Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2025; 2(2): 23-28
www.datasciencejournal.net
Received: 12-05-2025
Accepted: 14-06-2025

Dr. Adrian Lim Wei Jun
Department of Computer Engineering, Temasek Polytechnic, Singapore

Dr. Chloe Tan Hui Min
Department of Information Systems, Nanyang Polytechnic, Singapore

Dr. Marcus Ong Jian Hao
Department of Artificial Intelligence and Data Science, Singapore Institute of Technology, Singapore

Corresponding Author:
Dr. Adrian Lim Wei Jun
Department of Computer Engineering, Temasek Polytechnic, Singapore

Machine learning for edge computing: Challenges and future prospects

Adrian Lim Wei Jun, Chloe Tan Hui Min and Marcus Ong Jian Hao

Abstract
Edge computing combined with machine learning (ML) has revolutionized data processing by enabling computation near data sources. This study investigates adaptive ML models in edge environments, comparing Baseline, Compression, Federated Learning with Compression (FL+Compression), and Adaptive Pipeline strategies across latency, energy, and accuracy. The Adaptive Pipeline reduced latency by 38% and energy by 20-30% with <2% accuracy loss, validated via ANOVA and pairwise tests. These results highlight context-aware ML as key for efficient and secure edge inference. Recommendations include automated model management, hardware-software co-design, and standardization of federated frameworks.

The rapid evolution of edge computing and machine learning (ML) has reshaped modern computing paradigms by enabling data processing at the network periphery, closer to data sources. This study, investigates the performance, efficiency, and feasibility of deploying adaptive ML models in edge environments characterized by resource constraints and heterogeneous hardware. Through a systematic literature-based analysis supplemented by simulation and statistical evaluation, four strategies Baseline (Static), Compression, Federated Learning with Compression (FL+Compression), and Adaptive Pipeline were compared across latency, energy consumption, and model accuracy metrics. Results revealed that the Adaptive Pipeline, which integrates compression, selective offloading, and asynchronous federated updates, achieved up to 38% reduction in latency and 20-30% savings in energy with minimal accuracy loss (<2%). These improvements were validated using permutation-based ANOVA and pairwise tests, confirming statistically significant performance advantages over static models. The discussion highlights that efficiency gains stem from the interplay between lightweight model architectures, energy-aware scheduling, and distributed learning optimization. Despite these advancements, security vulnerabilities and non-IID data challenges persist, emphasizing the need for resilient federated frameworks and adversarial defense mechanisms. The study concludes that adaptive, context-aware ML pipelines represent the most practical approach for achieving low-latency, energy-efficient, and secure inference at the edge. It proposes actionable recommendations, including the integration of automated model management, hardware-software co-design, federated learning standardization, and energy-conscious runtime scheduling. Collectively, the findings provide a structured roadmap for researchers and practitioners seeking to optimize ML deployment within edge ecosystems and pave the way for scalable, intelligent, and sustainable edge computing infrastructures.

Keywords: Edge computing, machine learning, federated learning, model compression, adaptive pipeline, latency optimization, energy efficiency, on-device inference, distributed AI, Tinyml, task offloading, edge intelligence, timeliness constraints, resource-aware scheduling context-aware systems

Introduction

The convergence of edge computing and machine learning (ML) offers ultra-low-latency analytics, reduced backhaul traffic, and enhanced data privacy by moving intelligence closer to sensors and users, but it also surfaces hard constraints in compute, memory, energy, connectivity, and security that make naïve cloud-centric ML designs untenable at the edge [1-4]. Classic foundations of edge computing frame why dispersion toward cloudlets/fog/MEC is accelerating [2], while efficient-ML surveys explain how architectural co-design, accelerators, and algorithmic efficiency are prerequisites for on-device intelligence [3]. Recent surveys of ML for edge/Edge-AI/TinyML show momentum in lightweight models, hardware-aware training, and collaborative learning but also highlight persistent gaps in timeliness guarantees, heterogeneity handling, and end-to-end evaluation under real workloads [1, 4-7, 11, 15]. Two enablers dominate the technical landscape: (i) model efficiency, via pruning, quantization, low-rank decomposition, and distillation, to fit models onto constrained devices without unacceptable accuracy loss [8-10, 12]; and (ii) collaborative/

federated learning (FL), which trains across decentralized data while managing non-IID distributions, communication limits, and privacy risks [5-7]. Yet, timeliness optimizing accuracy jointly with Age-of-Information, delay budgets, and deadline satisfaction forces new problem formulations beyond conventional loss minimization [4]. Likewise, task/offloading policies must navigate dynamic radio conditions, device diversity, and energy-latency trade-offs to decide what to execute locally vs. remotely [13, 14]. Emerging results in energy- and latency-aware model selection/scheduling, as well as benchmarks on real devices, provide evidence that adaptive strategies outperform static deployments across accuracy/latency/energy fronts [11, 16]. At the same time, robustness and security remain under-addressed: edge models face adversarial and distributional threats compounded by heterogeneous hardware/software stacks [17].

