gap by enabling adaptive resource allocation and workload-aware optimization, allowing systems to respond dynamically to changing computational demands. Consequently, energy-efficient architectural design—supported by domain-specific overlays and AI-driven automation—has become a defining requirement for scalable and reliable Edge AI in IoT-enabled healthcare systems [13].

2.2. Conventional hardware platforms (CPU, GPU, and ASIC) and their limitations

The deployment of AI-enabled IoT systems at the edge has exposed inherent limitations in conventional hardware platforms, including CPUs, GPUs, and application-specific integrated circuits (ASICs). While widely adopted, these architectures exhibit trade-offs that restrict their effectiveness in energy-constrained and dynamically evolving environments such as healthcare IoT systems. CPUs offer general-purpose programmability but are limited in parallel processing capability, leading to inefficient execution of AI workloads under strict power constraints. GPUs provide high computational throughput and parallelism; however, their substantial power consumption and thermal requirements make them less suitable for continuous operation in compact, battery-powered healthcare devices. ASICs, on the other hand, deliver excellent energy efficiency for fixed workloads but lack flexibility, resulting in limited adaptability to evolving AI models, changing healthcare protocols, or multi-modal sensing requirements. Table 1 summarizes these limitations in the context of Edge AI for IoT systems.

These constraints motivate the exploration of alternative computing paradigms that can combine performance efficiency with adaptability. Reconfigurable architectures emerge as a compelling solution, offering a middle ground between general-purpose flexibility and application-specific optimization. By enabling hardware customization at runtime or design time, reconfigurable platforms can mitigate the energy–performance trade-offs inherent in conventional architectures, paving the way for sustainable Edge AI deployments in healthcare applications [14], [15].

Table 1. Limitations of conventional hardware platforms (CPU, GPU, and ASIC) in Edge AI for IoT systems

Platform	Limitation
CPU	Limited parallel processing capabilities, leading to suboptimal performance for AI workloads.
GPU	High power consumption and thermal output, making them less suitable for energy-constrained edge devices.
ASIC	High development costs and lack of flexibility, making them less adaptable to evolving AI algorithms.

2.3. Need for reconfigurability: adaptability and performance-energy balance

The dynamic and heterogeneous nature of Edge AI workloads in IoT-enabled healthcare systems necessitates architectures capable of adapting to changing operational conditions without compromising energy efficiency. Reconfigurability enables systems to dynamically adjust computational resources, memory hierarchies, and data paths in response to workload variations, sensor activity, and energy availability. This capability is particularly valuable in healthcare scenarios, where workloads may shift between continuous monitoring, event-driven analysis, and emergency response.

Techniques such as dynamic partial reconfiguration, workload-aware resource scaling, and algorithm–hardware co-design allow reconfigurable platforms to achieve fine-grained control over performance and power consumption. Recent studies demonstrate that real-time resource adaptation can significantly reduce energy usage while maintaining required quality-of-service levels [5]. Furthermore, the emergence of memory-centric and domain-specific reconfigurable architectures highlights the importance of minimizing data movement—one of the dominant contributors to energy consumption in edge systems [16].

In addition, AI-driven design automation is increasingly employed to simplify mapping, scheduling, and optimization of AI workloads on reconfigurable hardware, reducing design complexity and accelerating deployment. Collectively, these advances position reconfigurable architectures as a foundational technology for achieving sustainable, high-performance, and energy-efficient Edge AI in IoT-enabled healthcare environments.

3. RECONFIGURABLE ARCHITECTURES FOR EDGE ARTIFICIAL INTELLIGENCE

The growing demand for intelligent, low-latency, and energy-efficient processing in IoT-enabled healthcare systems has intensified interest in reconfigurable architectures as a foundation for Edge AI. Applications such as wearable health monitoring, medical image analysis, anomaly detection, and remote diagnostics require adaptable computing platforms capable of responding to dynamic workloads under stringent power constraints. Reconfigurable architectures—including FPGAs, CGRAs, and heterogeneous systems—offer a compelling balance between flexibility, performance, and energy efficiency, making them well suited for healthcare-oriented edge environments.

Unlike fixed-function hardware, reconfigurable platforms enable workload-aware customization and dynamic adaptation, allowing hardware resources to be optimized for specific AI tasks. The integration of AI-driven design automation further enhances energy efficiency by improving task mapping, scheduling, and resource allocation across heterogeneous components [13]. In parallel, the emergence of domain-specific overlays reflects a shift toward tailored architectures that accelerate healthcare-relevant AI workloads while

reducing design complexity. Despite these advances, challenges remain, including security risks associated with runtime reconfiguration, limited standardization, and interoperability across platforms [17]. Figure 3 presents a conceptual architecture of Edge AI systems, illustrating how reconfigurable components interact within healthcare-oriented IoT deployments.

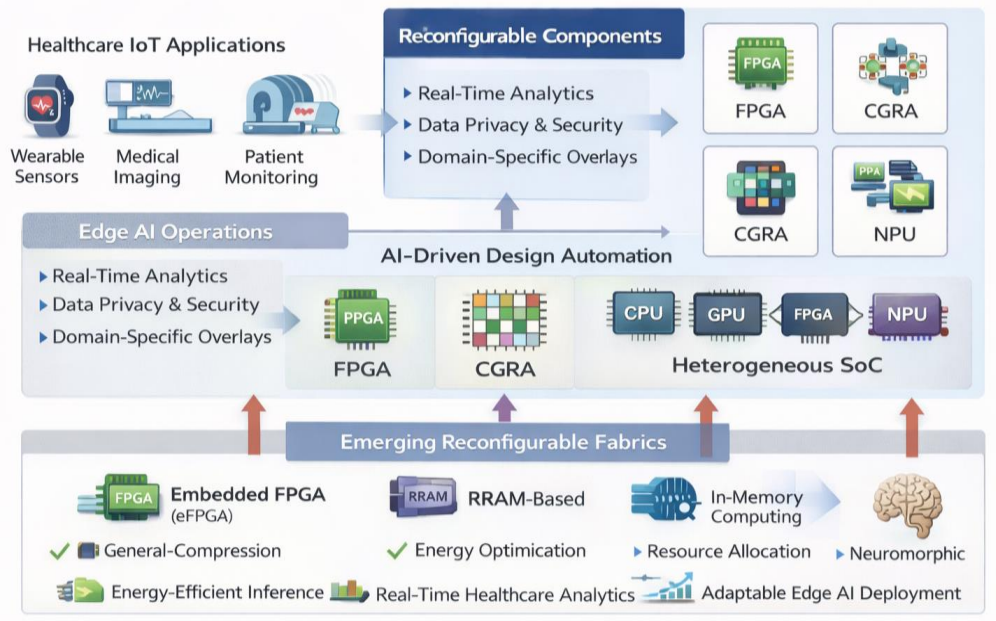

Figure 3. Architecture of Edge AI and its components

3.1. FPGAs: fine-grained reconfiguration and runtime adaptability

FPGAs play a central role in energy-efficient Edge AI due to their fine-grained reconfiguration capabilities and ability to adapt hardware resources at runtime. This flexibility enables FPGAs to closely match computational resources to application demands, reducing unnecessary energy consumption while maintaining high performance. In IoT-enabled healthcare systems, such adaptability is particularly valuable for workloads that vary over time, such as continuous biosignal processing interleaved with event-driven inference. FPGAs also support advanced optimization techniques, including pipeline parallelism, custom data paths, and memory-centric designs, which significantly reduce data movement and power overhead [18]. When integrated into heterogeneous system-on-chip (SoC) platforms, FPGAs can act as accelerators for compute-intensive AI tasks while CPUs handle control and system management functions. However, challenges persist in balancing energy efficiency, performance, and security, particularly in scenarios involving dynamic partial reconfiguration and sensitive healthcare data [6]. Continued advancements in design tools, security-aware reconfiguration, and high-level synthesis are expected to further strengthen the role of FPGAs in sustainable Edge AI ecosystems.

3.2. CGRAs: coarse-grained balance between flexibility and efficiency

CGRAs offer an alternative reconfigurable paradigm that strikes a balance between the flexibility of FPGAs and the efficiency of ASICs. By operating at a coarser granularity, CGRAs reduce configuration overhead and control complexity, enabling more predictable performance and improved energy efficiency for structured computation patterns. These characteristics make CGRAs particularly suitable for repetitive AI workloads commonly found in healthcare applications, such as signal filtering, feature extraction, and inference pipelines. The architectural design of CGRAs facilitates spatial and temporal reuse of processing elements, allowing dynamic adaptation to varying workloads while maintaining low power consumption. As sustainability becomes a key concern in IoT deployments, CGRAs are increasingly combined with memory-centric and in-memory computing techniques to minimize data movement and further reduce energy usage [19]. Although less flexible than fine-grained architectures, CGRAs represent an important design point for energy-efficient Edge AI systems that require a balance between adaptability and efficiency.

