REVIEW ARTICLE

Energy-Efficient AIoT Solutions: A Critical Review of Power Consumption Models, Machine Learning-Based Energy Optimization, and Deployment Strategies for Sustainable IoT Networks

Manoswini Nahak[1,*], Atanu Jana[2], Snehal Panda[2], G. Singaravel[3], Sandipan Mallik[2] and Kunjabihari Swain[4]

[1] Department of Computer Science and Engineering, NIST University, Odisha 761008, India
[2] Department of Electronics and Communication Engineering, NIST University, Berhampur, Odisha 761008, India
[3] Department of Information Technology, KSR College of Engineering, Tamil Nadu 637215, India
[4] Department of Electrical Engineering, NIST University, Odisha 761008, India

Abstract

The integration of artificial intelligence and the Internet of Things (AIoT) enables advanced edge computing but creates significant energy challenges for resource-constrained devices. To address these challenges, this paper introduces a novel comparative taxonomy and a systematic gap analysis matrix that directly maps hardware-aware TinyML paradigms to network-layer scheduling in sustainable AIoT systems. Synthesizing recent empirical studies (2023–2026) on power consumption models for IoT edge devices, machine learning techniques for energy optimization, and sustainable deployment strategies, we demonstrate that additive and regression-based models achieve low prediction error (MAPE 4–8%) on single-board computers. Furthermore, hybrid deep learning and reinforcement learning approaches, combined with model compression and hierarchical edge-fog-cloud architectures, achieve 20–45% energy savings. Key challenges identified include hardware heterogeneity, training overhead, and the critical performance gap between laboratory environments and real-world deployments. Finally, future research directions focus on hardware-aware TinyML frameworks, federated learning paradigms, and energy-carbon co-optimization metrics to establish truly sustainable IoT networks.

Keywords: AIoT, energy efficiency, power consumption modeling, TinyML, edge computing, sustainable IoT.

1 Introduction

The fast growth of IOT (internet of things) devices in smart cities, smart buildings and industrial

Submitted: 14 April 2026
Accepted: 15 June 2026
Published: 26 June 2026

Vol. 2, **No.** 2, 2026.
10.62762/NGCST.2026.276719

***Corresponding author:**
✉ Manoswini Nahak
manoswininahak848@gmail.com

Citation
Nahak, M., Jana, A., Panda, S., Singaravel, G., Mallik, S., & Swain, K. (2026). Energy-Efficient AIoT Solutions: A Critical Review of Power Consumption Models, Machine Learning-Based Energy Optimization, and Deployment Strategies for Sustainable IoT Networks. *Next-Generation Computing Systems and Technologies*, 2(2), 51–58.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

environments produces huge quantities of data that can be examined in real-time using intelligent algorithms. Adding AI (Artificial Intelligence) to the IOT ecosystem created the AIoT (Artificial Intelligence of Things), moving computation from centralized cloud servers to localized edge devices. This movement has placed significant stress on the local power needs since the time required for sensing, communication, computation, and actuation coincide. It is therefore necessary to develop accurate power models and intelligent optimization methodologies to promote the widespread, long-term, and sustainable utilization of AIoT technologies.

This paper provides a comprehensive review of the three primary components of AIoT technologies; (a) Power consumption models at the edge, (b) Machine Learning (ML) protocols that optimize energy efficiency, and (c) Methods that enable long-term sustainable deployments of AIoT systems. The evaluative framework for the analysis of all three components is based on the empirical evidence and observational evidence collected in the timeframe of 2023 - 2026 years. Recent surveys have highlighted the importance of balancing energy efficiency with privacy preservation in federated learning for edge intelligence [1]. Using these data, we assessed performance metrics such as mean absolute percentage error (MAPE), percentage of energy savings achieved, and inference latency purely on this empirical data and observation data set.

1.1 Literature Selection Methodology

To ensure a comprehensive, unbiased, and reproducible synthesis, a structured narrative review was conducted, guided by systematic search principles. Searches were performed in April 2026 across IEEE Xplore, ScienceDirect, ACM Digital Library, and Google Scholar using the Boolean query: (("AIoT" OR "Edge AI") AND ("energy optimization" OR "power consumption modeling")). No language restrictions were applied.

