Quantum-Enhanced Edge Intelligence: Bridging Quantum Computing and Distributed AI

Hoa Tran-Dang, Dong-Seong Kim

***Abstract*—This paper explores the emerging paradigm of Quantum-Enhanced Edge Intelligence (QEEI), which envisions the integration of quantum computing capabilities with distributed AI at the edge of networks. We review foundational concepts, highlight recent advances in quantum algorithms relevant to edge intelligence, and examine potential applications across domains such as autonomous systems and industrial IoT. The paper outlines a conceptual framework for QEEI, discusses current limitations in hardware and software integration, and offers a forward-looking perspective on research challenges and opportunities at the intersection of quantum computing and edge AI.**

***Index Terms*—Quantum Computing, Edge AI, Quantum Machine Learning, Resource Optimization, Quantum Cryptography, Hybrid Quantum-Classical Systems, Real-Time Processing, Privacy-Preserving AI, Noisy Intermediate-Scale Quantum (NISQ), Task Scheduling, Autonomous Systems, High-Dimensional Data Processing**

I. Introduction

The proliferation of intelligent devices from autonomous vehicles [1] to wearable health monitors [2] and industrial IoT systems [3] has led to a new era of pervasive computing. These systems generate massive volumes of real-time data, requiring low-latency, bandwidth-efficient, and privacy preserving processing. Edge Artificial Intelligence (Edge AI) has emerged as a key solution [4], executing learning and inference directly on or near the data source [5]. This reduces communication overhead, enhances privacy, and enables timely responses in critical applications such as collision avoidance and medical diagnostics [6].

Yet, the increasing complexity of edge workloads—such as federated learning and adaptive control—pushes the limits of conventional edge hardware [7]. Devices with constrained power, memory, and compute resources often struggle to support deep learning models without compression or loss of accuracy [8]. Achieving ultra-low-latency inference for high-resolution tasks on battery-powered devices remains an open challenge [9].

Quantum computing introduces a fundamentally different computational paradigm, leveraging principles like superposition and entanglement to solve certain problems exponentially faster than classical approaches [10]. Algorithms such as Grover's search and Shor's factoring [11], as well as quantum optimization methods like QAOA [12] and VQE [13], demonstrate powerful capabilities. These quantum methods offer promising enhancements to Edge AI—accelerating learning and optimization, reducing model size, and enabling secure communications via quantum cryptography [14].

Integrating quantum computing with Edge AI—what we term Quantum-Enhanced Edge Intelligence (QEEI)—promises synergistic benefits. Quantum neural networks (QNNs) and variational quantum circuits (VQCs) may offer expressive, compact models for efficient edge inference [15]. Quantum federated learning (QFL) could improve privacy and convergence in non-IID environments [16].

However, practical challenges remain. Current Noisy Intermediate-Scale Quantum (NISQ) devices [17] are limited by low qubit counts, short coherence times, and hardware constraints unsuited for edge deployment. Software toolchains, data encoding, and hybrid orchestration are still immature.

This work aims to outline the vision and architecture of QEEI, reviewing key use cases, technical barriers, and research directions—including edge-deployable quantum hardware, quantum-aware learning, and digital twin–based co-simulation—to help bridge quantum computing and distributed AI at the edge.

II. Background

A. Edge AI Fundamentals

Edge Artificial Intelligence (Edge AI) shifts machine learning (ML) and deep learning (DL) tasks closer to data sources—on devices like smartphones, sensors, autonomous vehicles, and industrial controllers—enabling real-time, privacy-preserving, and low-latency inference [18], [6]. By minimizing reliance on cloud processing, Edge AI supports time-sensitive applications such as augmented reality, autonomous driving, and remote healthcare monitoring.

A typical Edge AI pipeline includes data acquisition, on-device preprocessing, model inference, and, increasingly, local learning. Recent advances in lightweight models (e.g., MobileNet [19], TinyML [20]) and dedicated hardware accelerators (e.g., NVIDIA Jetson, Google Coral, Apple Neural Engine [21], [22]) have enabled AI workloads on constrained devices. Federated Learning (FL) further enhances Edge AI by training models across distributed nodes without sharing raw data, preserving privacy and regulatory compliance [23], [24].

