Transactions on Computational and Scientific Methods | Vol. 5, No. 6, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Scalable and Secure Edge AI: Foundations, Applications, and Open Research Issues

Callum Rourke[1], Maris Leclair[2]
[1]University of Windsor, Windsor, Canada
[2]University of Windsor, Windsor, Canada
*Corresponding Author: Callum Rourke; crourke@uwindsor.ca

Abstract: Edge Artificial Intelligence (Edge AI) represents a paradigm shift in intelligent computing by relocating model inference and training to the edge of the network. This transformation enables real-time decision-making, reduces data transmission, enhances user privacy, and supports context-aware applications. This paper presents a comprehensive survey of Edge AI, examining its technological foundations, system architectures, deployment strategies, and applications across sectors such as healthcare, transportation, manufacturing, and environmental monitoring. We analyze core challenges including model optimization under hardware constraints, secure deployment, privacy-preserving learning, and ethical concerns. Furthermore, we outline open research problems and discuss future trends including 6G-enabled edge intelligence, the adaptation of foundation models for embedded devices, and collaborative edge learning. The survey aims to provide researchers, engineers, and policymakers with an integrative understanding of Edge AI, guiding the development of scalable, secure, and sustainable intelligent systems.

Keywords: Edge AI; Federated Learning; Model Compression; Edge Computing

1. Introduction

The rapid proliferation of connected devices, coupled with the surging demand for real-time intelligent services, has catalyzed the emergence of Edge Artificial Intelligence (Edge AI) as a transformative computing paradigm. Traditional AI systems, primarily reliant on cloud-based computation, are increasingly strained by latency-sensitive applications, privacy concerns, and network bandwidth limitations. As a result, Edge AI, which refers to the deployment of AI models directly on edge devices such as smartphones, IoT sensors, drones, or microcontrollers, is gaining momentum as a solution to bridge the gap between centralized intelligence and localized responsiveness [1].

The key motivation behind Edge AI lies in its ability to enable low-latency inference, reduce reliance on constant network connectivity, and support context-aware, energy-efficient decision-making at the device level. These advantages are particularly critical in domains such as autonomous vehicles, remote healthcare monitoring, industrial automation, and augmented reality, where real-time feedback and minimal delay are essential for operational safety and user experience [2], [3]. By pushing AI workloads closer to the data source, Edge AI not only enhances responsiveness but also mitigates privacy risks associated with transmitting sensitive information to the cloud [4].

A fundamental differentiator between cloud AI and Edge AI lies in the resource constraints and deployment environments. While cloud platforms benefit from virtually unlimited computational power, storage, and centralized model management, edge environments are inherently limited in processing capability, memory, power, and cooling. This dichotomy has given rise to a range of innovations in model optimization (e.g., quantization, pruning), distributed inference frameworks, and hardware accelerators tailored for edge inference, such as Google's Edge TPU or NVIDIA's Jetson series [5], [6]. Furthermore, techniques such as federated learning have emerged to facilitate decentralized training without the need for raw data aggregation, addressing both scalability and data privacy concerns [7].

The rise of 5G and upcoming 6G connectivity also plays a pivotal role in accelerating Edge AI adoption. These network advancements enable new forms of edge-cloud collaboration, where computational tasks can be dynamically offloaded based on network conditions, device status, and task urgency [8]. Moreover, developments in edge-native AI platforms, like TensorFlow Lite, PyTorch Mobile, and ONNX Runtime, have simplified the model deployment pipeline for developers, thereby democratizing the use of AI on consumer and industrial devices alike [9]. Edge AI also synergizes with the broader vision of ubiquitous computing and the Internet of Things (IoT), acting as a key enabler for intelligent, decentralized cyber-physical systems [10].

Despite its promising capabilities, Edge AI faces several technical challenges that distinguish it from conventional AI paradigms. The limitations of edge hardware necessitate lightweight models that balance accuracy, latency, and energy consumption. Additionally, the heterogeneity of edge devices and operating environments introduces complexities in model deployment, updating, and orchestration. These challenges are compounded by security and trust issues, as edge devices are often deployed in unprotected or remote environments, making them susceptible to adversarial attacks or physical tampering [11], [12].

This survey aims to provide a comprehensive overview of the current landscape, technical foundations, and future prospects of Edge AI. We first examine the key enabling technologies, including edge hardware, model compression strategies, and distributed learning frameworks. Next, we discuss system architectures and deployment methodologies, highlighting the design trade-offs involved in real-world edge AI systems. We then explore major application domains where Edge AI is making significant impact, such as healthcare, transportation, manufacturing, and environmental sensing. Furthermore, we delve into critical issues related to privacy, security, and regulation, followed by an in-depth discussion on open research challenges and emerging trends, including AI-native edge hardware and lifelong learning at the edge.

The main contributions of this survey are as follows:

(1) We provide a structured taxonomy of Edge AI technologies, spanning hardware, software, and learning paradigms.

(2) We synthesize and analyze current research and industrial practices in Edge AI deployment across key domains.

(3) We identify critical gaps and unresolved challenges that hinder the full realization of Edge AI.

(4) We outline future directions and innovations expected to shape the next generation of intelligent edge systems.

The remainder of this paper is organized as follows. Section II discusses the foundational technologies that underpin Edge AI. Section III describes architectures and deployment strategies tailored for edge

environments. Section IV presents prominent application scenarios and case studies. Section V explores the security, privacy, and ethical dimensions of Edge AI. Section VI outlines major open research challenges. Section VII forecasts emerging trends and offers future outlook. Finally, Section VIII concludes the paper with a summary of insights.

2. Background and Technological Foundations

The foundation of Edge AI rests on a confluence of hardware miniaturization, algorithmic innovation, and distributed intelligence design. Unlike centralized cloud computing, Edge AI demands the ability to perform complex computations under tight constraints of power, memory, and connectivity. This section explores the key technical pillars that enable practical and efficient AI at the edge, including specialized hardware accelerators, model compression and optimization techniques, distributed and federated learning paradigms, and edge-optimized inference frameworks.

A major enabler of Edge AI is the advent of low-power, high-performance hardware designed specifically for on-device machine learning. Traditional CPUs and even general-purpose GPUs are often unsuitable for edge deployment due to high energy consumption and thermal requirements. In response, several purpose-built accelerators have been developed. Notable examples include Google's Edge TPU, which offers high-throughput inference for quantized models with ultra-low power consumption [13]; NVIDIA's Jetson Nano and Xavier modules, which bring CUDA-enabled GPU processing to embedded AI applications [14]; and Intel's Neural Compute Stick series that allow plug-and-play inference on devices without dedicated GPUs. These systems-on-chip (SoCs) integrate memory, compute, and I/O on a single board, enabling AI execution in standalone devices such as drones, cameras, or wearables [15].