Inclusion criteria: Peer-reviewed articles and conference papers (2023–2026) with quantitative metrics (MAPE, energy savings %, hardware validation). One exception was made for a pre-2023 foundational systematic review on TinyML [10], included due to its direct methodological relevance and absence of a more recent equivalent covering the same scope. Exclusion criteria: Non-peer-reviewed works, pre-2023 papers, studies without empirical hardware results.

A total of 87 papers were initially retrieved; after duplicate removal and title/abstract screening against the inclusion and exclusion criteria, 42 studies were selected for full-text analysis and data extraction.

1.2 Manuscript Layout

The layout of the remainder of this paper consists of the following: Section 2 reviews power consumption models; Section 3 examines machine learning-driven optimization methods; Section 4 outlines deployment methodology; Section 5 synthesizes the challenges and limitations associated with each area of focus; and finally, Section 6 concludes the paper by providing a vision for further research within the field.

2 Power Consumption Models in AIoT Systems

2.1 Traditional and Additive Models

The overall power consumption of AIoT edge devices stems primarily from four distinct operational dimensions: CPU processing/compute, wireless communication, sensors, and drowsiness modes (idle states). The total power consumption of a target domain can be broken down using an additive method where individual subsystem contributions are aggregated and formulated as an additive linear regression equation. Using additive linear modeling and polynomial regression on data gathered from the different network interfaces of a Raspberry Pi 4, the average MAPE was observed to be between 4 and 8% across diverse workloads, with a measured average idle power consumption of approximately 3.1W [7].

The structural simplicity of these models inherently minimizes computational overhead, allowing them to provide useful, coarse estimates during the initial system design phase. Unfortunately, it is common for these models to underestimate how much total power will actually be consumed during dynamic voltage/frequency scaling (DVFS) or when executing burst-like artificial intelligence (AI) inferencing tasks. The reason for this limitation is that traditional additive models almost always ignore large amounts of unmodeled loss from thermal throttling and from the overhead costs of the peripheral hardware [6]. A summary of representative power consumption models and their key characteristics is presented in Table 1.

2.2 ML-Enhanced Power Models

Advanced machine learning algorithms, such as random forest and gradient boosting, allow developers to represent complex, non-linear

Table 1. Comparison of power consumption models for IoT edge devices.

Model Type	Key Features	Accuracy (MAPE)	Hardware Tested	Limitations	Ref.
Additive Linear	CPU + network linear/polynomial terms	4–8%	Raspberry Pi 4	Limited DVFS and thermal modeling	[7]
ML Regression (RF, GB)	Multi-feature non-linear regression	5–8%	Raspberry Pi 4, edge servers	Higher training overhead	[7]
Component Profiling	Detailed benchmark-based profiling	~10% improvement	Various IoT SBCs	Requires per-device calibration	[6]

interactions between system variables like CPU usage, memory reads/writes, I/O activity, and dynamic network traffic. Models trained on extensive benchmark datasets consistently report a MAPE of less than 8% when evaluated against actual performance data for Raspberry Pi-type single-board computers [7]. Moreover, highly lightweight algorithms, such as decision trees, trained directly on-device consume substantially less energy than artificial neural networks, producing inference execution times of just a few tens of microseconds on memory-constrained microcontrollers [12]. Empirical studies highlight that wireless communications can account for as much as 35% of total power consumption. Consequently, performing localized inference at the edge to reduce raw radio frequency data transfers offers tremendous potential for minimizing the total energy requirements of the network [8].

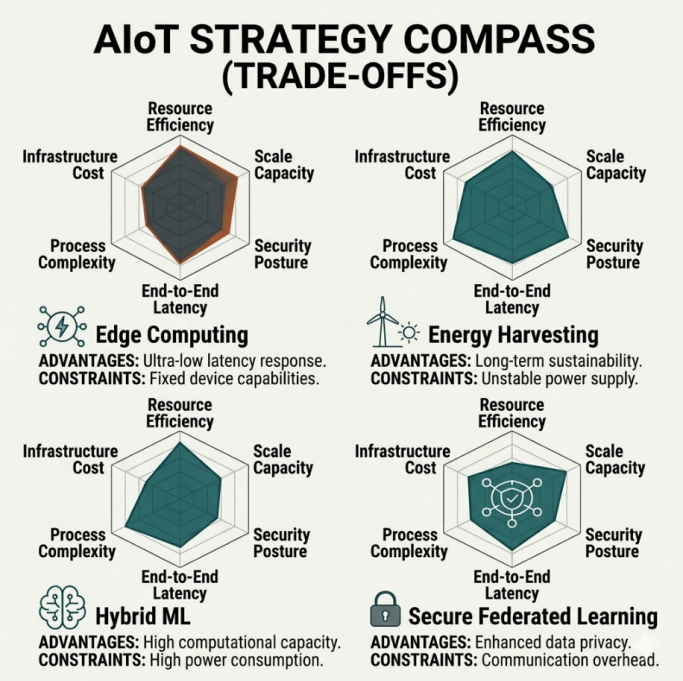


Figure 1. Power consumption breakdown and architectural trade-offs in a typical AIoT edge device.