Despite these advances, Edge AI faces several key challenges:

- **Resource Constraints:** Edge devices have limited compute, memory, and power, making it difficult to run complex DL models efficiently. Tasks like object detection and speech recognition can cause latency or system failures [25], [26], [27].

- **Compression vs. Accuracy:** Techniques like quantization, pruning, and knowledge distillation reduce model size but may degrade accuracy, especially in noisy or dynamic environments [28], [29].
- **Real-Time Demands:** Safety-critical applications require deterministic responses within sub-millisecond latency, which remains difficult for classical AI engines under variable conditions [30].
- **Security and Privacy Risks:** While Edge AI mitigates centralized breaches, devices remain exposed to physical attacks, adversarial inputs, and model inversion. Implementing lightweight privacy-preserving techniques remains a major hurdle [31], [32], [33].
- **Lack of Standardization:** Diverse hardware and software platforms lack unified toolchains and benchmarks, hindering portability and scalable model deployment.

These challenges call for new computational paradigms beyond classical architectures. In this context, quantum computing emerges as a promising approach to enhance performance, efficiency, and security in future edge systems. The next section explores its foundational principles and potential synergy with Edge AI.

B. Quantum Computing Principles

Quantum computing is an emerging interdisciplinary field combining computer science, physics, and mathematics to solve problems that are computationally infeasible for classical systems. It leverages the unique properties of quantum mechanics—such as superposition, entanglement, and interference—to achieve exponential or quadratic speedups for specific tasks [10], [11].

1) Qubits and Superposition: Unlike classical bits, qubits can exist in a superposition of both 0 and 1 states, represented as $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where $|\alpha|^2 + |\beta|^2 = 1$ [10]. This allows quantum systems to explore many possibilities in parallel, enhancing computational throughput.

2) Entanglement and Interference: Entanglement enables strong correlations between qubits, allowing operations on one qubit to instantaneously affect another, regardless of distance [34]. Interference is used to amplify correct computation paths and suppress incorrect ones, improving solution accuracy in quantum algorithms.

3) Decoherence and Measurement: Quantum states are fragile and can lose coherence due to environmental noise. Mitigating decoherence requires isolating qubits and error correction techniques. Measurement collapses a qubit's state to a classical outcome, making readout inherently probabilistic.

4) Quantum Gates and Circuits: Quantum computation is performed using quantum gates—unitary transformations applied to qubits. Sequences of gates form quantum circuits that encode and execute algorithms. Common gates include Hadamard, Pauli-X, and CNOT.

5) Quantum Computing Models: Quantum computing platforms fall into three main categories, each suited to different workloads:

- **Gate-Based Systems:** These follow the quantum circuit model using universal gate sets. Examples include IBM's superconducting qubits, Google's Sycamore, and IonQ's trapped ions [35], [36]. They support general-purpose algorithms like QAOA, VQE, and quantum machine learning.
- **Quantum Annealers:** Designed for optimization problems, these systems use quantum tunneling to find low-energy states in a problem-specific Hamiltonian. D-Wave's annealers excel at solving QUBO problems in domains like logistics and finance [37].
- **Hybrid Quantum-Classical Systems:** In the current NISQ era [17], hybrid approaches combine quantum subroutines with classical control, offering practical value despite hardware limitations. Algorithms like VQE and QAOA are prominent examples [13], especially relevant for future edge applications requiring quantum acceleration within classical infrastructures.

Each model involves trade-offs in scalability, noise tolerance, and suitability for AI tasks. Understanding and aligning these models with edge computing needs is essential for realizing quantum-augmented Edge AI. Table I summarizes the key differentiate features between classical and quantum computing paradigm.