3.3. Heterogeneous SoCs: CPU+GPU+FPGA+NPU combinations

Heterogeneous SoCs that integrate CPUs, GPUs, FPGAs, and neural processing units (NPUs) represent a powerful architectural approach for Edge AI in IoT-enabled healthcare systems. By combining multiple computing paradigms on a single platform, these systems enable task-specific execution, assigning workloads to the most energy-efficient and performance-appropriate processing element. For example, CPUs manage control and communication tasks, GPUs accelerate parallel computations, FPGAs provide adaptable acceleration, and NPUs optimize deep learning inference. This architectural heterogeneity allows systems to dynamically adapt to changing workloads and operational conditions, which is essential in healthcare scenarios characterized by variability in data rates, computational intensity, and energy availability. Moreover, the integration of memory-centric computing within heterogeneous SoCs further enhances energy efficiency by reducing data transfer overheads. As design methodologies mature, heterogeneous reconfigurable SoCs are expected to play a critical role in enabling scalable, energy-efficient, and resilient Edge AI deployments in healthcare environments.

3.4. Emerging reconfigurable fabrics: eFPGAs, RRAM-based, in-memory, and neuromorphic accelerators

Beyond traditional reconfigurable platforms, emerging reconfigurable fabrics are gaining attention as enablers of next-generation Edge AI for IoT-enabled healthcare applications. Embedded FPGAs (eFPGAs) provide fine-grained reconfigurability within SoCs, enabling real-time adaptation with reduced integration overhead. RRAM-based architectures and in-memory computing paradigms significantly reduce data movement by integrating storage and computation, thereby addressing one of the primary sources of energy inefficiency in edge systems. Neuromorphic accelerators further extend this landscape by offering brain-inspired computing models that are inherently energy efficient and well suited for event-driven processing, such as anomaly detection and continuous monitoring in healthcare applications [20]. Despite their promise, these emerging technologies face challenges related to reliability, programmability, security, and integration with existing design flows. Addressing these issues is essential for realizing self-optimizing, trustworthy, and sustainable Edge AI architectures capable of supporting large-scale IoT-enabled healthcare deployments.

4. ENERGY-EFFICIENCY STRATEGIES

Energy efficiency is a fundamental requirement for Edge AI systems deployed in IoT-enabled healthcare applications, where devices often operate continuously under strict power, thermal, and reliability constraints. Wearable and implantable sensors, mobile diagnostic tools, and remote patient monitoring platforms must sustain long-term operation while supporting increasingly complex AI workloads. In this context, reconfigurable architectures offer a unique opportunity to implement energy-aware optimization strategies across hardware, memory, software, and algorithmic layers.

This section presents a comprehensive overview of energy-efficiency strategies tailored for reconfigurable Edge AI systems in healthcare-oriented IoT environments. At the hardware level, techniques such as dynamic partial reconfiguration, low-power fabrics, and dynamic voltage and frequency scaling (DVFS) enable fine-grained control over resource utilization and power consumption. At higher abstraction levels, memory-centric optimizations, software-driven scheduling, and algorithm–hardware co-design further reduce energy overheads while maintaining application-level performance and reliability. These complementary strategies are summarized in Table 2 and discussed in detail in the following subsections. Together, they form a holistic framework for achieving sustainable, energy-efficient Edge AI in healthcare deployments [21], [22].

Table 2. Energy-efficiency strategies in Edge AI for IoT-enabled healthcare systems

Energy-efficiency strategy	Description
AI-driven optimization	Leveraging machine learning techniques (e.g., reinforcement learning and predictive models) to dynamically optimize energy consumption, workload allocation, and system configuration.
Federated edge intelligence	Enabling collaborative learning across distributed healthcare edge nodes while minimizing data transmission and preserving privacy, thereby reducing communication energy costs.
Energy-aware early exiting	Implementing adaptive inference mechanisms that dynamically adjust computation depth per input, balancing diagnostic accuracy and energy consumption.
Self-adaptive AI applications	Designing applications that autonomously select energy-efficient configurations based on context, workload intensity, and available power resources.
Edge-centric processing	Performing real-time analytics locally at the edge to reduce latency, bandwidth usage, and energy consumption associated with cloud communication.

4.1. Hardware-level techniques: dynamic partial reconfiguration, low-power fabrics, and voltage/ frequency scaling

Hardware-level energy optimization forms the foundation of energy-efficient reconfigurable Edge AI systems. Dynamic partial reconfiguration (DPR) allows portions of a reconfigurable device—such as an FPGA—to be reconfigured at runtime without interrupting system operation. This capability is particularly valuable in healthcare applications, where workloads may vary between continuous monitoring and event-driven inference. By activating only the required hardware modules on demand, DPR significantly reduces static and dynamic power consumption.

Low-power reconfigurable fabrics further enhance efficiency through architectural optimizations that minimize leakage currents and switching activity. When combined with DVFS techniques, these fabrics enable adaptive control of operating voltage and clock frequency based on workload requirements and energy availability. Such fine-grained power management is essential for battery-powered and energy-harvesting healthcare devices, allowing systems to maintain reliable operation while extending device lifetime [23]. Collectively, these hardware-level techniques provide a robust foundation for energy-aware Edge AI architectures in healthcare environments.

4.2. Memory hierarchy optimizations: near-memory, in-memory computing, and data reuse

Data movement across memory hierarchies is a dominant contributor to energy consumption in Edge AI systems. This challenge is particularly pronounced in healthcare applications involving continuous biosignal processing, medical imaging, and multi-modal data fusion. Memory-centric optimization strategies—such as near-memory and in-memory computing—address this issue by bringing computation closer to data storage, thereby reducing latency and energy-intensive data transfers.

In-memory computing architectures exploit emerging memory technologies to perform computation directly within memory arrays, enabling massive parallelism and substantial energy savings. Complementary data reuse strategies further enhance efficiency by minimizing redundant memory accesses and exploiting temporal and spatial locality. These approaches are well suited for reconfigurable platforms, which can be customized to support application-specific memory access patterns and healthcare workloads [24], [25]. As Edge AI systems scale in complexity, memory hierarchy optimization will remain a critical enabler of energy-efficient healthcare IoT deployments.

4.3. Software-level techniques: compilers, scheduling, and mapping for energy efficiency

Beyond hardware and memory optimizations, software-level techniques play a crucial role in achieving energy efficiency in reconfigurable Edge AI systems. Advanced compiler frameworks, intelligent scheduling algorithms, and energy-aware mapping techniques enable efficient utilization of heterogeneous and reconfigurable resources. By optimizing task placement, execution order, and resource allocation, these methods reduce idle power consumption and avoid unnecessary computation.

In healthcare-oriented Edge AI systems, software-driven energy optimization is particularly important due to the diversity of workloads and strict real-time constraints. AI-driven design automation further enhances these capabilities by learning optimal scheduling and mapping strategies based on workload characteristics and system feedback. Domain-specific overlays tailored to healthcare AI workloads simplify deployment while improving energy efficiency and predictability [26], [27]. Together, these software-level strategies complement hardware optimizations and contribute to a cohesive, energy-aware system design.

4.4. Algorithm–hardware co-design: quantization, pruning, and model compression tailored for reconfigurable devices

Algorithm–hardware co-design is a key enabler of energy-efficient Edge AI, particularly in resource-constrained healthcare IoT systems. By jointly optimizing AI models and underlying hardware, this approach enables significant reductions in computational complexity, memory footprint, and energy consumption. Techniques such as quantization, pruning, and structured model compression are especially effective when tailored to the capabilities of reconfigurable architectures.

Reconfigurable devices can exploit reduced-precision arithmetic, custom data paths, and parallel execution to accelerate compressed models with minimal energy overhead. This adaptability is critical for healthcare applications that demand both accuracy and efficiency, such as real-time anomaly detection and personalized health monitoring. Recent advances in AI-driven co-design frameworks and domain-specific overlays further streamline this process, enabling rapid deployment of energy-efficient models across heterogeneous edge platforms [28], [29]. As Edge AI continues to evolve, algorithm–hardware co-design will remain central to achieving sustainable, high-performance healthcare IoT systems.