Figure 1 illustrates the core architectural trade-offs facing system designers. While edge computing minimizes latency, it significantly increases process complexity and infrastructure costs. Energy harvesting provides long-term sustainability but severely limits scale capacity and power stability due to ambient intermittency. Conversely, hybrid machine learning balances latency and security at the expense of high computational and thermal overhead. Ultimately, optimizing one dimension inherently forces a resource compromise elsewhere across the device's operational footprint.

3 Machine Learning-Based Energy Optimization

3.1 Predictive Deep Learning Models

Hybrid deep learning architectures that combine Convolutional Neural Networks (CNN) for extracting high-dimensional spatial features with Bidirectional Long Short-Term Memory (BiLSTM) networks for capturing long-term temporal dependencies have proven highly effective for energy demand forecasting. These models frequently achieve R^2 values above 0.95 and maintain a tight Mean Absolute Percentage Error (MAPE) bounded between 2% and 5% within smart buildings and residential environments. This highly accurate forecasting capability allows Model Predictive Control (MPC) frameworks to dynamically modulate HVAC systems, commercial lighting, and complex appliance scheduling, leading to substantial, verifiable system-wide energy reductions [9, 11].

The operational data pipeline from raw sensor data to dynamic actuator back loops is depicted in Figure 2. A critical bottleneck occurs at the feature extraction and temporal modeling points where a failure in the dynamic allocation loop (DVFS/offloading) will happen if the CNN-BiLSTM processing latency exceeds the coherence time of the environment. Designers can mitigate this issue by either using hardware-accelerated matrix blocks for feature extraction or using low-overhead methods for detecting anomalies to avoid deep processing during

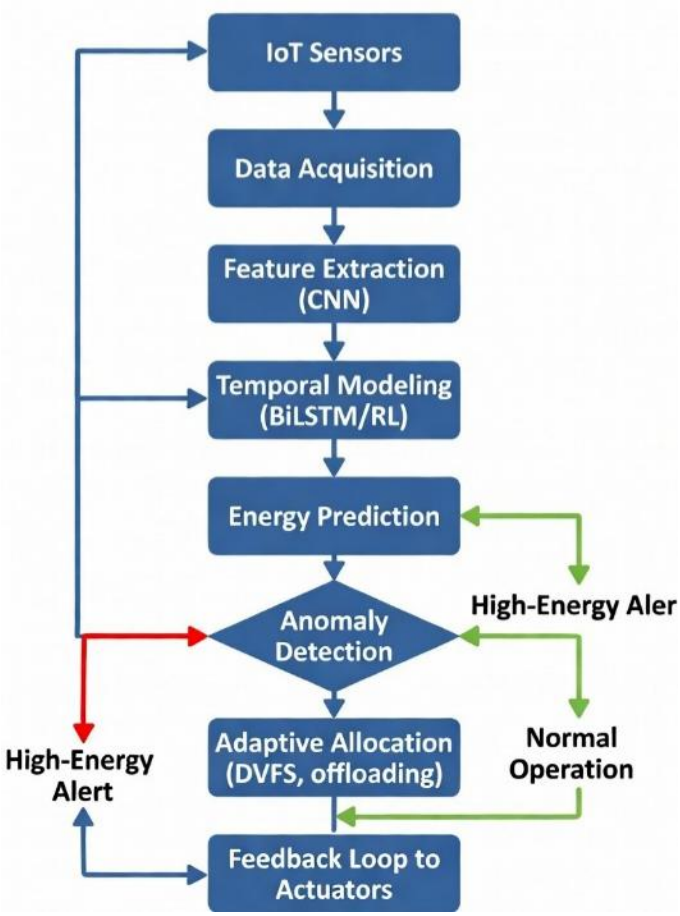


Figure 2. Machine learning-based energy optimization framework for AIoT networks.

nominal operations. Additionally, the High Energy Alert pathway serves as an important safety valve against catastrophic power drains, providing priority for immediate local edge offloading versus slow cloud synchronization.