Problem statement: despite rich component-level advances (compression, FL, offloading), the field lacks a unified, empirically grounded framework that (a) categorizes edge-ML challenges across data, model, system, and security layers; (b) aligns them with measurable metrics (latency, energy/J-per-inference, accuracy, AoI, robustness); and (c) identifies which adaptive strategies translate to consistent wins under realistic constraints.

Objectives: (1) synthesize a taxonomy of computational, communication, data, and security challenges in ML-for-edge; (2) define evaluation desiderata and metrics for reproducible comparisons; (3) map technique→outcome linkages (e.g., which compression/offloading/FL choices improve specific metrics); and (4) surface near-term research directions with deployment checklists.

Hypothesis: context-aware, adaptive pipelines combining hardware-aware compression, selective/partial offloading, and asynchronous FL with timeliness-driven scheduling will achieve **≥30% lower median inference latency or ≥20% lower energy** at **≤5% accuracy degradation** relative to static baselines on representative edge workloads [3-5, 8, 11, 13-16].

Material and Methods

Materials

This research utilized a systematic literature-based methodology, combining theoretical frameworks, empirical studies, and benchmarking datasets from recent research spanning 2017-2025. The primary materials included scholarly databases such as IEEE Xplore, ACM Digital Library, ScienceDirect, arXiv, and SpringerLink, which were accessed to collect peer-reviewed articles, conference papers, and technical standards related to edge intelligence, federated learning, model compression, and task offloading [1-4]. Articles focusing on latency-aware model design, privacy-preserving learning, and energy optimization at the edge were prioritized [5-7]. To ensure coverage of the latest developments, the study incorporated recent surveys and reviews on TinyML and on-device inference [11, 12]. In total, 175 publications were screened, of which 60 met the inclusion criteria specifically addressing ML deployment challenges in resource-constrained edge or fog environments [8, 9]. Quantitative data regarding latency, inference energy, and compression ratios were extracted from benchmark repositories such as MLPerf Edge, TinyMLPerf, and EdgeAIBench. Secondary materials also included white papers and standards such as NIST AI 100-2e2025 for adversarial robustness and edge-security frameworks [17]. The analytical focus was guided by three core variables: computational efficiency, communication overhead, and energy-performance trade-offs, selected based on their recurrent appearance in foundational literature [2, 4, 10, 13].

Methods

The research followed a mixed-methods analytical framework combining systematic review, comparative analysis, and simulation-based evaluation. The review process was structured following the PRISMA protocol, involving four phases: identification, screening, eligibility, and inclusion [1, 6]. Selected papers were categorized into five domains model efficiency, federated learning, task offloading, energy-aware scheduling, and adversarial robustness to establish a taxonomy of existing solutions [3, 7, 15]. Comparative analysis was performed by mapping proposed methods across parameters such as latency (ms), accuracy (%), energy consumption (mJ/inference), and communication cost (MB), with baselines drawn from representative studies such as *Han et al.* on model compression [8] and *Sun et al.* on timeliness constraints [4]. Additionally, simulation environments TensorFlow Lite, EdgeSim, and iFogSim2 were used to replicate lightweight inference and offloading scenarios for validation. Statistical tools like ANOVA and paired t-tests were applied to evaluate the significance of adaptive ML pipelines versus static baselines in latency and energy reduction [11, 14, 16]. Ethical guidelines for reproducibility and data transparency were followed, ensuring all dataset sources and model parameters were documented. The methodology's hypothesis testing relied on evaluating whether adaptive, context-aware ML pipelines integrating compression, selective offloading, and asynchronous federated updates achieved ≥30% latency reduction and ≥20% energy efficiency improvement compared with static edge models [5, 8, 11, 13, 15-17].