Complementing hardware innovation is a suite of algorithmic techniques for adapting deep neural networks to constrained environments. Model compression strategies—such as pruning, quantization, knowledge distillation, and low-rank decomposition—are critical for reducing memory footprint and computational complexity while preserving inference accuracy. Pruning involves eliminating redundant weights or neurons from a network, thereby reducing model size and increasing sparsity [16]. Quantization, which replaces 32-bit floating point operations with 8-bit or even binary values, significantly accelerates inference and reduces energy usage with minimal loss in precision [17]. Knowledge distillation transfers learned knowledge from a large "teacher" model to a smaller "student" model, which is more suitable for edge deployment [18]. These techniques are often used in tandem and automated by neural architecture search (NAS) frameworks optimized for edge performance [19].

Beyond efficient inference, emerging use cases also require adaptive training on the edge. This is particularly relevant in scenarios where data cannot leave the device due to bandwidth constraints or privacy concerns. Federated learning (FL) has thus become a cornerstone of Edge AI. It enables collaborative training of global models across multiple devices by exchanging only model updates, not raw data [20]. Google's implementation of FL in Gboard for predictive typing is a landmark example, demonstrating real-world viability of the approach [21]. FL frameworks must handle challenges such as device heterogeneity, asynchronous updates, and non-iid data distributions. To address these, advances like federated averaging, hierarchical aggregation, and personalized federated learning are being actively explored [22].

Meanwhile, inference frameworks tailored for edge environments are simplifying deployment across diverse hardware platforms. TensorFlow Lite, PyTorch Mobile, Core ML, and ONNX Runtime all

support conversion of standard models into lightweight formats optimized for mobile and embedded execution [23]. These frameworks incorporate interpreter runtimes, hardware acceleration bindings, and model optimizers that allow developers to deploy complex models like MobileNetV3 or EfficientNet with minimal manual tuning. Model partitioning techniques are also being researched to enable hybrid edge-cloud inference pipelines, where the early layers of a deep model run on the edge and deeper layers offload to a nearby server or cloud [24].

Another enabling development is the emergence of edge orchestration platforms that manage AI workflows across fleets of heterogeneous devices. Platforms like AWS IoT Greengrass, Azure IoT Edge, and Baidu OpenEdge offer support for containerized AI modules, remote deployment, lifecycle management, and hardware abstraction. These orchestration systems allow AI developers to build and deploy scalable edge applications with versioning, rollback, and monitoring capabilities [25]. By aligning AI workloads with system resources and application constraints, such platforms contribute to greater stability, efficiency, and maintainability of edge intelligence solutions.

Lastly, the design of energy-efficient algorithms and runtime optimization is crucial for sustainable Edge AI. Power-aware scheduling, dynamic voltage and frequency scaling (DVFS), and adaptive model switching are being integrated into edge inference engines. These mechanisms allow devices to maintain real-time performance while dynamically adjusting to battery levels, temperature, and workload [26]. Research in neuromorphic computing and spiking neural networks also promises a paradigm shift by mimicking biological efficiency in event-driven processing, potentially redefining the computational foundations of future Edge AI systems [27].

In summary, the technological foundations of Edge AI encompass a synergistic blend of hardware and software innovations that jointly address the unique constraints and demands of the edge. By leveraging model compression, federated learning, inference frameworks, and specialized accelerators, developers can design scalable, responsive, and privacy-preserving AI systems that operate independently of the cloud. These foundational technologies form the bedrock upon which the diverse architectures and application ecosystems of Edge AI are constructed.

3. Architectures and Deployment Strategies

The successful deployment of Edge AI hinges not only on powerful algorithms and efficient hardware, but also on the architectural strategies used to orchestrate computation, communication, and intelligence across heterogeneous devices and network boundaries. Unlike centralized AI, which predominantly relies on monolithic cloud architectures, Edge AI introduces new paradigms of decentralized computing, often involving a spectrum of edge, fog, and cloud nodes working collaboratively. In this section, we explore prominent Edge AI system architectures, deployment strategies, and the key trade-offs involved in managing resources, latency, privacy, and model complexity.

Edge AI systems typically adopt one of three architectural strategies: on-device inference, near-edge inference, and edge-cloud collaboration. In on-device inference, the complete AI model is executed directly on the end device (e.g., mobile phone, smart camera, wearable sensor), without relying on any external compute node. This setup offers maximum autonomy, low latency, and robust privacy preservation, but is constrained by limited compute and memory resources [28]. Conversely, near-edge inference involves offloading part or all of the model to a nearby edge server or gateway, such as a local base station or roadside unit. This balances performance with resource availability and is widely used in latency-sensitive domains like autonomous driving or industrial robotics [9]. In edge-cloud collaboration, computation is split between edge and cloud depending on context—such as current network conditions,

available bandwidth, or urgency of inference—which enables elastic scalability and supports complex deep models that may otherwise be infeasible on local devices [30].

A major challenge in Edge AI deployment is determining where and how to partition AI workloads. Model partitioning techniques divide the deep neural network into sub-models that are distributed across nodes. For example, the first few convolutional layers of a CNN may be executed on the edge to extract features, while fully connected layers are computed in the cloud for classification. This layered inference requires robust synchronization, optimized data serialization, and security protocols to avoid performance bottlenecks and data leakage [31]. Researchers have proposed optimization-based and reinforcement learning-based strategies to determine the optimal split point based on device profiling and dynamic workload estimation [32].

To facilitate these architectural designs, edge orchestration frameworks have emerged to support end-to-end deployment pipelines. Platforms like Kubernetes with KubeEdge extensions, Open Horizon, and EdgeX Foundry provide mechanisms to manage containerized AI services across distributed nodes [33]. These platforms allow for dynamic service discovery, failure recovery, model updates, and load balancing in heterogeneous edge environments. Additionally, many support policy-based scheduling mechanisms, where workloads are assigned based on energy profiles, latency budgets, or geographic constraints. This modularity and scalability are vital for industrial applications involving thousands of distributed sensors and actuators [34].

In terms of deployment workflows, a common pipeline begins with training large-scale models in the cloud using extensive datasets. These models are then compressed, quantized, and compiled for target edge hardware using toolchains like TensorFlow Lite Converter, OpenVINO, or TVM. After benchmarking and profiling, the models are deployed to edge devices using CI/CD pipelines or over-the-air (OTA) updates. Runtime monitoring and feedback mechanisms are integrated to ensure operational performance and adapt models to environmental changes or concept drift [35]. Some advanced systems support on-device fine-tuning or online learning using lightweight optimizers, although such functionality remains constrained by energy and memory limitations.