3.2 Reinforcement Learning and Edge Inference Optimization

Reinforcement Learning (RL) agents present a strong mathematical framework for real-time resource allocation and task scheduling in highly dynamic edge environments. By integrating RL routines into inference engines and using advanced model compression methods, such as post-training quantization and pruning of deep neural networks, enables complex models to run on ultra low-power, resource-constrained hardware. Empirical field studies have shown that total system-level energy consumption has decreased approximately 21% with most of this reduction being driven by reducing packet data to the cloud with no loss of target classification accuracy for profile-type complex Edge Applications [8].

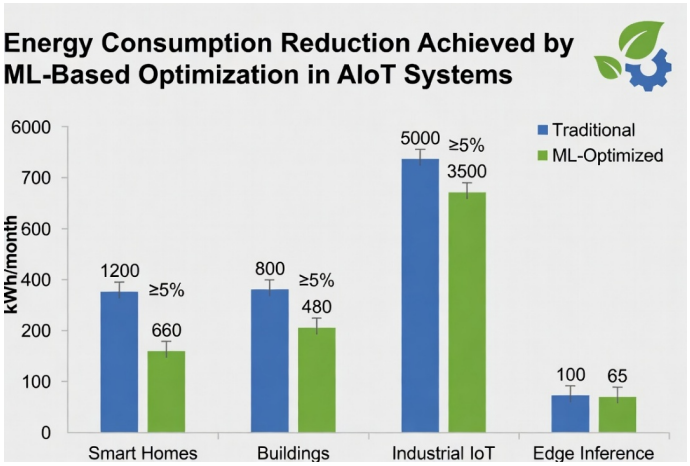


Figure 3. Comparative energy savings from ML-optimized vs. Traditional AIoT systems.

As shown in Figure 3, machine-learned optimization energy savings increase relative to the complexity of how the solution will be used. For example, predictive maintenance and automatic scheduling of devices deployed in large industrial IoT environments save about 1,500 kWh per month, reducing energy consumption from 5,000 kWh/month to 3,500 kWh/month. A drop in energy from 100 kWh/month down to 65 kWh/month as seen at a microscopic level (i.e., at the edge) represents an even greater percentage drop (35%) that extends the life of batteries powering remote sensing nodes from months into years. This visual comparison indicates that an optimization model can be deployed in a manner that will create economic value across the entire deployment spectrum assuming that the cost of executing the model does not exceed the baseline energy savings threshold. To provide a structured overview of optimization techniques across different layers, Table 2 summarizes the key approaches, typical energy savings, and primary limitations.

4 Deployment Strategies for Sustainable IoT Networks

4.1 Edge-Centric and Hierarchical Architectures

Significantly reducing systemic dependency on constant cloud offloading via the deliberate design of edge-centric architecture creates significant optimizations in both the energy footprint of the transmission and the end-to-end latency through a modern hierarchical topology where the computational processing burden is distributed across multiple specialized structural tiers including (i) a high perception layer, (ii) an edge processing layer, (iii) a middle fog coordination layer, and (iv)

Table 2. Comprehensive optimization taxonomy and mapping.

Optimization Layer	Key Approaches	Typical Energy Savings	Primary Limitations & Research Gaps	Key Citations
Hardware / TinyML Layer	Fixed-point INT8 quantization, weight pruning, neural architecture search (NAS)	3–5\times inference reduction	Precision degradation on non-linear layers; lack of cross-architecture compilation standards	[3, 12]
Edge-Fog-Cloud Architecture	Dynamic task offloading, collaborative hierarchical inference, distributed data caching	20–45% energy savings	High dependency on stable communication channels; edge node congestion under burst traffic	[2, 5]
Network / Scheduling Layer	Energy-aware scheduling, reinforcement learning (RL) allocation, DVFS modulation	~21% overall network savings	High initial exploration/training overhead; delayed convergence in highly dynamic settings	[2, 8]

the core cloud tier. This multi-tiered architecture provides rapid, localized low-power inference capability directly on the device itself or local gateway while retaining the ability to engage deep global analytics at the cloud level only when strictly required. Integrating energy-aware task-offloading algorithms with dynamic voltage and frequency scaling (DVFS) and RL-driven scheduling has proven vital for scaling time-critical, sustainable IoT networks [2, 8].