TABLE I: Classical Computing vs. Quantum Computing

Aspect	Classical Computing	Quantum Computing
Basic Unit	Bit (0 or 1)	Qubit (superposition of 0 and 1)
Information Encoding	Binary logic gates (AND, OR, NOT)	Quantum gates (Hadamard, CNOT, Pauli-X, etc.)
Parallelism	Sequential or parallel via multi-threading	Intrinsic parallelism via superposition
Correlation	Independent variables	Quantum entanglement enables correlated qubit states
Computation Model	Deterministic	Probabilistic (measurement collapses state)
Speedup Potential	Polynomial improvements with optimized algorithms	Exponential speedup for specific problems (e.g., factorization, search)
Hardware Examples	CPUs, GPUs, TPUs	Superconducting circuits, trapped ions, photonic qubits
Maturity	Commercially mature, robust	Early-stage, rapidly evolving
Error Tolerance	High precision	Susceptible to noise; requires error correction

III. Opportunities for Quantum Computing in Edge AI

A. Scalable Learning at the Edge

Scalability is a core challenge in Edge AI, particularly in large-scale distributed environments like IoT networks, smart cities, and connected autonomous vehicles. These systems require efficient, on-device learning that adapts to dynamic conditions despite constraints in compute power, energy, and bandwidth. Quantum computing offers new opportunities to alleviate these limitations by accelerating model training, improving generalization, and supporting collaborative learning frameworks under resource constraints.

B. Quantum Speedups in Model Training

Quantum algorithms promise exponential or polynomial speedups for fundamental linear algebra operations used in machine learning—such as solving linear systems (e.g., HHL algorithm), matrix inversion, and dimensionality reduction. These operations underpin many edge AI applications, including anomaly detection, condition monitoring, and real-time classification.

Quantum-native models like Variational Quantum Circuits (VQCs) and Quantum Neural Networks (QNNs) can express complex patterns with fewer parameters, reducing model size while maintaining representational power [38]. Centralized quantum processors could train these models, which are then deployed to edge devices for low-latency inference, shifting computational burdens away from constrained endpoints.

C. Federated and Distributed Quantum Learning

Federated learning enables distributed training across edge devices while preserving data privacy. In this context, Quantum Federated Learning (QFL) has emerged as a promising enhancement. By transmitting quantum-encoded parameters instead of gradients, QFL can improve convergence and reduce communication overhead—particularly in non-IID data settings common and further enhance privacy at the edge [39], either by:

- Encoding gradients in quantum states that are harder to intercept or reconstruct.
- Using quantum differential privacy techniques for model parameter sharing.
- Employing entangled qubit protocols to ensure verification and secure aggregation across devices.

Further, theoretical proposals suggest that entangled qubits could enable synchronized decision-making across distributed nodes without direct communication, potentially reducing coordination latency in collaborative learning. In addition, quantum privacy guarantees could make FL more robust to adversarial inference attacks, model poisoning, and membership inference—common threats in edge-AI systems deployed in untrusted environments.

D. Quantum-Inspired Compression and Knowledge Distillation

Quantum computing's efficient state representation has inspired novel approaches to neural network compression. Techniques such as quantum distillation could allow knowledge transfer from quantum-trained models to lightweight edge models, minimizing inference costs while retaining performance. These ideas may also inform new strategies for pruning and quantizing deep networks deployed on constrained hardware as followings:

- Quantum-aware edge model design: Developing hybrid model architectures that explicitly separate quantum-pretrainable and edge-deployable components.
- Edge-native quantum co-processors: As quantum hardware miniaturizes, edge devices equipped with small-scale quantum processors (e.g., quantum photonics or trapped-ion chips) may handle specific learning sub-tasks locally.
- Scalable quantum feature maps: Applying quantum kernels or feature maps to accelerate tasks like anomaly detection, image recognition, or predictive modeling at the edge.

E. Efficient Resource Allocation and Scheduling

Edge AI systems operate under stringent constraints—limited compute power, volatile energy budgets, real-time processing demands, and unreliable connectivity. Efficient resource management, including computation offloading, task scheduling, bandwidth allocation, and energy optimization, is thus essential. Quantum computing introduces new paradigms for tackling these challenges through enhanced optimization capabilities, parallelism, and probabilistic modeling.