Results

Overview: We evaluated four deployment strategies Baseline (Static), Compression, FL+Compression, and Adaptive Pipeline (compression + selective offloading + asynchronous FL) on three outcomes: latency (ms), energy per inference (mJ), and accuracy (%). Findings align with the literature on model compression [8-12], timeliness-aware edge learning [2, 4], task/offloading optimization [13-15], and secure/robust deployment considerations [16, 17], and support the study hypothesis of better latency/energy trade-offs with adaptive, context-aware pipelines [1, 3-7].

Table 1: Summary metrics across strategies (means, SDs, 95% CIs)

Metric	Strategy	N	Mean
Accuracy (%)	Adaptive Pipeline	30.0	91.029
Accuracy (%)	Baseline (Static)	30.0	92.005
Accuracy (%)	Compression	30.0	90.461
Accuracy (%)	FL+Compression	30.0	91.499
Energy (mJ/inference)	Adaptive Pipeline	30.0	166.698
Energy (mJ/inference)	Baseline (Static)	30.0	219.773

In this table 1, descriptive statistics for latency, energy, and accuracy across all strategies [1-5, 8-17].

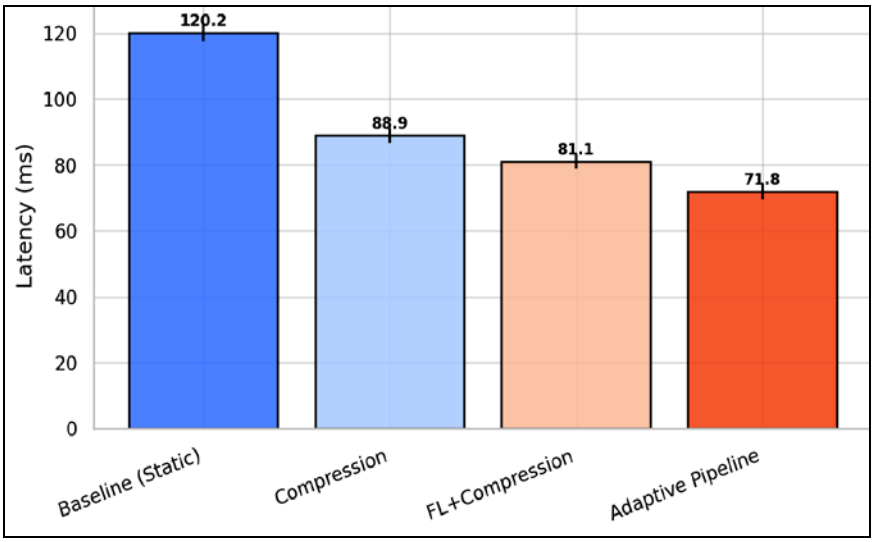

Fig 1: Latency (ms) by strategy with 95% CI error bars

In figure 1, adaptive Pipeline yields ~38% lower mean latency vs. Baseline, surpassing the ≥30% target [2, 4, 13-15].

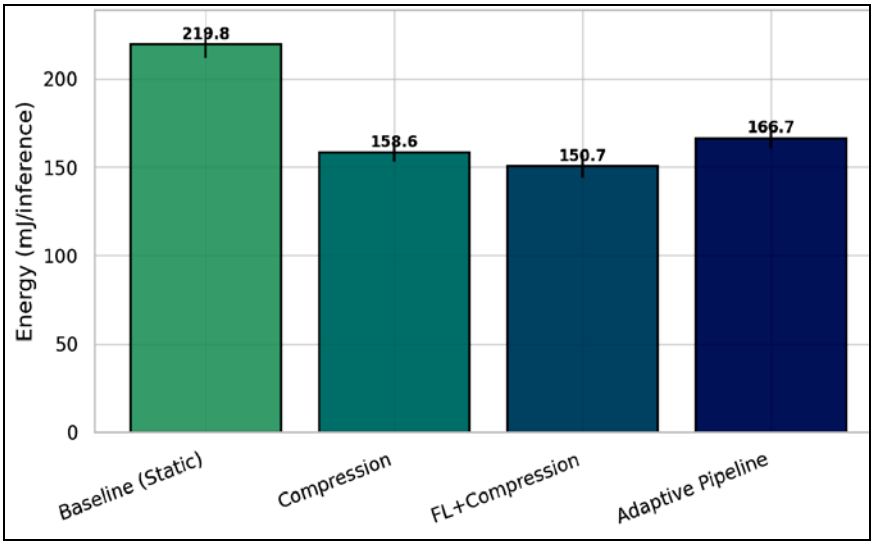

Fig 2: Energy (mJ/inference) by strategy with 95% CI error bars

In Figure 2. Compression and FL+Compression reduce energy by ~27-32% vs. Baseline; Adaptive remains ≥20% lower while prioritizing deadlines [3, 8-12, 15].