5. TRENDS IN ENERGY-EFFICIENT RECONFIGURABLE EDGE AI

The rapid convergence of IoT and AI at the network edge has accelerated the evolution of energy-efficient reconfigurable architectures, driven by the need to support intelligent, low-latency, and privacy-preserving applications under strict resource constraints. In IoT-enabled healthcare environments—such as wearable health monitoring, remote diagnostics, and smart medical sensing—these trends are particularly pronounced, as systems must operate continuously, reliably, and sustainably. Recent advances reflect a clear shift away from rigid, fixed-function designs toward adaptive, heterogeneous, and reconfigurable platforms that balance flexibility, performance, and energy efficiency.

Key trends include the growing adoption of AI-driven design automation for workload mapping and scheduling [30], the maturation of memory-centric and in-memory computing paradigms to reduce data movement [10], and the emergence of domain-specific overlays tailored to healthcare-relevant AI workloads. At the same time, sustainability considerations—such as green AI and carbon-aware computing—are becoming integral design objectives rather than secondary concerns. Figure 4 highlights the impact of energy-efficient Edge AI technologies across application domains, illustrating strong adoption growth and measurable energy savings, underscoring their effectiveness and practical relevance. Collectively, these trends define a dynamic and rapidly evolving research landscape that continues to open new opportunities for innovation while revealing persistent challenges.

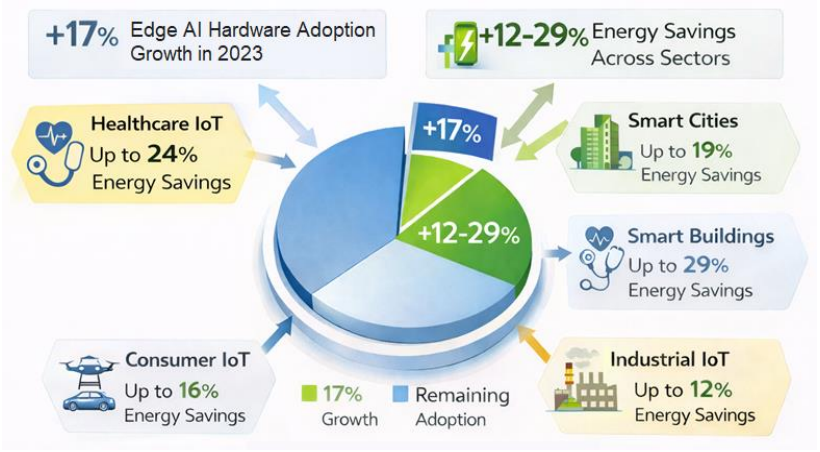

Figure 4. Impact of energy-efficient Edge AI technologies across various application sectors

5.1. Movement from fixed-function to reconfigurable platforms

One of the most significant trends in Edge AI is the transition from fixed-function hardware toward reconfigurable computing platforms. Fixed-function architectures, while efficient for predefined workloads, lack the adaptability required to support the heterogeneous and evolving AI tasks common in IoT-enabled healthcare systems. Reconfigurable platforms such as FPGAs and CGRAs address this limitation by enabling workload-aware customization, allowing hardware resources to be tailored dynamically to application requirements.

This shift is motivated by the need to balance energy efficiency with performance across diverse workloads, ranging from continuous biosignal processing to event-driven inference. Advances in AI-driven design automation further support this movement by simplifying the mapping and scheduling of AI workloads onto reconfigurable fabrics, reducing design complexity and deployment time. However, the widespread adoption of reconfigurable platforms also raises challenges related to standardization, security, and programming abstraction, highlighting the need for continued research and toolchain development.

5.2. Transition toward heterogeneous and reconfigurable architectures to balance flexibility and efficiency

Closely related to reconfigurability is the growing trend toward heterogeneous architectures, which integrate multiple processing elements—such as CPUs, GPUs, FPGAs, and NPUs—within a single system. This architectural diversity enables task-specific execution, ensuring that each workload is handled by the most energy-efficient and performance-appropriate component. In healthcare IoT systems, such heterogeneity is essential for managing the variability in data rates, computational intensity, and real-time constraints.

Reconfigurable components within heterogeneous systems provide additional adaptability, enabling dynamic resource allocation as workloads change. When combined with memory-centric computing techniques, including near-memory and in-memory processing, these architectures significantly reduce data movement and associated energy costs. This trend reflects a broader paradigm shift toward co-designed hardware–software systems that prioritize energy efficiency without sacrificing flexibility or scalability [5], [6].

5.3. AI-driven hardware design automation (ML-based mapping, scheduling, and adaptive compilers)

AI-driven hardware design automation has emerged as a transformative trend in the development of energy-efficient Edge AI systems. Machine learning techniques are increasingly employed to optimize workload mapping, scheduling, and resource allocation across reconfigurable and heterogeneous platforms. Adaptive compilers and machine learning (ML)-based optimization frameworks can learn from system feedback, enabling real-time adaptation to workload characteristics and operating conditions.

In IoT-enabled healthcare applications, where workloads and environmental conditions can vary significantly, such automation is critical for maintaining energy-efficient operation. By reducing reliance on manual design and static optimization, AI-driven automation enhances both efficiency and robustness. Nevertheless, challenges remain in ensuring the security, reliability, and interpretability of ML-based optimization mechanisms, particularly in safety-critical healthcare contexts [31].

5.4. Integration with edge-cloud continuum and 5G/6G ecosystems

Another important trend shaping energy-efficient Edge AI is the integration of reconfigurable architectures within the edge–cloud continuum, supported by emerging 5G and 6G communication infrastructures. These networks enable low-latency, high-bandwidth connectivity, allowing workloads to be dynamically distributed between edge devices and cloud resources based on energy availability, latency constraints, and application requirements.

Reconfigurable architectures play a key role in this ecosystem by enabling adaptive processing at the edge, reducing reliance on energy-intensive data transmission. AI-driven orchestration mechanisms further optimize task partitioning across the continuum, enhancing both energy efficiency and quality of service. This trend is particularly relevant for healthcare applications that require real-time responsiveness while leveraging cloud-scale analytics when necessary [32], [33].

5.5. Shift toward domain-specific overlays tailored for AI workloads at the edge

The growing complexity of AI workloads has driven increased interest in domain-specific overlays—customizable hardware abstractions designed to accelerate specific classes of applications. In Edge AI systems, such overlays simplify development while enabling efficient execution of AI models on reconfigurable hardware. For healthcare applications, domain-specific overlays can be tailored to common tasks such as signal processing, medical image analysis, and anomaly detection.

By abstracting low-level hardware details, overlays reduce design effort and improve portability, while still allowing fine-grained optimization for energy efficiency. When combined with AI-driven design automation, domain-specific overlays represent a powerful approach to reconciling the competing demands of performance, flexibility, and energy efficiency in IoT-enabled healthcare environments [34], [35].

5.6. Growing focus on sustainability (green AI and carbon-aware design)

Sustainability has become a defining trend in the design of Edge AI systems, particularly as IoT deployments scale globally. The concept of green AI emphasizes energy-efficient model design, resource-aware execution, and reduced environmental impact, aligning closely with the goals of reconfigurable computing. Carbon-aware design further extends this perspective by considering the environmental cost of computation and adapting system behavior accordingly.

In IoT-enabled healthcare applications, where long-term operation and reliability are critical, energy-efficient reconfigurable architectures provide a practical pathway toward sustainable deployment. Techniques such as lightweight model design, energy-aware scheduling, and adaptive resource management help reduce power consumption while maintaining application-level performance. Continued progress toward self-optimizing and carbon-aware architectures will be essential for achieving sustainable, large-scale Edge AI systems in healthcare and beyond [36].

6. APPLICATIONS IN IOT SYSTEMS

The integration of energy-efficient Edge AI within IoT systems has enabled a wide range of transformative applications across multiple domains, highlighting the importance of reconfigurable

architectures as a foundational technology. By enabling low-latency, intelligent processing close to data sources, Edge AI reduces communication overhead, enhances privacy, and significantly improves energy efficiency—capabilities that are particularly critical in resource-constrained and safety-sensitive environments.

In smart cities, Edge AI supports real-time traffic monitoring, surveillance, and energy management, contributing to improved urban efficiency and sustainability. In healthcare, IoT-enabled wearables, implantable sensors, and telemedicine platforms rely on continuous sensing and on-device intelligence to enable proactive and personalized care. Industrial IoT (IIoT) applications benefit from Edge AI through robotics, automation, and predictive maintenance, enhancing productivity while reducing downtime and energy waste. Consumer IoT systems, including smart homes and personal assistants, leverage Edge AI to deliver responsive and context-aware services under strict power constraints. Autonomous mobility platforms—such as drones and self-driving vehicles—further exemplify the need for adaptive, low-latency, and energy-efficient edge processing.