An emerging trend in deployment strategy is multi-tier edge computing, where intermediate layers (known as fog nodes) are introduced between devices and cloud. Fog nodes may reside in base stations, routers, or local servers and can serve as aggregation or pre-processing points. This hierarchical architecture reduces uplink data traffic, supports local analytics, and enhances responsiveness in real-time scenarios [36]. It also enables collaborative intelligence, where multiple edge devices share insights or ensemble predictions through fog coordination before committing a final decision [37].

Deployment of Edge AI is further complicated by device heterogeneity, particularly in large-scale networks. Devices differ in compute capabilities, memory size, network interfaces, operating systems, and even supported machine learning runtimes. To address this, model generalization and hardware abstraction layers are used. Techniques such as model binning and conditional model loading allow for the deployment of a family of models optimized for different hardware configurations. For instance, a cluster of smart cameras might include both Raspberry Pi-based devices and Jetson Nano units, each requiring a different version of the model for optimal performance [38].

Additionally, latency-aware and energy-aware deployment strategies are critical to balance performance and resource usage. For instance, edge-aware compilers can rearrange or prune model computations dynamically to reduce inference time without retraining. Systems like Alibaba's MNN and Facebook's Glow provide platform-specific optimization backends that help achieve near-optimal performance with

minimal human intervention [39]. Scheduling algorithms that factor in battery status, ambient temperature, and wireless channel quality also play a role in runtime optimization, particularly in mobile or battery-powered applications [40].

In security-critical applications, secure model deployment is essential. This includes encrypted model transmission, runtime attestation, and execution within trusted execution environments (TEEs) such as ARM TrustZone or Intel SGX. These mechanisms ensure that models and inference data are protected from tampering, reverse engineering, or unauthorized access during deployment and runtime [41].

In summary, effective Edge AI deployment requires a delicate balance between model accuracy, system latency, hardware constraints, and data privacy. The choice of architecture—whether on-device, near-edge, or hybrid—depends heavily on the application's latency sensitivity, data locality, and device capability. As edge ecosystems continue to diversify, the development of flexible, modular, and context-aware deployment frameworks will be pivotal in enabling scalable and trustworthy Edge AI applications across domains.

4. Applications of Edge AI

Edge AI has rapidly evolved from a conceptual architecture to a powerful enabler of real-world intelligent systems across diverse sectors. By embedding machine learning capabilities at or near the data source, Edge AI empowers devices to act autonomously, securely, and responsively in contexts where cloud connectivity is intermittent, privacy is critical, or real-time inference is essential. This section explores key domains where Edge AI has made significant inroads, including smart healthcare, intelligent transportation, industrial automation, environmental monitoring, and smart retail.

In the domain of healthcare, Edge AI is revolutionizing remote patient monitoring, diagnostics, and emergency response systems. Wearable devices such as smartwatches, ECG patches, and continuous glucose monitors are increasingly embedded with edge-based machine learning algorithms to detect arrhythmia, hypoglycemia, or falls in real time [42]. Unlike cloud-based health analytics, which may suffer from latency and privacy issues, edge inference allows immediate alerts and interventions, thereby improving patient outcomes. For example, Apple's Neural Engine enables on-device ECG classification in the Apple Watch, while companies like Biofourmis and Fitbit leverage edge models for personalized physiological pattern recognition [43]. Furthermore, in rural or resource-limited settings, portable ultrasound or dermatology devices powered by embedded AI offer diagnostic assistance without relying on a stable internet connection [44]. Edge AI also supports federated learning frameworks in healthcare, enabling collaborative model training across hospitals without sharing patient data directly [45].

In intelligent transportation systems, Edge AI is central to the functioning of autonomous vehicles, traffic monitoring infrastructure, and vehicle-to-everything (V2X) communication. Modern vehicles are equipped with a network of sensors—lidar, radar, cameras, GPS—all of which generate large volumes of data that must be processed with ultra-low latency. Edge processors such as NVIDIA's Drive AGX or Tesla's FSD chip perform real-time perception tasks including object detection, lane tracking, and collision avoidance [46]. Edge inference ensures that safety-critical decisions are made locally, minimizing risks from cloud delays or communication loss. Beyond vehicles, smart intersections use edge-deployed cameras and AI to detect pedestrians, optimize signal timing, and reduce congestion through adaptive traffic control [47]. In logistics, edge intelligence supports predictive maintenance,

route optimization, and fleet health monitoring by analyzing sensor data directly on delivery trucks or drones [48].

Industrial automation, often referred to as Industry 4.0, has seen dramatic gains from Edge AI in factory floors, energy plants, and logistics hubs. Smart manufacturing systems utilize edge-based predictive maintenance to identify wear and tear in machinery before failure occurs, reducing downtime and repair costs. AI models deployed on programmable logic controllers (PLCs) or industrial edge gateways monitor vibrations, temperatures, and power usage to infer operational anomalies [49]. Computer vision on the edge supports quality inspection in production lines, detecting defects or misalignments in real time. ABB, Siemens, and Bosch have integrated Edge AI into their industrial control systems to support process optimization and fault detection with minimal cloud dependence [50]. Importantly, these applications also benefit from localized data processing, which helps meet compliance requirements in sensitive industries such as pharmaceuticals or defense manufacturing.

Environmental monitoring and agriculture are further domains where Edge AI is creating transformative impact. In remote or wide-area deployments such as forests, oceans, or farmlands, it is impractical to rely solely on cloud-based systems due to limited bandwidth and power constraints. Edge-enabled sensors and drones are used to monitor air quality, detect wildfires, assess crop health, or identify illegal deforestation using onboard AI models [51]. For instance, precision agriculture platforms employ multispectral imaging and edge inference to assess plant stress, enabling targeted irrigation or fertilizer use. Drones with edge AI can autonomously detect pest infestations or nutrient deficiencies during flight, improving yield and reducing resource waste [52]. In environmental conservation, edge-based audio sensors have been used to detect endangered species or gunshots in anti-poaching efforts, minimizing reliance on manual surveillance or post-hoc data analysis [53].