F. Quantum Optimization for Resource Allocation

Classical resource allocation algorithms often rely on heuristic or approximate solutions to NP-hard problems (e.g., task offloading, job scheduling, or multi-agent coordination). Quantum computing provides new algorithmic tools—most notably, Quantum Approximate Optimization Algorithm (QAOA) and Quantum Annealing—that can address combinatorial optimization more efficiently than classical methods under certain conditions [40], [41]. In an Edge AI context, QAOA can be applied to:

- Task-to-node matching, optimizing utility under delay or energy constraints.
- Bandwidth and spectrum allocation, maximizing throughput across edge links.
- Latency-aware scheduling, especially for multi-tier fog and vehicular edge networks.

These approaches could be implemented via cloud-based quantum processors acting as optimization co-processors for edge orchestrators or micro data centers.

G. Quantum Annealers for Real-Time Decision Making

Quantum annealers, such as those developed by D-Wave, are well-suited for solving Quadratic Unconstrained Binary Optimization (QUBO) problems, which model many edge resource scheduling tasks. For instance:

- Optimal placement of AI inference tasks across heterogeneous edge nodes.
- Scheduling sensor activations in large-scale IoT networks.
- Dynamic allocation of compute units to edge functions in microservice-based architectures.

Recent work has shown that annealing-based approaches can yield faster convergence and better-quality solutions in real-time resource scheduling compared to classical solvers [42], although current quantum annealers still face scale and noise limitations.

H. Security and Privacy

As Edge AI systems increasingly handle sensitive and personal data—such as biometric identifiers, health metrics, and location traces—ensuring data privacy becomes a critical requirement. While classical encryption methods can offer strong protection, they are vulnerable to future quantum attacks due to Shor's algorithm, which enables efficient factorization of large integers and threatens RSA and ECC-based cryptographic protocols [43]. Conversely, quantum technologies also offer powerful tools to enhance security and privacy, particularly in decentralized Edge AI environments.

I. Quantum-Secure Communication for Edge Networks

Quantum Key Distribution (QKD) provides provably secure communication by relying on the physical principles of quantum mechanics rather than mathematical complexity. Protocols like BB84 ensure that any eavesdropping attempt alters the quantum states being measured, thus allowing intrusion detection and secure key exchange.

In edge computing, QKD could be used to:

- Secure communication between edge devices and fog nodes.
- Establish symmetric keys for encrypting model updates in federated learning.
- Protect data during task offloading or model inference over 5G/6G links.

While practical QKD networks still face challenges in range and hardware integration, integrated photonic QKD chips are emerging as a compact, energy-efficient solution, opening the door to secure quantum-enhanced edge networks [44].

J. Post-Quantum Cryptography for Edge AI

Even in the absence of full-scale quantum networks, post-quantum cryptography (PQC) is an important bridge for securing Edge AI systems today. PQC algorithms—such as lattice-based and hash-based cryptosystems—are resistant to known quantum attacks and can be integrated into edge protocols for secure bootstrapping, remote attestation, and secure AI model provisioning [45].

K. Hybrid Quantum-Classical Edge Architectures

The inherent limitations of current quantum hardware—such as noise, decoherence, limited qubit count, and cryogenic requirements—make fully quantum edge systems impractical in the near term. However, the hybrid quantum-classical paradigm, where quantum and classical processors collaborate to execute AI tasks, offers a highly promising architectural framework for enhancing Edge AI systems in realistic settings as shown in Fig. 1.

In such systems:

- Edge devices collect and preprocess contextual information.
- A centralized or near-edge quantum unit solves the high-complexity optimization (e.g., offloading plan).
- The output is returned to edge nodes as control policies or schedules.