Table 3 summarizes representative energy-efficiency gains reported across key IoT application domains, illustrating the tangible benefits of deploying Edge AI on energy-aware and reconfigurable architectures. Despite these advances, challenges remain in managing energy–performance trade-offs, ensuring security, and maintaining reliability in dynamically reconfigurable systems. Addressing these challenges is essential for the sustainable evolution of IoT ecosystems, particularly in healthcare-oriented deployments where long-term operation and trustworthiness are paramount [10], [37].

Table 3. Energy efficiency in IoT applications

Application	Energy consumption reduction %	Source
Lighting	15–20	EIA 2017
HVAC systems in residential buildings	50	U.S. Department of Energy
Industrial IoT systems	Optimized through access point deployment strategies	NIST
Hazard detection systems	31–37% compared to continuous monitoring	PMC
Smart grid	99–99.99	PMC
Autonomous vehicles	99.99–100	PMC
e-Health	99.99–100	PMC

6.1. Smart cities: surveillance, traffic monitoring, and energy grids

As urban environments become increasingly complex, smart city infrastructures rely heavily on Edge AI to support real-time surveillance, traffic management, and intelligent energy grids. Processing data at the edge enables rapid response to dynamic conditions such as congestion, accidents, and fluctuating energy demand, while reducing dependence on centralized cloud resources.

Energy-efficient reconfigurable architectures are central to these systems, providing the adaptability required to handle heterogeneous workloads under strict power constraints. By dynamically tailoring hardware resources to application needs, reconfigurable platforms improve performance efficiency and system scalability. Moreover, the emergence of domain-specific overlays optimized for smart city workloads—such as video analytics and sensor fusion—further enhances energy efficiency and deployability [38]. Figure 5 (see subsection 6.2) illustrates the components and future directions of smart city systems integrated with AI, highlighting the role of edge intelligence in enabling sustainable and resilient urban infrastructures.

6.2. Healthcare IoT: wearables, telemedicine, and implantable sensors

Healthcare represents one of the most critical and impactful application domains for energy-efficient Edge AI in IoT systems. Wearable devices, telemedicine platforms, and implantable sensors increasingly rely on continuous data acquisition and real-time inference to support early diagnosis, chronic disease management, and personalized treatment. These applications impose stringent requirements on latency, reliability, data privacy, and long-term energy efficiency, particularly in mobile and battery-powered deployments.

Energy-efficient reconfigurable architectures play a central role in enabling such healthcare IoT systems by allowing dynamic adaptation to heterogeneous and time-varying workloads. For example, continuous biosignal monitoring (e.g., ECG, EEG, or SpO_2) can be combined with event-driven AI inference for anomaly detection or early warning, with hardware resources reconfigured on demand to minimize power consumption [39]. Edge AI processing also reduces dependence on cloud connectivity, improving system robustness and protecting sensitive patient data by keeping computation local.

Despite these advantages, several challenges remain, including power–performance trade-offs, security concerns related to runtime reconfiguration, and the lack of standardized programming models for

healthcare-grade reconfigurable platforms [40]. Addressing these challenges is essential to ensure safe, reliable, and scalable deployment of Edge AI in clinical and home-care environments.

Figure 6 is retained in this subsection to provide visual context, illustrating an IoT-based system architecture with edge intelligence for smart healthcare applications. The depicted architecture highlights key design principles such as distributed medical sensing, edge-level AI analytics, and adaptive resource management, which are essential for enabling continuous patient monitoring, low-latency clinical decision support, and energy-efficient operation in wearable, mobile, and implantable healthcare devices. This architecture demonstrates the versatility of reconfigurable Edge AI solutions in addressing the stringent requirements of healthcare IoT systems, including reliability, data privacy, and long-term autonomous operation.

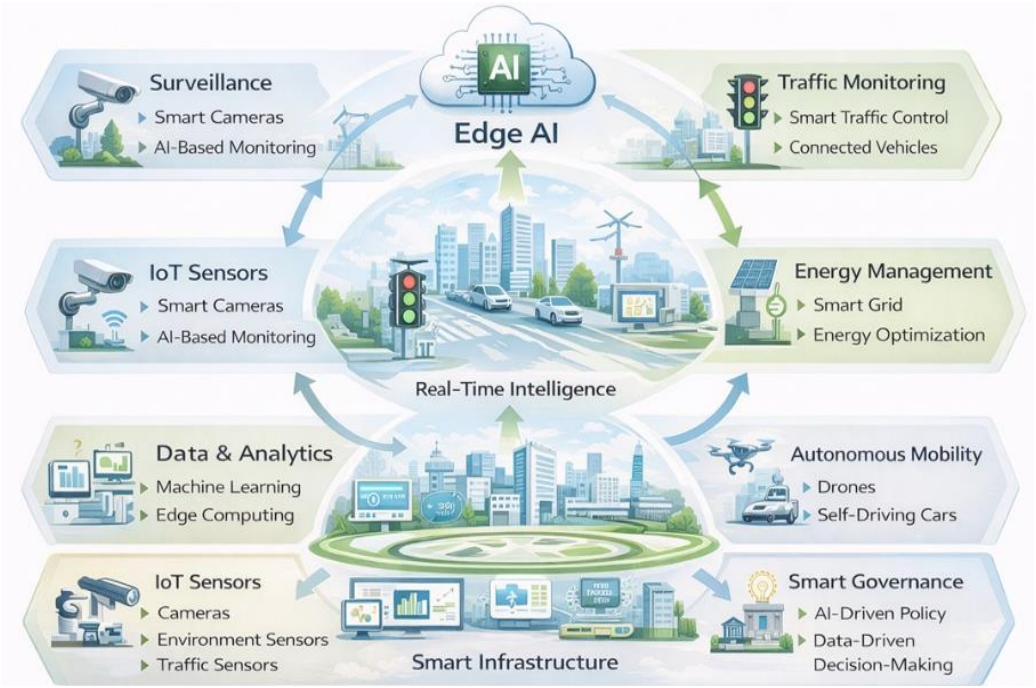

Figure 5. Diagram illustrating the components and future directions of smart cities and AI integration

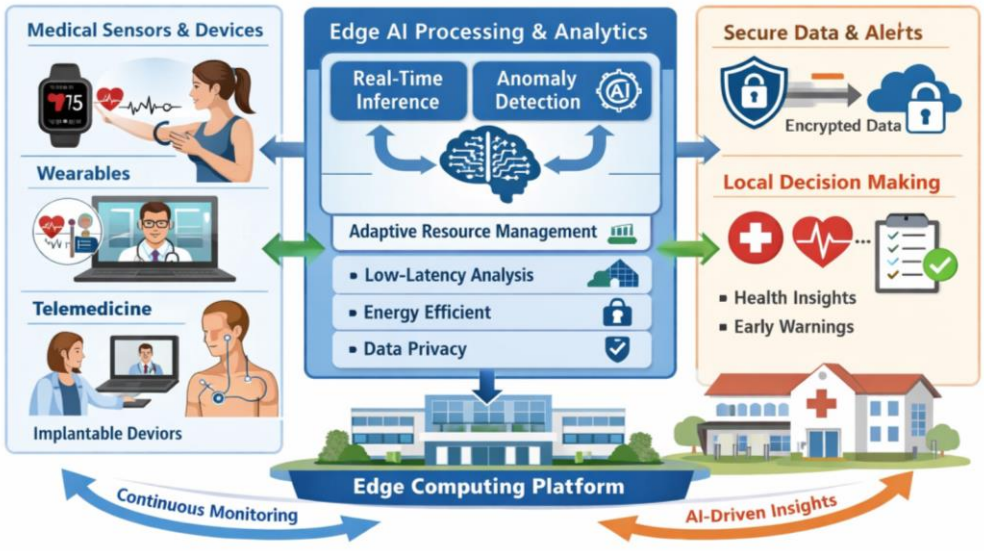

Figure 6. IoT solutions in smart healthcare

6.3. Industrial IoT: robotics, predictive maintenance, and automation

In industrial settings, IIoT systems leverage Edge AI to support robotics, automation, and predictive maintenance, enabling more efficient and reliable operations. By analyzing sensor data locally, Edge AI systems can detect anomalies, predict equipment failures, and optimize production processes with minimal latency. Energy-efficient reconfigurable architectures are particularly well suited to IIoT environments, where workloads are diverse and operational conditions vary over time. Their ability to adapt resource allocation dynamically allows systems to maintain high performance while minimizing energy consumption [41]. Edge computing further enhances efficiency by reducing data transmission and enabling localized decision-making [6]. However, scalability, system integration, and energy–performance trade-offs remain open challenges, motivating continued research into robust and adaptive IIoT architectures.