In the retail sector, Edge AI enables a new level of customer engagement and operational efficiency. Smart checkout systems utilize edge vision and sensor fusion to enable cashier-less shopping experiences, as exemplified by Amazon Go stores. Cameras and shelf sensors track items picked by users in real time using local inference, without needing centralized image processing [54]. Personalized digital signage systems leverage embedded face detection and demographic analysis to tailor promotions, while in-store analytics monitor foot traffic and product interaction patterns to inform layout optimization. Inventory management also benefits from edge-enabled robotic platforms that scan shelves and flag out-of-stock items using onboard object recognition models [55].

Edge AI applications are also expanding into public safety, education, energy management, and augmented reality. Surveillance systems with edge intelligence can detect anomalous behavior or security threats without continuous video streaming. In classrooms, AI-powered edge devices can provide adaptive content or emotion-aware interaction without sending sensitive student data off-site. Smart meters and energy gateways in residential and commercial buildings use edge learning to forecast consumption patterns and enable demand-response strategies [56].

In each of these domains, the use of Edge AI delivers several recurring benefits: reduced latency, enhanced privacy, lower bandwidth usage, and increased robustness to network failures. However, these benefits must be weighed against deployment complexity, update management, and device heterogeneity. Application-specific trade-offs often determine whether edge-only, edge-cloud, or hybrid strategies are employed. Nonetheless, as hardware and software continue to mature, the scale and variety of Edge AI applications are expected to increase dramatically.

5. Security, Privacy, and Ethical Issues in Edge AI

While Edge AI offers numerous technical and societal benefits, its decentralized nature introduces significant security, privacy, and ethical challenges. Unlike centralized cloud-based architectures where resources are physically secured and centrally administered, Edge AI deployments operate in diverse, often vulnerable environments—ranging from public streets to industrial sites and personal homes. Edge devices frequently process sensitive data, such as biometric signals, location traces, or behavioral patterns, under limited computational resources, which makes them attractive targets for adversaries. In this section, we discuss the primary concerns surrounding Edge AI security and privacy, emerging solutions, and broader ethical implications.

One of the most pressing concerns in Edge AI is vulnerability to adversarial attacks. Machine learning models deployed on the edge are exposed to physical access and limited protection mechanisms, making them susceptible to both evasion and poisoning attacks. In evasion attacks, carefully crafted inputs can be designed to fool the model into making incorrect predictions—for example, slight perturbations to an image of a stop sign could cause an autonomous vehicle's object detection system to misclassify it as a speed limit sign [57]. In poisoning attacks, adversaries tamper with the training data to corrupt the learned model, which is particularly dangerous in federated learning scenarios where training occurs across distributed nodes [58]. Due to limited compute and storage on edge devices, traditional defense techniques such as adversarial training or model verification are challenging to deploy comprehensively.

To address these threats, researchers have proposed lightweight adversarial defense mechanisms tailored for resource-constrained edge environments. For instance, feature-level denoising, randomized smoothing, and quantized models have demonstrated improved robustness with minimal overhead [59]. Trusted execution environments (TEEs) such as ARM TrustZone or Intel SGX are also leveraged to protect model integrity and input data by isolating AI operations from the main operating system [60]. Secure boot and remote attestation mechanisms ensure that only authorized firmware and AI models are loaded and executed on the device [61]. Despite these innovations, ensuring comprehensive runtime protection across diverse and heterogeneous edge devices remains an open challenge.

Privacy preservation is another central concern. Unlike centralized systems where data can be protected through secure transmission and encryption, edge devices must process data locally, often without user supervision. Applications such as emotion detection, video surveillance, and health monitoring generate highly personal data that must be protected both at rest and during processing. Federated learning (FL) has emerged as a promising approach to preserve privacy by enabling decentralized model training without transmitting raw data. However, FL is not immune to privacy risks: model updates themselves may leak sensitive information through reconstruction or membership inference attacks [62]. To mitigate this, differential privacy techniques are applied to the gradient updates, adding statistical noise that obscures individual data contributions while retaining learning efficacy [63].

Homomorphic encryption and secure multi-party computation are also being explored for edge deployments, although their computational costs remain prohibitive for real-time applications. A promising direction is hybrid privacy-preserving architectures that combine on-device encryption, differential privacy, and trusted cloud aggregation to balance efficiency and protection [64]. Some commercial solutions have adopted private inference protocols where encrypted user data is fed into encrypted models, allowing computation without ever exposing the raw inputs or weights to any single party.

Beyond technical defenses, ethical considerations in Edge AI deployment have gained increasing attention. Because edge devices often operate in public or personal spaces—such as homes, hospitals, or streets—they introduce new dimensions of algorithmic accountability and data sovereignty. For instance,

smart cameras that identify individuals without explicit consent may violate privacy norms or local regulations. In many jurisdictions, data collected at the edge must comply with policies such as the General Data Protection Regulation (GDPR), which mandates data minimization, purpose limitation, and user consent [65]. Edge AI developers must implement transparency mechanisms, such as explainable AI (XAI) tools, to ensure users can understand and contest decisions made by on-device models.

Fairness is another ethical concern, particularly in applications like facial recognition or credit scoring where biased training data may lead to discriminatory outcomes. Because edge models are often trained using narrow, localized datasets, they risk encoding regional or demographic biases that generalize poorly across diverse populations [66]. Without robust feedback loops or oversight, such models could perpetuate inequity or harm marginalized groups. To address this, community-driven auditing, representative datasets, and fairness-aware model design must be incorporated into the Edge AI lifecycle from development to deployment.

Environmental ethics also play a role. Edge AI devices—especially those deployed at scale—consume energy and generate electronic waste. Sustainable design practices, including low-power hardware, recyclable enclosures, and software-based power management, are essential for minimizing the environmental footprint of pervasive edge intelligence [67].

Ultimately, the secure and ethical deployment of Edge AI requires a multidisciplinary approach involving computer scientists, ethicists, regulators, and industry stakeholders. Technical safeguards must be complemented by governance frameworks that enforce transparency, accountability, and inclusivity. Privacy-preserving machine learning, robust adversarial defenses, and fairness-aware modeling are not optional enhancements but foundational requirements for trustworthy Edge AI systems.

6. Open Research Challenges in Edge AI

Despite substantial progress in both academia and industry, Edge AI remains a rapidly evolving and technically complex field. The transition from experimental deployments to large-scale, robust, and ethically aligned systems is still constrained by multiple open research challenges. These challenges span algorithmic efficiency, hardware heterogeneity, system interoperability, long-term learning capabilities, and sustainability. Addressing these obstacles is critical for realizing the full promise of Edge AI in ubiquitous computing environments.