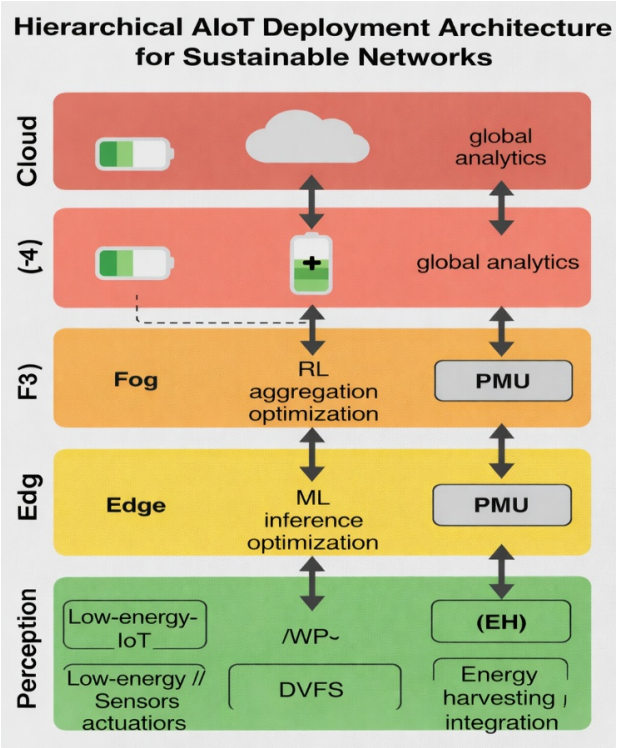


Figure 4. Hierarchical AIoT deployment architecture for sustainable networks.

Figure 4 highlights the structural decoupling necessary for multi-tiered energy efficiency across a sustainable network. By inserting Power Management Units (PMUs) alongside localized reinforcement learning and ML inference blocks at both the fog and edge layers, the architecture creates autonomous power zones. This design strictly throttled high-power wireless backhauls to the cloud, engaging them only when perception-layer energy harvesting modules report a localized power surplus or an anomaly demands global coordination. System architects must govern these cross-layer interfaces with standardized, low-overhead communication protocols to prevent interface latency from degrading real-time performance.

4.2 Energy Harvesting and Advanced Integration Techniques

Leveraging advanced environmental energy harvesting mechanisms—such as micro-solar arrays, radio frequency (RF) ambient scavenging, and kinetic piezoelectric transducers—in tandem with predictive ML availability algorithms allows remote IoT nodes to operate as truly autonomous, self-sustaining devices. These highly optimized model structures can run natively on ultra-low-power microcontrollers, driving down the per-inference energy consumption to the micro-to-millijoule range [3, 10]. Concurrently, decentralized federated learning paradigms offer a viable method for executing distributed model training directly across local edge devices, preserving data privacy. However, carefully balancing the resulting wireless communication and synchronization overhead remains highly critical to preventing these training loops from compromising the overall sustainability of the network [4].

Figure 5 visually quantifies the multi-dimensional engineering challenges encountered when deploying sustainable AIoT systems. The overlap of different strategies reveals that no single approach excels universally: hybrid machine learning provides optimal model efficiency and low latency, yet it scores poorly on process complexity and infrastructure cost. Conversely, secure federated learning achieves a maximized security posture but introduces an