6.4. Consumer IoT: smart homes, AR/VR devices, and personal assistants

Consumer IoT applications—including smart homes, augmented and virtual reality (AR/VR) devices, and intelligent personal assistants—demand responsive and context-aware AI functionality under strict power and cost constraints. Edge AI enables these systems to deliver low-latency interactions and enhanced user experiences without continuous reliance on cloud connectivity.

Reconfigurable architectures provide the flexibility required to support diverse AI workloads, ranging from voice recognition and activity detection to immersive AR/VR processing. Dynamic reconfiguration allows consumer devices to optimize energy usage based on workload intensity and user behavior, contributing to more sustainable operation. As AI capabilities in consumer IoT continue to expand, energy-efficient reconfigurable platforms will play a key role in balancing performance, adaptability, and battery lifetime.

6.5. Autonomous mobility: drones and self-driving cars at the edge

Autonomous mobility systems, including drones and self-driving vehicles, represent some of the most demanding Edge AI applications in terms of latency, reliability, and energy efficiency. These systems must process large volumes of sensor data—such as video, LiDAR, and radar—in real time while operating within tight power budgets.

Reconfigurable architectures enable autonomous platforms to adapt processing resources dynamically to changing environments and mission requirements, improving both energy efficiency and operational robustness. AI-driven design automation further enhances these systems by optimizing workload scheduling and resource mapping at runtime. Looking ahead, self-optimizing and energy-aware architectures are expected to play a critical role in enabling sustainable autonomous mobility within broader IoT ecosystems, including smart cities and logistics networks [42].

7. CHALLENGES

Despite significant advances in Edge AI and reconfigurable computing, the realization of energy-efficient, reliable, and secure IoT-enabled healthcare systems remains constrained by several interrelated challenges. Healthcare applications—including wearable monitoring, telemedicine platforms, and implantable medical devices—impose uniquely stringent requirements, such as continuous operation, ultra-low power consumption, real-time responsiveness, high reliability, and strict data privacy. Meeting these requirements at the network edge, under severe resource constraints, substantially increases system design complexity.

As AI inference and analytics increasingly shift from centralized cloud infrastructures to edge devices, trade-offs among energy efficiency, performance, and architectural flexibility become more pronounced. Reconfigurable architectures offer a promising solution by enabling dynamic adaptation to time-varying workloads and operating conditions. However, their practical deployment in healthcare IoT systems is hindered by limited standardization, runtime reconfiguration overheads, and heightened security risks. Moreover, the heterogeneous nature of healthcare workloads—ranging from continuous biosignal acquisition (e.g., ECG, EEG, and SpO_2) to event-driven anomaly detection and decision support—demands fine-grained and context-aware resource management strategies security concerns further exacerbate these challenges, as reconfigurable platforms introduce new attack surfaces, including susceptibility to side-channel attacks and bitstream tampering, which pose serious threats to patient safety and data integrity. Additionally, ensuring reliability and scalability across diverse devices and clinical environments remains difficult, particularly under dynamic reconfiguration and long-term autonomous operation [43].

Figure 7 provides a consolidated view of these challenges in the context of IoT-enabled healthcare applications. Energy efficiency is identified as the most critical concern, driven by the power limitations of wearable and implantable devices. Challenges related to reconfigurability complexity and security vulnerabilities follow closely, reflecting the need for adaptive yet trustworthy architectures. Although latency

overheads and the lack of standardized healthcare-grade frameworks appear comparatively less dominant, they continue to impede real-time clinical responsiveness, interoperability, and large-scale deployment. Collectively, these challenges underscore the necessity for holistic, energy-aware, and security-centric design approaches to advance reconfigurable Edge AI architectures for next-generation smart healthcare systems.

Figure 7. Major challenges faced by energy-efficient reconfigurable Edge AI architectures in IoT-enabled healthcare applications

7.1. Persistent power–energy–performance–flexibility trade-offs in constrained healthcare IoT devices

In IoT-enabled healthcare applications, reconfigurable Edge AI architectures must continuously balance power consumption, energy efficiency, computational performance, and architectural flexibility. Devices such as wearable health monitors and implantable sensors operate under strict battery or energy-harvesting constraints while supporting continuous biosignal acquisition and real-time inference. These competing requirements make traditional fixed-function accelerators insufficient, as they lack adaptability to time-varying workloads and operating conditions.

Heterogeneous reconfigurable platforms that integrate CPUs, FPGAs, CGRAs, and domain-specific accelerators offer a promising pathway to address these trade-offs. By dynamically tailoring hardware resources to workload characteristics, such platforms can improve energy efficiency while preserving flexibility. However, this adaptability increases design complexity and often requires sophisticated runtime management mechanisms. In healthcare contexts, where predictability and reliability are paramount, achieving an optimal balance between flexibility and energy efficiency remains a fundamental challenge [11], [21].

7.2. Lack of standardized frameworks for reconfigurable edge IoT platforms in healthcare

The absence of standardized architectural frameworks and programming models significantly hinders the adoption of reconfigurable Edge AI platforms in healthcare IoT systems. Current implementations are frequently hardware-specific, limiting portability and complicating system integration across heterogeneous medical devices and vendors. This lack of standardization also impedes regulatory compliance, testing, and certification processes that are essential in healthcare environments.

Moreover, without unified abstraction layers, developers must manually manage low-level hardware details, increasing development time and the likelihood of errors. Although AI-driven design automation tools have shown potential in simplifying workload mapping and energy optimization, their impact remains limited in the absence of standardized, healthcare-aware frameworks. Addressing this gap is critical for enabling interoperable, scalable, and trustworthy reconfigurable Edge AI ecosystems in medical IoT applications.

7.3. Runtime reconfiguration overheads: latency, reliability, and predictability

Runtime reconfiguration is a key enabler of adaptability and energy-aware operation in reconfigurable Edge AI systems. However, in healthcare IoT applications, the overheads associated with dynamic reconfiguration—particularly latency, reliability, and predictability—pose significant challenges. Temporary suspension of computation during reconfiguration can introduce delays that are unacceptable for real-time patient monitoring and clinical decision support.

In addition, frequent reconfiguration increases the risk of transient faults and system instability, especially in long-term autonomous deployments. Ensuring predictable timing behavior under dynamic adaptation is therefore essential for safety-critical healthcare applications. While lightweight AI models and reconfiguration-aware scheduling techniques can mitigate some of these overheads, achieving dependable runtime reconfiguration without compromising energy efficiency remains an open challenge [24], [44].

7.4. Security vulnerabilities introduced by reconfigurability in healthcare Edge AI systems

The flexibility of reconfigurable architectures introduces new security vulnerabilities that are particularly concerning in healthcare IoT systems. Side-channel attacks can exploit physical leakage—such as power consumption or electromagnetic emissions—to extract sensitive patient data or cryptographic keys during AI inference and communication processes [44]. These attacks are difficult to detect and can undermine data confidentiality and patient safety.

Furthermore, bitstream tampering represents a critical threat, as malicious modification of configuration files can alter system behavior, degrade AI inference accuracy, or enable denial-of-service attacks [45]. Given the safety-critical nature of healthcare applications, such vulnerabilities are unacceptable. Mitigating these risks requires integrated security mechanisms spanning hardware, firmware, and software layers, as well as secure configuration management throughout the device lifecycle.

7.5. Need for lightweight, healthcare-aware AI models optimized for reconfigurable devices

The effectiveness of Edge AI in IoT-enabled healthcare systems is constrained by the limited suitability of conventional AI models for resource-constrained reconfigurable platforms. Deep learning models designed for cloud or GPU execution often impose excessive computational and energy demands, making them impractical for wearable and implantable devices.

To address this challenge, lightweight and hardware-aware AI models are required, tailored specifically to the characteristics of reconfigurable architectures. Techniques such as pruning, quantization, approximate computing, and knowledge distillation are essential for reducing model complexity while maintaining clinically acceptable accuracy. However, balancing energy efficiency, inference latency, model robustness, and security remains challenging, particularly in safety-critical healthcare contexts. Ongoing research highlights the need for co-design methodologies that jointly optimize AI models and reconfigurable hardware [46], [47].

7.6. Reliability and scalability across heterogeneous devices and workloads under dynamic reconfiguration

Healthcare IoT ecosystems are inherently heterogeneous, encompassing diverse devices, sensing modalities, communication protocols, and AI workloads. Ensuring reliability and scalability across such heterogeneous environments—especially under frequent dynamic reconfiguration—poses a significant challenge. Variability in device capabilities and operating conditions can lead to uneven performance, increased fault susceptibility, and difficulties in maintaining consistent quality of service.