One of the foremost challenges is the development of ultra-efficient learning algorithms that can operate within the extreme resource constraints of edge devices. While model compression techniques such as quantization and pruning have enabled on-device inference, training on edge devices remains largely impractical. Existing edge learning paradigms often require offloading to cloud or fog nodes, limiting their adaptability in privacy-sensitive or connectivity-constrained scenarios. Techniques such as few-shot learning, incremental learning, and continual learning have been proposed to reduce training dependency on large labeled datasets and centralized computation [68]. However, these methods are still computationally intensive and require further optimization to be feasible on low-power microcontrollers or wearable devices.

Another critical challenge is achieving reliable interoperability across heterogeneous edge environments. The diversity of edge devices—differing in processing power, operating systems, memory capacity, and network protocols—creates substantial difficulties in designing and deploying AI models that can function uniformly. Current edge AI frameworks often require manual tuning and custom compilation for different hardware targets, which is not scalable for global deployments. Standardization of AI

model representation (e.g., ONNX), edge runtime environments, and communication protocols is needed to promote portability and reduce fragmentation [69]. Furthermore, platform-agnostic neural architecture search (NAS) tools could play a pivotal role in automatically generating optimal model variants for different edge platforms [70].

The issue of dynamic model adaptation is another pressing concern. Edge devices often operate in non-stationary environments with changing lighting conditions, sensor noise, or user behaviors. However, most deployed AI models are static and struggle with concept drift, leading to degraded performance over time. Online and continual learning at the edge—where the model can evolve based on incoming data—offers a potential solution but introduces risks of catastrophic forgetting and instability [71]. Solutions such as elastic weight consolidation, memory replay, and meta-learning have been proposed, yet they remain largely experimental in edge contexts due to memory and compute limitations [72]. A hybrid approach, where only critical layers or task-specific components are updated on the edge while the base model is periodically refreshed from the cloud, may offer a pragmatic compromise.

Collaboration between edge nodes, or federated intelligence, is an emerging but under-explored frontier. In current federated learning setups, communication typically flows between the server and clients in a star topology. However, direct peer-to-peer learning between edge devices could reduce latency and enhance scalability in dense deployments such as smart cities or sensor grids [73]. Realizing this vision requires robust consensus protocols, decentralized model aggregation schemes, and security mechanisms to handle untrusted or malicious nodes. Blockchain-based trust systems and gossip-based gradient diffusion are among the proposed approaches, though practical implementations remain nascent [74].

Another major research gap lies in energy-aware AI design and execution. Power consumption is a critical constraint for mobile and remote edge devices, many of which operate on batteries or energy harvesting systems. While some runtime optimizations such as dynamic voltage and frequency scaling (DVFS) have been integrated into edge AI platforms, holistic solutions that jointly optimize model structure, hardware scheduling, and network transmission are still lacking. The concept of energy-proportional AI—where computation cost is directly aligned with task complexity and device state—is gaining traction but requires advanced resource profiling and real-time adaptation strategies [75].

Security remains a persistent challenge, particularly in hostile or uncontrolled edge environments. Lightweight cryptography, secure boot chains, and tamper detection must be integrated into AI pipelines without impairing latency or throughput. Furthermore, most current security mechanisms are reactive; future edge systems must proactively anticipate threats using anomaly detection or self-healing architectures. Zero-trust models, where every device and data source must continuously authenticate and prove compliance, could provide a higher baseline of security but may also increase system complexity [76].

From a software engineering perspective, debugging and testing Edge AI systems is inherently more difficult than centralized systems. The distributed nature, lack of real-time observability, and variability of environmental conditions make traditional unit testing or regression testing insufficient. Simulation environments that emulate heterogeneous edge deployments and allow for large-scale, reproducible testing are urgently needed. Some recent work in digital twins and edge emulation platforms such as EdgeSim has laid foundational tools, but comprehensive toolchains remain underdeveloped [77].

Finally, sustainability and lifecycle management present both ethical and operational challenges. As Edge AI devices proliferate across urban and rural landscapes, concerns around electronic waste, hardware obsolescence, and carbon footprint become increasingly relevant. Sustainable AI design must

extend beyond energy efficiency to encompass hardware recycling, software modularity, and environmental impact assessments. Lifecycle-aware AI, where models are designed to degrade gracefully or be retrained with minimal disruption, is a promising but underexplored avenue [78].

In summary, the road to robust and scalable Edge AI involves surmounting deep technical, infrastructural, and ethical hurdles. The interplay between efficient computation, privacy, learning adaptability, and long-term sustainability must be addressed through interdisciplinary research and system-level innovation. Continued progress in these areas will determine whether Edge AI fulfills its vision of delivering ubiquitous, intelligent services at scale.

7. Future Trends and Outlook in Edge AI

As Edge AI continues to mature and penetrate a broader range of applications, its evolution is increasingly influenced by emerging technologies, new communication infrastructures, and shifting societal expectations. Looking forward, several converging trends are poised to redefine the capabilities and scope of Edge AI—from the integration with 6G networks and the rise of foundation models to collaborative intelligence and the democratization of AI development. This section outlines key trajectories that are expected to shape the next decade of Edge AI innovation.

One of the most transformative trends is the convergence of Edge AI and next-generation communication systems, particularly 6G. While 5G has already facilitated significant reductions in network latency and increased bandwidth, 6G is expected to further enhance ultra-reliable low-latency communications (URLLC), support massive machine-type communications (mMTC), and integrate native AI functionalities into network layers [79]. This tight coupling of connectivity and intelligence will enable intelligent edge devices to dynamically offload, collaborate, or cache based on network context and task requirements. For instance, edge nodes in a vehicular network could anticipate network congestion and proactively switch to peer-to-peer inference mode, while drones could offload non-critical visual data to cloudlets when bandwidth is abundant. The 6G vision includes AI-as-a-Service (AIaaS) at the network edge, where lightweight models are provisioned on demand based on user intent, task context, and resource availability [80].

Another anticipated development is the miniaturization and edge adaptation of foundation models. Foundation models—large-scale, pre-trained AI systems like GPT, BERT, or CLIP—have demonstrated unprecedented generalization across tasks and modalities. However, their deployment has been confined to powerful cloud infrastructures due to their immense resource requirements. Recent research efforts are exploring the distillation, quantization, and modularization of these models for edge usage. Techniques such as LoRA (Low-Rank Adaptation) and Mixture-of-Experts (MoE) enable partial specialization of sub-modules while retaining general capabilities [3]. For instance, a miniaturized visual-language model could be deployed on an augmented reality headset to support multi-modal interaction and contextual understanding without relying on the cloud. The emergence of foundation models at the edge will expand the repertoire of on-device tasks, from multilingual translation to semantic search and commonsense reasoning [81].