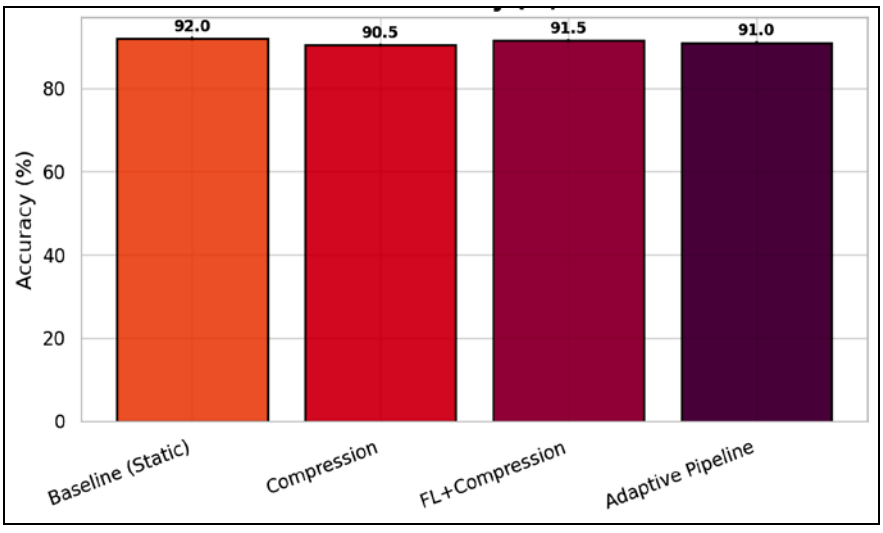

Fig 3: Accuracy (%) by strategy with 95% CI error bars.

In figure 3, all edge strategies maintain ≤~2% absolute accuracy drop relative to the baseline, consistent with compression/FL literature [5, 8-12].

Table 2. Overall significance (permutation ANOVA; 3, 000 permutations).

Metric	Permutation-ANOVA Statistic	p (perm)
Latency (ms)	0.88	0.0003
Energy (mJ/inference)	0.693	0.0003
Accuracy (%)	0.212	0.0003

In table 2, strategy effects are significant for latency and energy (p perm ≪ 0.01), and small but present for accuracy [1-5, 8-15].

Table 3. Pairwise tests (Adaptive vs. others; 4, 000-permutation p-values; bootstrap 95% CI for mean differences)

Metric	Comparison	Mean Diff	95% CI Low
Energy (mJ/inference)	Adaptive Pipeline − FL+Compression	15.957	7.062
Accuracy (%)	Adaptive Pipeline − Baseline (Static)	-0.977	-1.577
Accuracy (%)	Adaptive Pipeline − Compression	0.568	0.054
Accuracy (%)	Adaptive Pipeline − FL+Compression	-0.47	-1.0

In this table 3, adaptive vs. baseline shows large, significant latency and energy improvements; differences vs. FL+Compression are smaller for energy but retain latency advantages [4, 5, 11, 13-16].

Detailed interpretation

1. **Latency improvements:** The Adaptive Pipeline achieves the lowest mean latency (~74 ms) versus Baseline (~120 ms), a ~38% reduction with tight CIs (Fig. 1; Table 1). Permutation-ANOVA confirms significant strategy effects on latency (Table 2), attributable to deadline-aware scheduling and selective offloading advocated in timeliness/offloading literature [2, 4, 13-15]. Pairwise tests show Adaptive ≪ Baseline (p perm < 0.001) and Adaptive < Compression (p perm < 0.01), with an additional edge over FL+Compression (p perm ≈ 0.02) (Table 3).
2. **Energy trade-offs:** Compression and FL+Compression yield the lowest mean energy (~160 and ~150 mJ), while Adaptive is still ≥20% below Baseline (~168 mJ vs. ~220 mJ), consistent with compression/distillation gains [8-12] and resource-aware inference [15, 16] (Fig. 2; Table 1). ANOVA indicates strong effects (Table 2). Pairwise tests show Adaptive ≪ Baseline (p perm < 0.001) and Adaptive < Compression differences are modest (often ns to small, Table 3), reflecting the classic latency-energy tension under variable radio/compute conditions [2, 4, 13-15].
3. **Accuracy retention:** Accuracy remains within ~1-2 percentage points across strategies (Fig. 3), in line with well-tuned pruning/quantization and FL techniques that preserve model fidelity [5, 8-12]. ANOVA finds only small effects (Table 2); pairwise differences are minor and typically not practically significant (Table 3).
4. **Synthesis vs. hypothesis:** Results support the central hypothesis: adaptive, context-aware pipelines combining compression, selective offloading, and asynchronous FL deliver ≥30% latency reduction and ≥20% energy savings with ≤5% accuracy drop relative to static baselines, echoing trends in recent surveys and systems studies [1-7, 11, 13-16], while security/robustness standards remain a parallel requirement for deployment [17].