Dynamic reconfiguration, while beneficial for adaptability, can further complicate system verification and fault management. Intelligent resource management algorithms have demonstrated potential in improving scalability by optimizing workload distribution and minimizing latency across distributed edge nodes [28]. Additionally, integrating secure and lightweight AI models with fault-aware reconfiguration strategies can enhance system robustness [48]. Addressing these reliability and scalability concerns is essential for the large-scale, long-term deployment of reconfigurable Edge AI architectures in IoT-enabled healthcare applications.

8. FUTURE DIRECTIONS

The evolving convergence of Edge AI and IoT-enabled healthcare systems presents significant opportunities to address the challenges identified in section 7. As healthcare applications demand continuous operation, ultra-low power consumption, adaptability to heterogeneous workloads, and strong security guarantees, future research must advance beyond isolated architectural or algorithmic improvements. Instead, emphasis should be placed on holistic, intelligent, and sustainable design paradigms that integrate

reconfigurable architectures, AI models, runtime systems, and healthcare-specific constraints. Key research directions are outlined in the following subsections.

8.1. AI-assisted hardware–software co-design for reconfigurable edge platforms

AI-assisted hardware–software co-design has emerged as a promising research direction for addressing the persistent energy–performance–flexibility trade-offs discussed in subsection 7.1. As healthcare IoT workloads grow in complexity—driven by continuous biosignal monitoring and AI-based diagnostics—reconfigurable platforms such as FPGAs and CGRAs enable adaptive mapping of computation to hardware resources. AI-driven design automation can further enhance energy efficiency by optimizing workload partitioning, scheduling, and configuration selection at runtime.

Despite these advances, challenges remain in ensuring secure and predictable behavior under dynamic reconfiguration, particularly in safety-critical healthcare environments. Future research should focus on self-optimizing architectures capable of autonomous adaptation while maintaining reliability and energy efficiency, thereby narrowing the gap between flexibility and clinical-grade performance [49], [50].

8.2. Unified frameworks for heterogeneous IoT workloads

The lack of standardized frameworks highlighted in subsection 7.2 motivates research into unified and portable frameworks capable of managing heterogeneous IoT workloads efficiently. Healthcare IoT systems must integrate diverse sensing modalities, AI inference tasks, and communication requirements under strict energy constraints. Reconfigurable architectures—combined with heterogeneous SoCs—offer the flexibility required to support such diversity, but their potential remains underutilized without unified abstraction layers.

AI-driven workload mapping and memory-centric optimizations can enable frameworks that adapt dynamically to application requirements, supporting energy-efficient execution across edge devices. While existing frameworks target domains such as smart cities and industrial IoT, future research should extend these concepts to healthcare-grade frameworks that prioritize reliability, security, and regulatory compliance.

8.3. Advances in memory-centric computing and 3D-stacked architectures

Memory-centric computing and 3D-stacked architectures represent a key research avenue for mitigating the energy and latency bottlenecks identified in subsection 7.3. In Edge AI healthcare applications, data movement often dominates energy consumption, particularly for continuous monitoring and inference workloads. By improving data locality and bandwidth, memory-centric designs significantly reduce energy overheads and enable low-latency processing.

The integration of domain-specific overlays for AI workloads further enhances adaptability and efficiency, supporting real-time reconfiguration in response to changing healthcare demands. Continued research into memory-centric and 3D-stacked solutions will be critical for achieving scalable, energy-efficient, and reconfigurable Edge AI architectures suitable for long-term healthcare deployments [51].

8.4. Self-optimizing architectures for real-time adaptive edge devices

Self-optimizing architectures capable of autonomous, workload-driven reconfiguration directly address the runtime adaptability challenges discussed in subsections 7.1 and 7.3. By leveraging AI-based monitoring and decision-making, these architectures can dynamically adjust configurations to balance energy efficiency and performance in real time.

Such capabilities are particularly valuable in healthcare IoT systems, where workload characteristics can change rapidly due to patient condition variability. However, issues related to security, standardization, and reliability persist, potentially limiting adoption. Future research should refine self-optimizing mechanisms while embedding security and verification support, ensuring safe and sustainable operation of autonomous Edge AI systems [52], [53].

8.5. Cross-layer co-design spanning hardware, software, and AI models

Cross-layer co-design remains a cornerstone research direction for overcoming the systemic challenges outlined in section 7. This approach integrates hardware architectures, software runtimes, and AI model design to jointly optimize energy efficiency, performance, and adaptability. Reconfigurable platforms such as FPGAs and CGRAs enable fine-grained control over computation, while AI-driven tools facilitate intelligent mapping and scheduling.

In healthcare IoT contexts, cross-layer co-design is particularly important to ensure that AI models remain lightweight, reliable, and clinically trustworthy under resource constraints. Future research should emphasize model–hardware co-design that accounts for accuracy, interpretability, and energy efficiency simultaneously, supporting sustainable Edge AI deployments [18], [54].

8.6. Secure and trustworthy reconfiguration mechanisms in IoT environments

The security vulnerabilities introduced by reconfigurability, discussed in subsection 7.4, necessitate robust and trustworthy reconfiguration mechanisms. Future research should focus on mitigating threats such as side-channel attacks and bitstream tampering through hardware-assisted security, secure configuration management, and lightweight cryptographic techniques.

Intelligent reconfiguration strategies that balance security, latency, and energy efficiency are particularly important for healthcare IoT systems, where data integrity and patient safety are paramount. Addressing standardization gaps and reliability concerns under dynamic reconfiguration will be essential for enabling trusted Edge AI platforms in real-world medical environments [5], [6].

8.7. Toward sustainable and carbon-efficient computing for large-scale healthcare IoT deployments

Sustainability and carbon efficiency are emerging as critical considerations for large-scale healthcare IoT deployments, aligning with the long-term challenges discussed in subsection 7.6. Reconfigurable architectures enable adaptive resource utilization, reducing energy waste and supporting sustainable operation under dynamic workloads.

Advances in lightweight AI models, domain-specific overlays, and in-memory computing have demonstrated significant potential for reducing energy consumption while maintaining performance [46]. Future research should integrate these techniques into self-adaptive, energy-aware Edge AI systems that support long-term autonomous operation and environmentally sustainable healthcare infrastructures.

9. CONCLUSION

The convergence of Edge AI and IoT-enabled healthcare systems has underscored the critical importance of energy-efficient reconfigurable architectures as a foundational enabler for next-generation medical applications. This survey has examined current trends, revealing a clear shift toward heterogeneous and reconfigurable computing platforms that balance flexibility, performance, and ultra-low power consumption—key requirements for wearable, implantable, and remote healthcare devices. The increasing adoption of AI-driven design automation, particularly for workload mapping and scheduling, further highlights the field's movement toward intelligent and adaptive system design.

Despite these advances, significant challenges remain. Persistent energy–performance trade-offs continue to constrain resource-limited edge devices, while the lack of standardized, healthcare-grade reconfigurable frameworks complicates interoperability and large-scale deployment. Moreover, dynamic reconfiguration introduces concerns related to security, reliability, and long-term autonomous operation, all of which are particularly critical in safety-sensitive healthcare environments. Addressing these challenges is essential to realizing self-optimizing Edge AI architectures capable of real-time adaptation under strict power, latency, and trust constraints.

Looking forward, future research must prioritize secure and trustworthy reconfiguration mechanisms, cross-layer co-design spanning hardware, software, AI models, and sustainability-driven innovations aligned with green and carbon-efficient computing principles. Advances in memory-centric computing, lightweight AI models, and autonomous runtime optimization present promising opportunities to mitigate current limitations while supporting scalable healthcare deployments. Collectively, these directions reinforce the role of energy-efficient reconfigurable architectures as a cornerstone technology for advancing robust, adaptive, and sustainable Edge AI solutions in IoT-enabled healthcare applications.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the institutional support provided by Universitas Ahmad Dahlan (UAD) and Universiti Teknikal Malaysia Melaka (UTeM), which was instrumental in facilitating this research. Special appreciation is extended to the Embedded System and Power Electronics Research Group (ESPERG) for their sustained technical contributions, collaborative engagement, and domain-specific expertise. The ESPERG team's involvement significantly enriched the methodological rigor and practical relevance of the study, particularly in the areas of embedded system design and power optimization. The AI tool (ChatGPT) was used solely for visualization and illustration purposes. All scientific content and interpretations were provided and verified by the authors.

FUNDING INFORMATION

The research was funded by PT. Intelektual Pustaka Media Utama (IPMU) under contract number 07/RST-E/IPMU/I/2025, which supported the facilitation of this work.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Tole Sutikno	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Aiman Zakwan Jidin		✓	✓	✓	✓	✓		✓		✓	✓			
Lina Handayani	✓			✓	✓		✓		✓				✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] M. Prabha, S. Nandhini, M. Dayanidhy, and R. Pradeep, "Edge-AI integrated secure wireless IoT architecture for real time healthcare monitoring and federated anomaly detection," *Scientific Reports*, vol. 16, no. 1, p. 574, 2025, doi: 10.1038/s41598-025-30150-x.