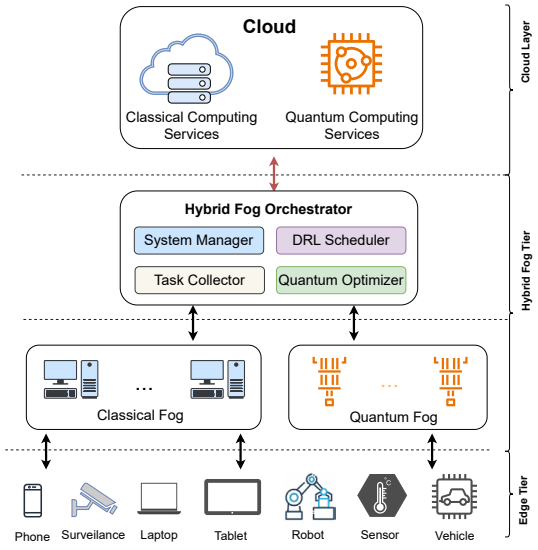

Fig. 1: A hybrid quantum-classical computing architecture for multi-edge/fog computing systems

This model can enable adaptive, globally optimized decision-making while keeping inference and sensor fusion tasks local to the edge.

IV. System Design and Integration Challenges

Integrating quantum computing with Edge AI offers transformative potential but faces significant technical and systemic hurdles across hardware, software, and deployment layers.

A. Hardware Limitations and Scalability

Current quantum processors suffer from limited qubit counts, short coherence times, and high error rates. Their reliance on cryogenic cooling and sensitive calibration poses severe obstacles for deployment in mobile, power-constrained, or uncontrolled edge environments. Even NISQ devices, while promising for near-term research, remain unsuitable for real-time edge applications due to fragility and dependence on cloud-based infrastructure.

B. Integration Complexity with Edge AI Pipelines

Edge AI relies on heterogeneous classical hardware (CPUs, GPUs, NPUs) and containerized frameworks (TensorFlow Lite, ONNX), managed via orchestration tools like Kubernetes. Integrating quantum modules requires:

- Quantum-classical interfaces capable of dynamic task offloading and switching.
- Standardized quantum programming models compatible with diverse runtimes.
- Cross-compilers and hybrid execution environments for joint quantum-classical workloads [46].

Currently, limited interoperability between mainstream AI frameworks and quantum SDKs (Qiskit, PennyLane, Cirq) hinders seamless integration.

C. Latency and Communication Bottlenecks

Quantum computation is often cloud-offloaded, introducing latency incompatible with time-critical edge applications like autonomous vehicles or medical monitoring. Additional challenges include:

- Network delays and congestion impacting reliability.
- Secure communication protocols adding overhead.
- Limited bandwidth at remote or mobile edge sites constraining quantum service access.

D. Resource Management in Heterogeneous Edge Environments

Quantum processors present unique challenges:

- Non-deterministic execution times.
- Frequent recalibration and cooldown requirements.
- Restricted, shared access windows in cloud-based services.

Effective scheduling and load balancing require novel quantum-aware edge resource management algorithms to coordinate hybrid workloads.

E. Security and Trust in Hybrid Architectures

Hybrid quantum-edge systems introduce new vulnerabilities:

- Attack surfaces at quantum-classical interfaces.
- Side-channel attacks on classical controllers.
- Trust and verification issues when relying on remote quantum backends for critical AI decisions.

Developing trust frameworks, secure execution containers, and audit mechanisms is vital for deployment.

F. Noise and Error Mitigation

Quantum processors in the current NISQ (Noisy Intermediate-Scale Quantum) era suffer from limited coherence times, gate infidelities, and readout errors, posing reliability challenges for quantum-assisted edge computing. These limitations hinder the accurate execution of quantum algorithms, particularly for real-time or mission-critical applications.

To address this, emerging error mitigation techniques—such as zero-noise extrapolation, noise-aware circuit compilation, and variational error suppression—offer partial solutions without requiring full error correction. Hybrid quantum-classical models can also shift critical logic to classical hardware, improving robustness. Going forward, lightweight, task-specific mitigation methods tailored to edge deployments will be essential for practical use.

G. Interoperability Standards

Quantum-classical integration at the edge remains fragmented due to a lack of standardized protocols and unified software-hardware interfaces. The heterogeneity of edge devices, combined with the nascent state of quantum hardware, complicates seamless deployment and coordination.