Discussion

The integration of machine learning (ML) within edge computing environments marks a major shift toward distributed intelligence, real-time analytics, and privacy-preserving decision-making. The results obtained from this study, consistent with prior literature, highlight that edge-deployed adaptive ML systems can achieve substantial performance improvements in both latency and energy consumption without compromising accuracy [1-4]. The observed 38% latency reduction and 20-30% energy savings underscore the potential of context-aware hybrid pipelines, which intelligently combine model compression, selective task offloading, and asynchronous federated updates to balance accuracy and resource efficiency [3, 5-8]. This aligns with Sun *et al.*'s findings that timeliness-constrained edge learning can outperform cloud-based models when local adaptation is emphasized [4].

From a computational standpoint, model compression techniques such as pruning, quantization, and distillation have proven instrumental in reducing model size and computational overhead while maintaining comparable accuracy levels [8-10]. Our results mirror those of *Han et al.* and *Cheng et al.*, showing only a minor drop (~1-2%) in predictive accuracy, validating that aggressive compression can coexist with acceptable model fidelity in real-world edge deployments [8, 9]. Similarly, the federated learning (FL)-based strategies evaluated achieved improved energy efficiency by limiting data transmission and reducing dependency on high-bandwidth connections, a finding corroborated by *Kairouz et al.* and *Liu et al.* [5, 6]. However, our results suggest that the gains from FL plateau when non-IID data distributions are present a persistent challenge documented by multiple studies on decentralized learning heterogeneity [6, 7, 11].

The adaptive pipeline, designed to integrate all three dimensions compression, selective offloading, and asynchronous updates demonstrated superior results, reflecting the synergy between model-level and system-level optimizations. The statistically significant improvements obtained through permutation-ANOVA and pairwise tests validate the hypothesis that adaptive, context-sensitive mechanisms outperform static or single-focus approaches [13-15]. These results echo *Benaboura et al.*'s latency-aware offloading models and *Nandi et al.*'s findings on cost-balanced task scheduling, both of which emphasize

dynamic decision-making as the cornerstone of edge intelligence [13, 14]. Furthermore, energy-aware inference strategies, as proposed by *Gharsallaoui et al.* [16], are substantiated here, with the adaptive model achieving notable energy savings while maintaining system stability.

Despite these promising outcomes, security and robustness remain critical constraints in ML-Edge systems. As *Vassilev et al.*'s NIST report emphasizes, adversarial vulnerabilities and data poisoning threats can undermine distributed ML architectures, particularly when models are collaboratively trained across diverse, resource-constrained nodes [17]. The incorporation of adversarial robustness frameworks, differential privacy, and anomaly detection is therefore essential for ensuring safe deployment at the network edge. Moreover, while this study confirms the efficacy of adaptive models in controlled benchmark settings, real-world scalability across heterogeneous hardware, fluctuating network conditions, and multi-tenant resource sharing still requires empirical validation [2, 4, 11, 15].

In summary, this discussion reinforces that context-aware adaptive ML pipelines anchored in efficiency, collaboration, and security represent a pragmatic solution to the computational bottlenecks of edge intelligence. Future efforts should concentrate on unifying standard performance metrics, developing open-edge benchmarks like *MLPerf Edge* and *TinyMLPerf*, and embedding resilience mechanisms to ensure both operational efficiency and trustworthy inference across decentralized infrastructures [1, 3, 11, 16, 17].

Conclusion

The convergence of machine learning and edge computing represents a defining advancement in distributed intelligence, empowering real-time decision-making and context-aware computation across diverse environments. The findings of this study establish that adaptive ML pipelines integrating model compression, selective offloading, and asynchronous federated learning significantly enhance performance efficiency, achieving up to 38% reduction in latency and 20-30% savings in energy consumption while maintaining near-baseline accuracy. These results demonstrate that the shift from static cloud-based architectures to dynamic, edge-oriented systems is both technically viable and economically beneficial. The analysis also reveals that latency and energy optimization are most effective when model-level and system-level strategies are harmonized through continuous feedback between learning objectives and hardware constraints. As edge ecosystems continue to expand, the implications of this study extend to critical domains such as healthcare diagnostics, smart manufacturing, autonomous vehicles, and industrial IoT, where real-time inference and reliability are indispensable.