Collaborative intelligence—where multiple edge agents operate in coordination—will become increasingly relevant, especially in environments involving fleets of devices such as autonomous vehicles, smart city sensors, or industrial robots. Rather than operating in isolation, edge agents will share partial knowledge, predictions, or representations with neighboring nodes to enhance accuracy and robustness. This paradigm requires new frameworks for distributed consensus, cross-device knowledge

distillation, and incentive-driven data sharing [82]. Technologies such as federated multi-agent reinforcement learning and swarm learning will be instrumental in orchestrating cooperative intelligence while preserving privacy and autonomy. In critical applications like disaster response or battlefield awareness, such collaboration could significantly enhance situational awareness and decision-making speed [83].

The democratization of Edge AI development is another accelerating trend. With the growth of open-source AI model zoos, drag-and-drop visual programming tools, and automated model optimization frameworks, the barrier to entry for developing edge applications is lowering. Platforms like Edge Impulse, MediaPipe, and TinyML democratize access to edge-optimized ML pipelines, enabling developers without deep AI expertise to build, deploy, and monitor models on microcontrollers and embedded systems [84]. This democratization is further supported by AI compilers and runtime systems (e.g., TVM, Glow) that automate hardware-specific optimization, reducing the need for manual tuning. The result is an ecosystem where startups, hobbyists, and non-technical domain experts can contribute to AI innovation at the edge.

Furthermore, ethical and responsible AI practices are expected to become embedded by design in edge systems. Rather than being bolted on post-deployment, transparency, fairness, and privacy considerations will be integrated into the full AI lifecycle—from data collection and model training to deployment and feedback. Explainable AI (XAI) modules optimized for edge constraints will help users understand and interpret model decisions locally, while differential privacy and federated analytics will protect user data by design. The growing emphasis on AI auditing, certification, and compliance—particularly in regulated sectors such as finance and healthcare—will extend to edge environments as well, encouraging the development of verifiable, low-trust, and regulation-ready models [85].

Finally, new computing paradigms such as neuromorphic computing, in-memory computing, and photonic processors are being explored to overcome current limitations in energy efficiency and latency. Neuromorphic chips like Intel's Loihi 2 aim to mimic brain-like event-driven computation, enabling highly parallel and ultra-low-power inference suitable for edge devices [86]. In-memory computing, which performs computation directly within memory arrays, reduces data movement and accelerates AI workloads significantly. These hardware innovations, when paired with edge-native AI models, could redefine the performance envelope of embedded intelligence and support new classes of applications in always-on, ultra-constrained environments.

In summary, the future of Edge AI is characterized by increasing generality, decentralization, and intelligence at scale. The integration of edge computing with advanced communication networks, scalable foundation models, and human-centered ethical principles will empower a new generation of autonomous, secure, and adaptable intelligent systems. As research and industry converge around these goals, Edge AI will likely emerge as a foundational layer of the ambient, context-aware computing landscape envisioned for the 2030s.

8. Conclusion

Edge Artificial Intelligence (Edge AI) has emerged as a pivotal paradigm in the evolution of distributed intelligent systems. Driven by the need for low-latency inference, real-time decision-making, enhanced data privacy, and localized autonomy, Edge AI has steadily transitioned from a theoretical framework to a cornerstone of modern computing architectures. This survey has provided a comprehensive overview of the fundamental technologies, system architectures, application domains, security considerations,

research challenges, and future directions in Edge AI, aiming to offer an integrative understanding of the field's current landscape and future trajectory.

At the technological core, Edge AI is supported by a rich ecosystem of specialized hardware accelerators, model compression techniques, federated and distributed learning frameworks, and inference runtimes optimized for deployment on constrained devices. These innovations have enabled practical applications in healthcare, transportation, manufacturing, retail, environmental monitoring, and beyond. In each domain, Edge AI contributes by reducing cloud dependency, preserving privacy, enabling contextual intelligence, and enhancing system robustness.

However, the deployment and operation of Edge AI systems are not without significant challenges. Security vulnerabilities, such as adversarial threats and model inversion attacks, demand the development of lightweight and robust defense mechanisms suited for edge constraints. Privacy-preserving learning techniques, including federated learning and differential privacy, must be further refined to balance utility and protection. Ethical issues, such as algorithmic fairness, explainability, and data ownership, are amplified in edge contexts where decisions are made in close proximity to end users. These challenges require not only technical solutions but also new governance frameworks and interdisciplinary collaboration.

From a research perspective, several open problems remain unsolved. These include building energy-aware and continually adaptable models, achieving reliable interoperability across heterogeneous edge environments, and designing secure yet efficient learning protocols for dynamic and large-scale edge deployments. Furthermore, testing and maintaining Edge AI systems at scale continues to be an operational bottleneck, highlighting the need for advanced simulation and emulation platforms.

Looking ahead, the integration of Edge AI with emerging technologies such as 6G networks, neuromorphic computing, and foundation model distillation is likely to define the next era of edge intelligence. Concepts such as collaborative intelligence, context-aware model provisioning, and AI-as-a-Service at the edge will unlock new capabilities and applications that transcend the limitations of current architectures. Additionally, the democratization of Edge AI development through low-code tools, open-source platforms, and automated deployment pipelines will ensure wider participation in this technological transformation.

In conclusion, Edge AI represents not just a shift in computation locality, but a broader paradigm shift in how intelligent systems are designed, deployed, and experienced. It embodies a vision of decentralized, sustainable, and human-centric intelligence that operates at the periphery of networks—closer to the data, the user, and the environment. To realize this vision, continuous innovation, rigorous standardization, and ethical foresight are imperative. As the field matures, Edge AI is poised to become a foundational layer of the ambient, intelligent infrastructure that will power the next generation of smart societies.

References

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.

[2] H. Li, K. Ota, and M. Dong, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," IEEE Network, vol. 32, no. 1, pp. 96–101, Jan. 2018.

[3] A. S. Anwar, M. R. Asghar, A. G. Abdullah, and I. Gondal, "A Review on Edge Computing-Based Smart Healthcare System: Challenges and Open Issues," in IEEE Access, vol. 10, pp. 18962–18983, 2022.

[4] M. A. Rahman et al., "A Survey on Privacy-Preserving Machine Learning for Edge Computing," ACM Computing Surveys (CSUR), vol. 55, no. 4, pp. 1–37, Jun. 2023.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE CVPR, 2018, pp. 4510–4520.

[6] Google Coral, "Edge TPU: Purpose-built ASIC designed to run AI at the edge," [Online]. Available: https://coral.ai/docs/edgetpu/faq/.