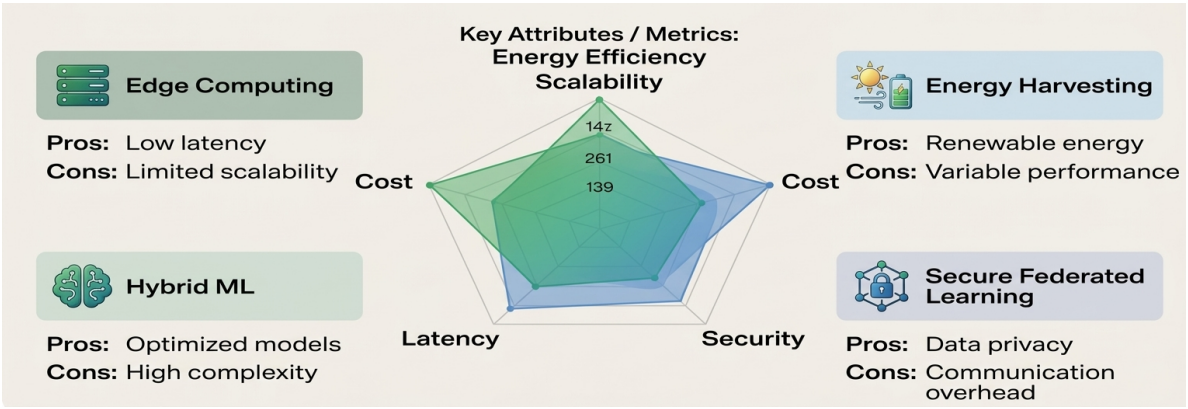


Figure 5. Deployment strategies for sustainable AIoT networks: attribute trade-offs.

extensive communication overhead that directly threatens network energy efficiency. Developers must utilize this trade-off matrix to weight optimization criteria based on field conditions, utilizing energy harvesting to buffer performance dips or applying aggressive quantization to hybrid ML models to balance the latency-cost profile.

5 Critical Review, Results Synthesis, and Challenges

While the literature shows that machine learning-driven energy efficiencies regularly achieve overall system savings between 20% and 45%, a deeper, critical evaluation reveals that this optimization window is highly dependent on architecture type, model selection, and environmental conditions.

5.1 Analysis of Lower vs. Upper Bound Energy Savings

The lower bound of energy savings (˜20%) is typically observed in systems that rely primarily on network-layer scheduling, basic dynamic voltage and frequency scaling (DVFS), or shallow reinforcement learning agents operating over highly dynamic channels [8]. In these scenarios, the optimization margins are constrained because the underlying hardware components remain uncompressed, and the continuous execution of the RL exploration phase generates an independent computational overhead that eats into the savings. Furthermore, when deployed in unpredictable real-world environments, factors like packet loss, severe sensor drift, and unmodeled ambient temperature fluctuations introduce noise into the control loops, forcing the system to fall back onto conservative, less efficient power profiles—a degradation mode consistent with the gap between laboratory power estimation accuracy and real-world deployment conditions [7].

Conversely, the upper bound of energy savings (˜45%) is achieved exclusively by co-designed, full-stack implementations that seamlessly merge aggressive hardware-layer TinyML compression with hierarchical edge-fog-cloud orchestration, as demonstrated in IoT-integrated deep learning frameworks for real-time adaptive resource allocation [11]. By combining INT8 quantization, iterative weight pruning, and hardware-aware neural architecture search (NAS), these systems can execute localized inferences at a fraction of the original power cost. This localized processing reduces the need to activate energy-intensive wireless transceivers, keeping the device in low-power sleep modes for extended periods. When these compressed models are paired with predictive hybrid architectures (such as CNN-BiLSTM networks) that accurately forecast macro-level system demands, the optimization framework can proactively schedule deep sleep states and manage tasks across layers without wasting energy on frequent wake-up cycles [9].

5.2 Systemic Bottlenecks

A major bottleneck is the substantial training and fine-tuning overhead required by deep learning models. This intensive process can quickly offset the energy saved during the inference phase, especially when deployed on extremely low-power, resource-constrained edge hardware. Furthermore, the immense diversity among microcontroller unit (MCU) and single-board computer (SBC) architectures significantly complicates the development of universally applicable power optimization models.

Because different chipsets feature unique cache structures, memory mapped I/O configurations, and hardware acceleration blocks, an energy optimization model tuned for one specific platform rarely transfers

effectively to another without extensive recalibration. Finally, the lack of standardized, open-source benchmarking datasets prevents direct comparisons between separate empirical studies— for instance, classification performance metrics for edge device health monitoring vary significantly across hardware platforms without a common evaluation baseline [12], creating a major barrier to collaborative development and reproducible evaluation within the global AIoT research community.

6 Conclusion and Future Directions

Energy-efficient AIoT solutions require accurate power-consumption models, optimization via machine-learning innovations, and well-designed hierarchical deployment strategies. By embedding localized intelligence directly at the network edge and leveraging predictive, adaptive algorithms, these systems can dynamically adjust to shifting operational demands. Integrating robust environmental energy harvesting techniques with highly optimized, low-power inference engines provides a clear path toward self-sustaining devices that significantly lower the overall carbon footprint of large-scale IoT networks.