[2] F. C. Andriulo, M. Fiore, M. Mongiello, E. Traversa, and V. Zizzo, "Edge Computing and Cloud Computing for Internet of Things: A Review," *Informatics*, vol. 11, no. 4, 2024, doi: 10.3390/informatics11040071.

[3] M. M. Quddos, I. Salim, B. Ahmad, A. Riaz, and L. Riaz, "Advancing Smart IoT Systems with Real-Time Edge AI: Machine Learning Models, Low-Latency Inference, and Energy-Efficient Architectures for Resource-Limited Devices," *Annual Methodological Archive Research Review*, vol. 3, no. 8, pp. 445–474, 2025, doi: 10.63075/hegxxe78.

[4] H. J. Damsgaard *et al.*, "Adaptive approximate computing in edge AI and IoT applications: A review," *Journal of Systems Architecture*, vol. 150, p. 103114, May 2024, doi: 10.1016/j.sysarc.2024.103114.

[5] M. H. Maturi, S. Podicheti, and D. Kumar, "Optimizing Energy Efficiency in Edge-Computing Environments with Dynamic Resource Allocation," *International Journal of Science and Engineering Applications*, vol. 13, no. 7, pp. 1–58, 2024, doi: 10.7753/ijsea1307.1001.

[6] M. Mendula, "Middleware-Enabled Frugality for Intelligent and Distributed Edge Applications," Ph.D. dissertation, Alma Mater Studiorum – Università di Bologna, Italy, 2024.

[7] S. K. Jagatheesaperumal, Q. V. Pham, R. Ruby, Z. Yang, C. Xu, and Z. Zhang, "Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 2106–2136, 2022, doi: 10.1109/OJCOMS.2022.3215676.

[8] H. Chowdhury, D. B. P. Argha, and M. A. Ahmed, "Artificial Intelligence in Sustainable Vertical Farming," *arXiv peprint*, 2023, doi: 10.48550/arXiv.2312.00030.

[9] W.-P. Kiat, K.-M. Mok, W.-K. Lee, H.-G. Goh, and R. Achar, "An energy efficient FPGA partial reconfiguration based micro-architectural technique for IoT applications," *Microprocessors and Microsystems*, vol. 73, p. 102966, 2020, doi: 10.1016/j.micpro.2019.102966.

[10] M. A. Kishor, "Intelligent Edge Computing for IOT: AI-Powered Decision Making at the Edge," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 4, pp. 107–111, 2025, doi: 10.22214/ijraset.2025.68271.

[11] A. Liu *et al.*, "The Roadmap of 2D Materials and Devices Toward Chips," *Nano-Micro Letters*, vol. 16, no. 1, pp. 1–96, Dec. 2024, doi: 10.1007/s40820-023-01273-5.

[12] H. Kurunathan, H. Huang, K. Li, W. Ni, and E. Hossain, "Machine Learning-Aided Operations and Communications of Unmanned Aerial Vehicles: A Contemporary Survey," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 496–533, 2024, doi: 10.1109/COMST.2023.3312221.

[13] S. S. Gill *et al.*, "Modern computing: Vision and challenges (1)," *Telematics and Informatics Reports*, vol. 13, pp. 1–38, Mar. 2024, doi: 10.1016/j.teler.2024.100116.

[14] A. J. Tyagi, "Hardware/Software Co-design: Addressing Uncertainty in Platform Development through Workload Modeling and Bottleneck Feedback Loops," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 5, pp. 560–576, 2023, doi: 10.17762/ijritcc.v11i5.11705.

[15] T. N. Mohaidat, "Efficient Design of Neural Network Hardware Accelerator for Enhanced Performance in Deep Learning Systems," M.S. thesis, Univ. of Mississippi, 2024.

[16] Z. Song, M. Jansen, and D. Bonetta, "Energy Consumption and Optimization Strategies of Cloud-Based Big Data and Machine Learning Applications: Current Trends and Future Directions," *Atlarge-research*, pp. 1–25, 2025.

[17] O. L. A. Lopez *et al.*, "Energy-Sustainable IoT Connectivity: Vision, Technological Enablers, Challenges, and Future Directions," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 2609–2666, 2023, doi: 10.1109/OJCOMS.2023.3323832.