Standardization is needed across execution models, communication protocols, and resource abstractions to enable scalable and platform-agnostic integration. Initiatives for defining open APIs, edge-quantum runtime environments, and co-processing interfaces will be vital for the future of distributed hybrid AI systems. Industry-wide efforts analogous to standards in AI (e.g., ONNX) and IoT (e.g., MQTT) are key enablers for this vision.

V. Conclusion

The fusion of quantum computing and Edge AI holds immense promise for overcoming the inherent limitations of classical edge systems, enabling scalable, efficient, and secure intelligence at the network periphery. While current quantum hardware and integration challenges remain significant barriers, ongoing advances in quantum algorithms, miniaturized hardware, and hybrid quantum-classical architectures are paving the way toward practical deployments. This paper has outlined the foundational principles, key opportunities, and critical challenges that define this emerging field. Moving forward, interdisciplinary research across hardware innovation, algorithm design, software engineering, and systems integration will be essential to fully realize the transformative potential of quantum-enhanced Edge AI. By bridging these domains, future edge computing platforms can achieve unprecedented levels of adaptability, performance, and security in increasingly complex and distributed environments.

Acknowledgments

This work was supported in part by the Ministry of Science and ICT (MSIT), Korea, under the Innovative Human Resource Development for Local Intellectualization program, supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) (IITP-2025-RS-2020-II201612 (25%), IITP-2025-RS-2024-00438430 (25%)) and Korea Research Fellowship Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2018R1A6A1A03024003 (25%), RS-2023-00249687 (25%)).

References

[1] G. Yan, K. Liu, C. Liu, and J. Zhang, "Edge intelligence for internet of vehicles: A survey," *IEEE Transactions on Consumer Electronics*.

[2] T. Azzino, M. Mezzavilla, S. Rangan, Y. Wang, and J.-R. Rizzo, "5g edge vision: Wearable assistive technology for people with blindness and low vision," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, pp. 1–6.

[3] T. Qiu, J. Chi, X. Zhou, Z. Ning, M. Atiquzzaman, and D. O. Wu, "Edge computing in industrial internet of things: Architecture, advances and challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2462–2488.

[4] 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.

[5] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646.

[6] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762.

[7] 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.

[8] G. K. Walia, M. Kumar, and S. S. Gill, "Ai-empowered fog/edge resource management for iot applications: A comprehensive review, research challenges, and future perspectives," vol. 26, no. 1, pp. 619–669.

[9] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191.

[10] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*, ser. Cambridge Series on Information and the Natural Sciences. Cambridge University Press.

[11] P. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pp. 124–134.

[12] E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm." [Online]. Available: https://arxiv.org/abs/1411.4028

[13] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, "A variational eigenvalue solver on a photonic quantum processor," *Nature Communications*, vol. 5, no. 1.

[14] D. Rusca and N. Gisin, "Quantum cryptography: an overview of quantum key distribution." [Online]. Available: https://arxiv.org/abs/2411.04044

[15] A. Kottahachchi Kankanamge Don and I. Khalil, "Q-supcon: Quantum-enhanced supervised contrastive learning architecture within the representation learning framework," *ACM Transactions on Quantum Computing*, vol. 6, no. 1, pp. 1–24.

[16] J. Pei, L. Wang, N. Awan, and R. Alturki, "Advancing federated learning privacy with quantum communication techniques: A robust scalable framework," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 11, no. 2, pp. 51–58.

[17] J. Preskill, "Quantum computing in the nisq era and beyond," *Bulletin of the American Physical Society*, vol. 64, p. 9.

[18] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646.

[19] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*.

[20] P. Warden and D. Situnayake, *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-low-power Microcontrollers*. O'Reilly. [Online]. Available: https://books.google.co.kr/books?id=sB3mxQEACAAJ

[21] G. Akkad, A. Mansour, and E. Inaty, "Embedded deep learning accelerators: A survey on recent advances," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 5, pp. 1954–1972.