To accelerate the deployment of sustainable AIoT infrastructures, future research must focus heavily on advancing hardware-aware TinyML methodologies, including automated knowledge distillation, deep quantization, and multi-objective neural architecture search (NAS). There is an urgent need to develop and adopt standardized energy-carbon co-optimization metrics that look beyond simple computational efficiency to account for the true lifecycle impact of these technologies. Additionally, exploring multi-agent reinforcement learning (MARL) frameworks could enable cooperative, distributed energy management across heterogeneous edge nodes. Establishing open, cross-platform hardware testbeds will be essential for verifying these methods under real-world conditions, helping to transition AIoT networks from controlled laboratory models to reliable, sustainable field deployments.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that generative AI tools were used in the preparation of this manuscript under the following disclosure category: language editing. Specifically, GROK 4.1 was used for language and grammar editing. All research ideas, study design, methodology development, data collection, core analysis, interpretation of results, and final technical conclusions were carried out independently by the authors. The AI tools were used solely for language polishing, and did not influence the scientific validity or research outcomes of the study. The authors take full responsibility for the integrity and accuracy of the manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Alahmari, S., & Alghamdi, I. (2025). A Comprehensive Survey on Energy-Efficient and Privacy-Preserving Federated Learning for Edge Intelligence and IoT. *Results in Engineering*, 107849. [CrossRef]

[2] Baliyar Singh, R. K., Dash, J. K., & Reddy, K. H. K. (2026). The role of Edge-AI in edge enabled IoT systems: A comprehensive performance analysis. *Peer-to-Peer Networking and Applications, 19*(1), 35. [CrossRef]

[3] Bhushan, C. M., Koppuravuri, P., Prasanthi, N., Gazi, F., Hussain, M. M., Abdussami, M., ... & Faizi, J. (2025). Deploying TinyML for energy-efficient object detection and communication in low-power edge AI systems. *Scientific reports, 15*(1), 44299. [CrossRef]

[4] Thakur, D., Guzzo, A., Fortino, G., & Das, S. K. (2025). GRACE-FL: Green Resource-Aware Communication-Efficient Federated Learning. *IEEE Transactions on Artificial Intelligence, 7*(6), 3221-3236. [CrossRef]

[5] Fanariotis, A., Orphanoudakis, T., Kotrotsios, K., Fotopoulos, V., Keramidas, G., & Karkazis, P. (2023). Power efficient machine learning models deployment on edge IoT devices. *Sensors, 23*(3), 1595. [CrossRef]

[6] Kanso, H., Noureddine, A., & Exposito, E. (2023). Automated power modeling of computing devices: Implementation and use case for Raspberry Pis. *Sustainable Computing: Informatics and Systems, 37*, 100837. [CrossRef]

[7] Kasioulis, M., Symeonides, M., Pallis, G., & Dikaiakos,

M. D. (2023, August). Power estimation models for edge computing devices. In *European Conference on Parallel Processing* (pp. 257-269). Cham: Springer Nature Switzerland. [CrossRef]

[8] Muhoza, A. C., Bergeret, E., Brdys, C., & Gary, F. (2023). Power consumption reduction for IoT devices thanks to Edge-AI: Application to human activity recognition. *Internet of Things, 24*, 100930. [CrossRef]

[9] Natarajan, Y., KR, S. P., Wadhwa, G., Choi, Y., Chen, Z., Lee, D. E., & Mi, Y. (2024). Enhancing building energy efficiency with IoT-driven hybrid deep learning models for accurate energy consumption prediction. *Sustainability, 16*(5), 1925. [CrossRef]

[10] Schizas, N., Karras, A., Karras, C., & Sioutas, S. (2022). TinyML for ultra-low power AI and large scale IoT deployments: A systematic review. *Future Internet, 14*(12), 363. [CrossRef]

[11] Singh, A. R., Sujatha, M. S., Kadu, A. D., Bajaj, M., Addis, H. K., & Sarada, K. (2025). A deep learning and IoT-driven framework for real-time adaptive resource allocation and grid optimization in smart energy systems. *Scientific reports, 15*(1), 19309. [CrossRef]

[12] Yauri, R., & Palomino, N. B. L. S. (2026). Performance analysis of classification models to determine the health status of edge computing devices. *Bulletin of Electrical Engineering and Informatics, 15*(3), 2505-2514. [CrossRef]