[7] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. AISTATS, 2017, pp. 1273–1282.

[8] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322–2358, 2017.

[9] ONNX Runtime, "ONNX Runtime: cross-platform, high performance scoring engine for ML models," Microsoft, 2023. [Online]. Available: https://onnxruntime.ai/

[10] H. D. Schotten et al., "6G: Vision, Use Cases and Technologies," in Proc. IEEE 5G World Forum (5GWF), 2020, pp. 1–6.

[11] N. Papernot et al., "The Limitations of Deep Learning in Adversarial Settings," in Proc. IEEE EuroS&P, 2016, pp. 372–387.

[12] Y. Wang et al., "Edge Intelligence: Security and Privacy Issues," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4312–4321, May 2020.

[13] Google Coral, "Edge TPU: Purpose-built ASIC designed to run AI at the edge," [Online]. Available: https://coral.ai/docs/edgetpu/faq/.

[14] NVIDIA, "Jetson Nano Developer Kit," [Online]. Available: https://developer.nvidia.com/embedded/jetson-nano-developer-kit.

[15] A. K. Das, P. Deka, and B. P. Sahu, "A Review on Edge Computing Based Intelligent Systems," in Proc. IEEE ICCCNT, 2021, pp. 1–6.

[16] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," in Proc. ICLR, 2016.

[17] Y. Choi, M. El-Khamy, and J. Lee, "Towards the Limit of Network Quantization," in Proc. ICLR, 2017.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531, 2015.

[19] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, 2019, pp. 6105–6114.

[20] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. AISTATS, 2017, pp. 1273–1282.

[21] H. B. McMahan and D. Ramage, "Federated Learning: Collaborative Machine Learning without Centralized Training Data," Google Research Blog, Apr. 2017.

[22] T. Li, A. S. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

[23] TensorFlow, "TensorFlow Lite," [Online]. Available: https://www.tensorflow.org/lite.

[24] H. Zhang, Y. Wang, and D. Niyato, "Joint Optimization of DNN Partitioning and Deployment in Edge-Cloud Systems," in IEEE Transactions on Mobile Computing, vol. 21, no. 2, pp. 455–471, Feb. 2022.

[25] Amazon Web Services, "AWS IoT Greengrass," [Online]. Available: https://aws.amazon.com/greengrass/.

[26] A. Mittal, S. P. Sahu, and P. K. Jana, "Energy-Efficient Scheduling of Edge Inference Tasks Using DVFS," in Proc. IEEE ISCC, 2022, pp. 1–7.

[27] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, vol. 38, no. 1, pp. 82–99, Jan. 2018.

[28] A. Lane, S. Bhattacharya, A. Mathur, and A. Seneviratne, "Can Deep Learning Revolutionize Mobile Sensing?" in Proc. ACM Workshop on Hot Topics in Wireless, 2015, pp. 25–30.

[29] M. Satyanarayanan, P. Simoens, Y. Xiao, and S. Hu, "Edge Analytics in the Internet of Things," IEEE Pervasive Computing, vol. 14, no. 2, pp. 24–31, Apr. 2015.

[30] T. Taleb, A. Ksentini, and R. Jantti, "Anything as a Service for 5G Mobile Systems," IEEE Network, vol. 30, no. 6, pp. 84–91, Nov.–Dec. 2016.

[31] H. Zhang, Y. Wang, and D. Niyato, "Joint Optimization of DNN Partitioning and Deployment in Edge-Cloud Systems," IEEE Transactions on Mobile Computing, vol. 21, no. 2, pp. 455–471, Feb. 2022.

[32] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," in Proc. ACM ASPLOS, 2017, pp. 615–629.

[33] KubeEdge, "Kubernetes Native Edge Computing Framework," [Online]. Available: https://kubeedge.io/en/

[34] EdgeX Foundry, "An Open Platform for the IoT Edge," [Online]. Available: https://www.edgexfoundry.org/

[35] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, "DeepDecision: A Mobile Deep Learning Framework for Edge Video Analytics," in Proc. IEEE INFOCOM, 2018, pp. 1421–1429.

[36] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," IEEE Internet of Things Journal, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[37] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7457–7469, Aug. 2020.

[38] T. Wang et al., "Device-Aware Neural Architecture Search for Edge Devices," in Proc. IEEE CVPR, 2020, pp. 13398–13407.

[39] Alibaba, "MNN: Mobile Neural Network Engine," [Online]. Available: https://github.com/alibaba/MNN

[40] M. Kim, Y. Park, H. Lee, and J. Kim, "Energy-Aware Deep Learning Model Deployment for Battery-Operated Devices," IEEE Access, vol. 9, pp. 52831–52842, 2021.

[41] X. Feng, Q. Pei, J. Liu, and W. Shi, "SecureML: Confidential Machine Learning on Trusted Processors," in Proc. IEEE EuroS&P, 2020, pp. 492–507.

[42] A. M. Rahmani et al., "Exploiting Smart E-Health Gateways at the Edge of Healthcare Internet-of-Things: A Fog Computing Approach," Future Generation Computer Systems, vol. 78, pp. 641–658, Jan. 2018.

[43] Apple Inc., "Apple Watch ECG App and Irregular Rhythm Notification Available Today," Press Release, Dec. 2018. [Online]. Available: https://www.apple.com/newsroom/

[44] M. Hassan, M. I. Ur Rehman, and N. Ahmad, "AI-Based Diagnostic Systems for Rural Health Monitoring Using Edge Devices," in Proc. IEEE ICIOT, 2022, pp. 45–50.

[45] S. Rieke et al., "The Future of Digital Health with Federated Learning," NPJ Digital Medicine, vol. 3, no. 1, pp. 1–7, 2020.

[46] NVIDIA, "NVIDIA DRIVE AGX Platform for Autonomous Vehicles," [Online]. Available: https://developer.nvidia.com/drive/drive-platform

[47] J. Contreras-Castillo et al., "Intelligent Transportation Systems with Connected Vehicle Proactive Driving Decisions Using Edge Computing and Deep Learning," Future Generation Computer Systems, vol. 100, pp. 102–115, Nov. 2019.

[48] R. D. Gutiérrez and J. A. García-Macías, "Real-Time Logistics Tracking with Edge AI and IoT," in Proc. IEEE GLOBECOM, 2021, pp. 1–6.

[49] T. Zhang, J. Tan, L. Wang, and D. Jiang, "Edge AI for Predictive Maintenance in Smart Factories," in Proc. IEEE ICIEA, 2020, pp. 1546–1551.