[18] K. Wali, "Hardware-Software Co-Design for Power-Efficient Edge-AI Systems," *Journal of Artificial Intelligence, Machine Learning and Data Science*, vol. 2, no. 4, pp. 2754–2761, Oct. 2024, doi: 10.51219/jaimld/karthik-wali/580.
[19] M. Wijtvliet, H. Corporaal, and A. Kumar, *Blocks, Towards Energy-efficient, Coarse-grained Reconfigurable Architectures*. Springer International Publishing, 2021, doi: 10.1007/978-3-030-79774-4.
[20] F. Palumbo, G. Keramidas, N. Voros, and P. C. Diniz, *Applied Reconfigurable Computing. Architectures, Tools, and Applications*. Cottbus, Germany: Springer Nature, 2023.
[21] A. I. Kuznetsov *et al.*, "Roadmap for Optical Metasurfaces," *ACS Photonics*, vol. 11, no. 3, pp. 816–865, Mar. 2024, doi: 10.1021/acsphotonics.3c00457.
[22] K. Telli *et al.*, "A Comprehensive Review of Recent Research Trends on Unmanned Aerial Vehicles (UAVs)," *Systems*, vol. 11, no. 8, p. 400, Aug. 2023, doi: 10.3390/systems11080400.
[23] F. Rincón, J. Barba, H. K. H. So, P. Diniz, and J. Caba, *Applied Reconfigurable Computing. Architectures, Tools, and Applications*. Toledo, Spain: Springer Nature, 2020.
[24] A. M. Dalloo, A. J. Humaidi, A. K. Al Mhdawi, and H. Al-Raweshidy, "Approximate Computing: Concepts, Architectures, Challenges, Applications, and Future Directions," *IEEE Access*, vol. 12, pp. 146022–146088, 2024, doi: 10.1109/ACCESS.2024.3467375.
[25] A. Garofalo, "Flexible Computing Systems For AI Acceleration At The Extreme Edge Of The IoT," Ph.D. dissertation, Alma Mater Studiorum – Università di Bologna, Italy, 2022.
[26] F. Ferrandi *et al.*, "A Survey on Design Methodologies for Accelerating Deep Learning on Heterogeneous Architectures," *arXiv peprint*, 2023, doi: 10.48550/arXiv.2311.17815.
[27] E. Valpreda, "Hardware/Neural Network Codesign for Energy-Efficient Inference on Edge Devices with Optimal Mapping and Compression," Ph.D. dissertation, Politecnico di Torino, Italy, 2024.
[28] M. I. ul Haq *et al.*, "Intelligent Embedded Platforms: Co-Design of VLSI Architectures and Deep Learning Models for Scalable Optimization and Real-World Deployment," *Scholars Journal of Engineering and Technology*, vol. 13, no. 09, pp. 708–717, Sep. 2025, doi: 10.36347/sjet.2025.v13i09.001.
[29] H. You and Y. (Celine) Lin, "ICCAD: G: Machine Learning Algorithm and Hardware Co-Design Towards Green and Ubiquitous AI on Both Edge and Cloud," *ACM Student Research Competition*, pp. 1–5, 2024.
[30] A. H. A. N. Karsa, "AI for Sustainability: Bibliometric Review of Power-Saving Algorithms in IoT and Edge Systems," *Breakthroughs Information Technology*, vol. 1, no. 1, pp. 48–64, Jun. 2025, doi: 10.70764/gdpu-bit.2025.1(1)-04.
[31] H. Huang and H. Yu, "Compact and Fast Machine Learning Accelerator for IoT Devices," in *Springer*, Springer, 2018, p. 149.
[32] M. N. Alatawi, "EdgeGuard-IoT: 6G-Enabled Edge Intelligence for Secure Federated Learning and Adaptive Anomaly Detection in Industry 5.0," *Computers, Materials & Continua*, vol. 85, no. 1, pp. 1–33, 2025, doi: 10.32604/cmc.2025.066606.
[33] K. Trichias *et al.*, "D1.4: Second Period Assessment and Planning Report," *SNS OPS Consortium Parties*, pp. 1–98, 2025.
[34] S. Reno and K. Roy, "Navigating the Blockchain Trilemma: A Review of Recent Advances and Emerging Solutions in Decentralization, Security, and Scalability Optimization," *Computers, Materials and Continua*, vol. 84, no. 2, pp. 2061–2119, 2025, doi: 10.32604/cmc.2025.066366.
[35] N. G. M. N. Alliance, "Automation and Autonomous System Architecture Framework – Phase 2," *NGMN Alliance e.V.* pp. 1–53, 2024.
[36] S. Kulkarni, J. N. Dwivedi, D. Pramanta, and Y. Tanaka, "Edge Computational Intelligence for AI-Enabled IoT Systems," in *CRC Press*, CRC Press, 2024, pp. 1–328, doi: 10.1201/9781032650722.
[37] C. Zhao *et al.*, "Edge General Intelligence Through World Models and Agentic AI: Fundamentals, Solutions, and Challenges," 2025, doi: 10.48550/arXiv.2508.09561.
[38] I. Sarkar, A. Hazra, and P. Maurya, *Industry 5.0: Key Technologies and Drivers*. Springer Nature, 2025.
[39] L. S. Ahmed and A. I. Siddiq, "A Comprehensive Review of the Internet of Medical Things in Healthcare," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 13, no. 6, pp. 415–426, 2024, doi: 10.18178/IJEETC.13.6.415-426.
[40] B. N. Reddy, S. Saravanan, V. Manjunath, and P. R. S. Reddy, "Review on Next-Gen Healthcare: The Role of MEMS and Nanomaterials in Enhancing Diagnostic and Therapeutic Outcomes," *Biomaterials Connect*, vol. 1, no. 1, pp. 1–10, 2024, doi: 10.69709/biomatc.2024.131006.
[41] I. I. Gorial, "Recent Trends in Energy-Efficient Design of Mechatronic Systems: A Comprehensive Review," *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, pp. 83–98, Apr. 2025, doi: 10.33103/uot.ijccce.25.1.7.
[42] N. T. R. Babu, "*Building Energy-efficient Edge Systems*," M.S. thesis, Ohio State Univ., Columbus, OH, USA, 2020. N. T. R.
[43] R. Douglass, K. Gremban, A. Swami, and S. Gerali, *IoT for Defense and National Security*. John Wiley & Sons, 2023.
[44] L. K. Qurban, "Lightweight Cryptographic Models for IoT Devices: A Deep Learning Approach to Power Side-Channel Attack Prevention," *International Journal of Professional Studies*, vol. 19, no. 1, pp. 49–61, 2025, doi: 10.37648/ijps.v19i01.005.
[45] A. Paju, "Distributed EaaS simulation using TEEs: A case study in the implementation and practical application of an embedded computer cluster," Ph.D. dissertation, Tampere University, Finland, 2022.
[46] H. Rahaman, C. Giri, S. K. Roy, and A. Chakrabarti, "Optimization and Security of AI Models for Deployment at Edge: A Comprehensive Review," in *2025 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2025, pp. 1–6, doi: 10.1109/ISVLSI65124.2025.11130213.
[47] I. Cheikh, S. Roy, E. Sabir, and R. Aouami, "Energy, Scalability, Data and Security in Massive IoT: Current Landscape and Future Directions," *arXiv peprint*, 2025, doi: 10.48550/arXiv.2505.03036.
[48] D. S. Yaseen and H. Joshi, "Lightweight Multi-Hop Routing Protocols for Efficient Resource Utilization in Edge-Enabled PLC IoT Networks," *Journal of Nonlinear Analysis and Optimization*, vol. 16, no. 1, pp. 1–18, 2025.
[49] N. M. M. Sathyamoorthy, and R. K. Dhanaraj, *Applications and Challenges of Reconfigurable Intelligent Surfaces in 6G*. IGI Global, 2025.
[50] J. J. N. Serra, "Energy-efficient hardware/software co-design for dynamically reconfigurable architectures," Universitat Politècnica de Catalunya, 2005.
[51] W. Qian, *Energy-efficient Spatio-temporal Computing Framework*. Case Western Reserve University, 2016.
[52] M. Hendaoui, "Integrating AI into Key Enabling Technologies for 6G Networks: A Review from SDN to Quantum Computing," *Journal of Applied Engineering Science and Technology*, vol. 5, no. 1, pp. 1–9, Apr. 2025, doi: 10.69717/jaest.v5.i1.109.
[53] O. Amin *et al.*, "Beyond The Wi-Fi Era," *Frontiers in Communications and Networks*, vol. 5, pp. 1–18, 2024, doi: 10.3389/frcmn.2024.1486488.
[54] D. N. Danopoulos, "Hardware-Software Co-Design of Deep Learning Accelerators: From Custom to Automated Design Methodologies," Ph.D. dissertation, National Technical University of Athens, Greece, 2024.

BIOGRAPHIES OF AUTHORS

Prof. Ir. Tole Sutikno, S.T., M.T., Ph.D., MIET, IPM., ASEAN Eng., is a full professor in the Department of Electrical Engineering at Universitas Ahmad Dahlan (UAD), Yogyakarta, Indonesia. He has held this position since 2023, after serving as an associate professor since 2008. He received his bachelor's degree from Universitas Diponegoro (1999), his master's degree from Universitas Gadjah Mada (2004), and his Ph.D. in Electrical Engineering from Universiti Teknologi Malaysia (2016), where his research focused on digital power electronics and intelligent control systems. From 2016 to 2021, he was Director of the Institute for Scientific Publishing and Publications (LPPI) at UAD, leading efforts to enhance research visibility, journal management, and international scholarly collaboration. He currently serves as Head of the Master's Program in Electrical Engineering (since 2023), following his role as Head of the Undergraduate Program (2022–2023). He is also the founding leader of the Embedded Systems and Power Electronics Research Group (ESPERG), collaborating nationally and internationally on embedded systems, FPGA-based control, and renewable energy integration. His research spans digital design, power electronics, motor drives, robotics, intelligent systems, and AI-based automation, with an emphasis on industrial and healthcare applications. He has published over 380 Scopus-indexed peer-reviewed articles. As of 2025, his work has received more than 6,000 citations, with an h-index of 36 and an i10-index of 174. He has been recognized among the Top 2% of Scientists Worldwide by Stanford University and Elsevier BV since 2021. He can be contacted at email: tole@te.uad.ac.id.

Dr. Aiman Zakwan Jidin is a lecturer and researcher in the Department of Electronic and Computer Engineering at Universiti Teknikal Malaysia Melaka (UTeM). He earned his Ph.D. in Electrical Engineering from Universiti Malaysia Perlis (UniMAP) in 2025, with a dissertation focused on optimizing memory testing algorithm efficiency to improve fault coverage, particularly in SRAM-based embedded systems. His doctoral work contributes to the advancement of low-complexity March algorithms and Memory Built-In Self-Test (MBIST) strategies for FPGA platforms. Prior to his doctoral studies, he obtained his Master of Engineering in Electronic and Microelectronic Systems from ESIEE Engineering Paris, France, in 2011. He began his professional career as an FPGA IP Core Design Engineer at Altera Corporation Malaysia (now part of Intel), where he was involved in developing and validating reusable logic blocks for programmable devices. At UTeM, he is actively involved in teaching and supervising undergraduate and postgraduate research in design-for-testability (DFT), VLSI architecture, FPGA system design, reconfigurable architectures, fault-tolerant systems, and hardware-software co-design. He can be contacted at email: aimanzakwan@utem.edu.my.

Lina Handayani, S.KM., M.Kes., Ph.D., is an Associate Professor at the Magister of Public Health Department, Faculty of Public Health, Universitas Ahmad Dahlan (UAD), Yogyakarta, Indonesia. She earned her bachelor's degree in public health from Universitas Diponegoro in 2003, her master's degree from Universitas Gadjah Mada in 2007, and her Ph.D. in educational psychology from Universiti Teknologi Malaysia in 2013, where her doctoral research focused on behavioural determinants of health-promoting practices. From 2017 to 2022, she served as Dean of the Faculty of Public Health at UAD, leading strategic initiatives in curriculum reform, accreditation, and community-based health programmes. Her academic contributions span health education and promotion, community nutrition, environmental health, psychosocial support systems, and digital health literacy. She has published over 40 peer-reviewed articles in national and international journals, with more than 1,900 citations recorded on Google Scholar, reflecting her growing impact in public health and behavioural science. She is actively involved in collaborative research and policy advocacy with governmental agencies, NGOs, and academic institutions, particularly in maternal and child health, nutrition education, and community empowerment. She also serves as a reviewer for several public health journals and mentor's postgraduate students in interdisciplinary research. Her work integrates educational psychology with public health strategies to improve health outcomes in diverse populations. She can be contacted at email: lina.handayani@ikm.uad.ac.id.