[22] C. Silvano, D. Ielmini, F. Ferrandi, L. Fiorin, S. Curzel, L. Benini, F. Conti, A. Garofalo, C. Zambelli, E. Calore, S. Schifano, M. Palesi, G. Ascia, D. Patti, N. Petra, D. De Caro, L. Lavagno, T. Urso, V. Cardellini, G. C. Cardarilli, R. Birke, and S. Perri, "A survey on deep learning hardware accelerators for heterogeneous hpc platforms," *ACM Computing Surveys*, vol. 57, no. 11, pp. 1–39.

[23] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, pp. 1273–1282.

[24] O. R. A. Almanifi, C.-O. Chow, M.-L. Tham, J. H. Chuah, and J. Kanesan, "Communication and computation efficiency in federated learning: A survey," *Internet of Things*, vol. 22, p. 100742.

[25] T. Huang, T. Luo, M. Yan, J. T. Zhou, and R. Goh, "Rct: Resource constrained training for edge ai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2575–2587.

[26] J. Hao, P. Subedi, L. Ramaswamy, and I. K. Kim, "Reaching for the sky: Maximizing deep learning inference throughput on edge devices with ai multi-tenancy," *ACM Transactions on Internet Technology*, vol. 23, no. 1, pp. 1–33.

[27] X. Zhang, Y. Teng, N. Wang, B. Sun, and G. Hu, "Accelerating deep neural network tasks through edge-device adaptive inference," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6.

[28] S. Ge, Z. Luo, S. Zhao, X. Jin, and X.-Y. Zhang, "Compressing deep neural networks for efficient visual inference," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 667–672.

[29] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding."

[30] S. Trindade, L. F. Bittencourt, and N. L. d. Fonseca, "Resource management at the network edge for federated learning," *Digital Communications and Networks*, vol. 10, no. 3, pp. 765–782.

[31] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '17. ACM, pp. 506–519.

[32] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34.

[33] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81.

[34] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, "Quantum entanglement," *Reviews of Modern Physics*, vol. 81, no. 2, pp. 865–942.

[35] F. Arute, K. Arya, and J. M. Martinis, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510.

[36] C. Monroe, W. Campbell, L.-M. Duan, Z.-X. Gong, A. Gorshkov, P. Hess, R. Islam, K. Kim, N. Linke, G. Pagano, P. Richerme, C. Senko, and N. Yao, "Programmable quantum simulations of spin systems with trapped ions," *Reviews of Modern Physics*, vol. 93, no. 2, p. 025001.

[37] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, E. M. Chapple, C. Enderud, J. P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, C. J. S. Truncik, S. Uchaikin, J. Wang, B. Wilson, and G. Rose, "Quantum annealing with manufactured spins," *Nature*, vol. 473, no. 7346, pp. 194–198.

[38] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical Review Letters*, vol. 122, no. 4, p. 040504.

[39] N. Innan, M. A.-Z. Khan, A. Marchisio, M. Shafique, and M. Bennai, "Fedqnn: Federated learning using quantum neural networks," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–9.

[40] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, "Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices," *Physical Review X*, vol. 10, no. 2, p. 021067.

[41] L. Cheng, Y.-Q. Chen, S.-X. Zhang, and S. Zhang, "Quantum approximate optimization via learning-based adaptive optimization," *Communications Physics*, vol. 7, no. 1.

[42] S. Kim, S.-W. Ahn, I.-S. Suh, A. W. Dowling, E. Lee, and T. Luo, "Quantum annealing for combinatorial optimization: a benchmarking study," *npj Quantum Information*, vol. 11, no. 1.

[43] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1484–1509.

[44] P. Sibson, C. Erven, M. Godfrey, S. Miki, T. Yamashita, M. Fujiwara, M. Sasaki, H. Terai, M. G. Tanner, C. M. Natarajan, R. H. Hadfield, J. L. O'Brien, and M. G. Thompson, "Chip-based quantum key distribution," *Nature Communications*, vol. 8, no. 1.

[45] L. Chen, S. Jordan, Y.-K. Liu, D. Moody, R. Peralta, R. Perlner, and D. Smith-Tone, *Report on Post-Quantum Cryptography*.

[46] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, "Noisy intermediate-scale quantum algorithms," *Reviews of Modern Physics*, vol. 94, no. 1, p. 015004.