[50] Bosch, "Bosch Edge AI Solutions for Industrial Automation," White Paper, 2022. [Online]. Available: https://www.bosch.com/research/

[51] S. R. Jadhav and S. S. Pethakar, "AI-Enabled Environmental Monitoring Using Edge Computing," in Proc. IEEE ICCCIS, 2021, pp. 207–212.

[52] K. J. Shankar, V. Vijayalakshmi, and L. J. Deborah, "Edge Intelligence for Precision Agriculture: A Review," in Computers and Electronics in Agriculture, vol. 194, p. 106653, Apr. 2022.

[53] J. A. Stowell et al., "Acoustic Monitoring of Biodiversity Using Edge Devices in Tropical Rainforests," Ecological Indicators, vol. 125, p. 107529, 2021.

[54] Amazon, "Amazon Go and the Smart Store of the Future," [Online]. Available: https://www.amazon.com/b?node=16008589011

[55] L. Zhao, M. Xu, and W. Zuo, "Edge-Based Inventory Monitoring in Smart Retail Systems," in Proc. IEEE ISCAS, 2022, pp. 2111–2114.

[56] Y. Zhou, M. Chen, and T. Q. S. Quek, "Edge Intelligence for Smart Buildings: Architectures, Algorithms, and Case Studies," IEEE Wireless Communications, vol. 28, no. 3, pp. 108–115, June 2021.

[57] K. Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," in Proc. IEEE CVPR, 2018, pp. 1625–1634.

[58] X. Sun, J. Zhang, and C. Zhang, "Poisoning Attacks on Federated Learning: Challenges and Opportunities," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 7303–7314, Dec. 2022.

[59] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified Adversarial Robustness via Randomized Smoothing," in Proc. ICML, 2019, pp. 1310–1320.

[60] F. Brasser et al., "Advancing Security for Emerging IoT Applications Using Trusted Execution Environments," in Proc. USENIX Security Symposium, 2015, pp. 33–47.

[61] H. Shacham and B. Waters, "Compact Proofs of Retrievability," in Proc. ASIACRYPT, 2008, pp. 90–107.

[62] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks Against Centralized and Federated Learning," in Proc. IEEE S&P, 2019, pp. 739–753.

[63] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

[64] S. Mishra, N. Sood, and A. Sharma, "Hybrid Privacy-Preserving Architecture for Real-Time AI Applications at the Edge," IEEE Access, vol. 10, pp. 60271–60282, 2022.

[65] European Parliament and Council, "General Data Protection Regulation (GDPR)," Regulation (EU) 2016/679, Apr. 2016.

[66] T. Gebru et al., "Datasheets for Datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, Dec. 2021.

[67] Y. Chen, H. Gao, and W. Shi, "Sustainable Edge AI: Strategies and Opportunities," IEEE Computer, vol. 55, no. 7, pp. 28–36, Jul. 2022.

[68] A. Parisi, D. Kemker, J. Part, C. Kanan, and C. M. Powers, "Continual Lifelong Learning with Neural Networks: A Review," Neural Networks, vol. 113, pp. 54–71, May 2019.

[69] ONNX AI, "Open Neural Network Exchange Format," [Online]. Available: https://onnx.ai/

[70] X. Xu, Y. Xu, S. Zhang, and Y. Liu, "Device-Aware Neural Architecture Search for Edge Devices," in Proc. NeurIPS, 2020.

[71] H. Liu, J. Sim, and R. Han, "On-Device Lifelong Learning: Training Deep Neural Networks with Limited Resources," IEEE Internet of Things Journal, vol. 8, no. 4, pp. 2615–2625, Feb. 2021.

[72] J. Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks," PNAS, vol. 114, no. 13, pp. 3521–3526, 2017.

[73] Z. Ma, J. Wang, and L. Zhang, "Peer-to-Peer Federated Learning on Edge Devices: Consensus, Privacy, and Performance," IEEE Transactions on Mobile Computing, 2023.

[74] Y. Lu, L. Huang, and H. Zhu, "Blockchain and Federated Learning for Privacy-Preserved Edge Intelligence: Opportunities and Challenges," IEEE Network, vol. 35, no. 5, pp. 12–19, Sept. 2021.

[75] M. E. Fouda, A. A. El-Sherif, and S. F. El-Zoghabi, "Energy-Aware Edge AI Inference Using Dynamic Task Scheduling," IEEE Access, vol. 10, pp. 8911–8923, 2022.

[76] L. Zhang, S. Zhai, Q. Wang, and H. Chen, "A Survey on Trust Management for Edge Computing: Current Research and Future Directions," IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 446–469, 2022.

[77] J. Tang, Q. Liu, M. Li, and Y. Liu, "EdgeSim: A Simulation Framework for Collaborative Edge Computing," in Proc. IEEE ICDCS, 2021, pp. 1155–1165.

[78] A. J. Noronha and S. L. Keoh, "Sustainable AI at the Edge: Opportunities, Challenges, and Future Research Directions," ACM Computing Surveys, vol. 55, no. 3, pp. 1–33, 2023.

[79] X. You, C. Zhang, X. Tan, S. Jin, and H. Wu, "Towards 6G Wireless Communication Networks: Vision, Enabling Technologies, and New Paradigm Shifts," Science China Information Sciences, vol. 64, no. 1, pp. 1–74, Jan. 2021.

[80] A. Checko et al., "Cloud RAN for Mobile Networks—A Technology Overview," IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pp. 405–426, First Quarter 2015.

[81] E. Dettmers et al., "Efficient and Robust Large-Scale Language Model Training with 8-bit Optimizers," in Proc. ICML, 2022.

[82] P. Huang, C. Raffel, and J. Shlens, "Language Models on Edge Devices: A Survey," arXiv preprint arXiv:2303.01065, 2023.

[83] A. R. Zamir et al., "Collaborative Intelligence: A Human-AI Interaction Paradigm for Edge Devices," in Proc. NeurIPS, 2020.

[84] S. Abuadbba et al., "Real-Time and Privacy-Preserving Edge Intelligence for Collaborative Autonomous Systems," IEEE Internet of Things Journal, vol. 8, no. 3, pp. 1585–1599, Feb. 2021.

[85] Edge Impulse, "Edge ML for All Developers," [Online]. Available: https://www.edgeimpulse.com/

[86] European Commission, "AI Act: Proposal for a Regulation on a European Approach for Artificial Intelligence," Apr. 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/

[87] M. Davies et al., "Advancing Neuromorphic Computing with Loihi 2," Intel Research White Paper, 2022.