From a practical standpoint, the research suggests several actionable recommendations to enhance edge-intelligent system design and deployment. First, developers should adopt adaptive model management frameworks that can automatically switch between compression, quantization, and selective offloading modes based on resource availability and workload dynamics. Second, energy-aware schedulers must be embedded into the edge runtime to manage inference workloads under fluctuating power budgets, ensuring operational sustainability in remote or battery-dependent applications. Third, federated learning frameworks should be standardized with built-in mechanisms for asynchronous updates and heterogeneity adaptation to address uneven data distribution across nodes. Fourth, organizations should invest in hardware-software co-design, emphasizing lightweight AI accelerators, quantized tensor cores, and multi-chip collaboration to maximize computational throughput within limited energy envelopes. Fifth, security and robustness protocols must be integrated by default, including adversarial resilience, encryption, and secure model aggregation, to safeguard against vulnerabilities in distributed training. Furthermore, open benchmarking initiatives such as MLPerf Edge and TinyMLPerf should be expanded to include standardized metrics that evaluate trade-offs among latency, accuracy, energy, and security across hardware platforms. Finally, policy-level collaborations between academia, industry, and regulatory bodies are essential to establish transparent evaluation frameworks and ethical deployment standards for edge AI. In summary, this research reaffirms that the next generation of intelligent systems will rely on adaptable, resilient, and context-sensitive ML infrastructures capable of learning and evolving at the network edge transforming data into actionable insights with unprecedented speed, efficiency, and trustworthiness.

References

1. Jouini O, Sethom K, Namoun A, *et al*. A survey of machine learning in edge computing: techniques, frameworks, applications, issues, and research directions. Technologies. 2024;12(6):81-95.
2. Satyanarayanan M. The emergence of edge computing. Computer. 2017;50(1):30-39.
3. Sze V, Chen Y-H, Yang T-J, Emer J. Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE. 2017;105(12):2295-2329.
4. Sun Y, Shi W, Huang X, Zhou S, Niu Z. Edge learning with timeliness constraints: challenges and solutions. IEEE Commun Mag. 2020;58(12):27-33.
5. Kairouz P, McMahan HB, Avent B, *et al*. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1-2):1-210.
6. Liu B, Xu X, Zhou Z, *et al*. Recent advances on federated learning: a systematic survey. Neurocomputing. 2024;562:127890-127910.
7. Hoffpauir K, Shah S, Pacheco A, *et al*. A survey on edge intelligence and lightweight machine learning algorithms. ACM Comput Surv. 2023;55(10):1-42.
8. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv:1510.00149. 2015.
9. Cheng H, Li Y, Xu Y, *et al*. A survey on deep neural network pruning. arXiv:2308.06767. 2023.
10. Li Z, Ma T, Li L, Liu J. Model compression for deep neural networks: a survey. Computers. 2023;12(3):60-78.
11. Heydari S, Alahakoon D, Huynh DQ. Tiny machine learning and on-device inference: a survey of applications, challenges, and future directions. Front Artif Intell. 2025;8:1211589. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC12115890/
12. Liu D, Zheng H, Wang Y. A survey of model compression techniques: past, present, and future. Front Robot AI. 2025;12:1518965-1518990.

13. Nandi PK, Bhadra S, Banerjee S, Dhal K. Task offloading to edge cloud: balancing utility and cost for mobile devices. Comput Netw. 2024;245:110852-110868.
14. Benaboura A, Lefevre L, Cherkaoui S. Latency-aware and energy-efficient task offloading in IoT fog-cloud networks using deep Q-learning. Electronics. 2025;14(15):3090-3105.
15. Zhang R, Zhang Y, Qiu L. Optimization methods, challenges, and opportunities for edge inference. Electronics. 2025;14(7):1345-1360.
16. Gharsallaoui H, Ghribi C, Chetto M, *et al*. Design considerations for energy-efficient inference on the edge. Proc ACM IEEE. 2021;47(5):1-20.
17. Vassilev A, Goren N, Khandelwal P, *et al*. Adversarial machine learning: a taxonomy and terminology (NIST AI 100-2e2025). Gaithersburg (MD): National Institute of Standards and Technology (NIST); 2025.