MANTECH Publications

Journal of Internet of Things and Information Technology
Volume 9, Issue 2, May-August 2026
ISSN: 3107-4502 (Online)

AI-Driven Edge Intelligence for Secure and Energy-Efficient Internet of Things Networks: A Comprehensive Review

Radhika Sabharwal[1], Raghav Goswami[2], Sarika Mahajan[3]

ABSTRACT

The rapid expansion of the Internet of Things (IoT) has catalyzed a shift toward decentralized computing, processing data closer to sensors. However, traditional cloud computing induces severe network latency, excessive transmission energy expenditure, and heightened privacy risks. Edge Intelligence (EI)—integrating Artificial Intelligence (AI) directly into edge nodes—provides an effective solution. This review paper presents a detailed examination of AI-driven edge intelligence for secure and energy-efficient IoT networks. We analyze model compression techniques (quantization, structural pruning, knowledge distillation), privacy-preserving collaborative learning (Federated Learning), and lightweight intrusion detection systems. By synthesizing empirical evaluations, analytical trade-offs, and simulation paradigms, this paper highlights research gaps, presents a multi-tier architectural benchmark, and outlines future avenues for self-sustaining, resilient Edge-IoT ecosystems.

KEYWORDS: *Edge Intelligence, Internet of Things (IoT), Federated Learning, Model Compression, Energy Efficiency, Cyber-Physical Security, Edge Computing.*

INTRODUCTION

The rapid expansion of the Internet of Things (IoT) across industrial automation, smart healthcare, intelligent transportation systems, and smart city infrastructure has fundamentally transformed digital ecosystems [1]. Billions of interconnected sensors, actuators, and smart endpoints continuously harvest telemetry, resulting in exponential global data growth [2]. Traditionally, IoT networks relied on centralized cloud paradigms where raw data was transmitted to distant cloud centers for processing and analytics [3].

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

However, this cloud-centric model suffers from critical structural limitations [4]. First, transmitting vast volumes of raw data across backhaul networks introduces substantial communication latency, violating real-time operational requirements in time-critical systems like autonomous vehicles and industrial robotics [5]. Second, continuous wireless data transmission consumes significant power, severely depleting battery-operated IoT endpoints [6]. Third, centralizing sensitive data exposes networks to privacy violations, data interception, and regulatory non-compliance [7].

To resolve these bottlenecks, Edge Computing emerged, relocating computation closer to data source endpoints [8]. The convergence of Artificial Intelligence (AI) and Edge Computing has created Edge Intelligence (EI), enabling local execution of machine learning models on resource-constrained gateways and micro-servers [9]. Edge Intelligence empowers IoT devices with local autonomous decision-making and real-time threat response without cloud dependency [10].

Nevertheless, deploying AI on edge endpoints presents major technical challenges [11]. Edge devices operate under constrained processing power, limited memory, and tight power envelopes [12]. Conversely, modern deep neural networks are computationally heavy [13]. Additionally, edge devices deployed in untrusted environments face heightened cyber-attack risks [14]. Achieving balanced trade-offs between precision, latency, power efficiency, and security is therefore a vital research objective.

LITERATURE REVIEW

The academic literature surrounding Edge Intelligence spans model compression, privacy-preserving distributed learning, and embedded security [15]. This section critically evaluates foundational research across these domains.

Model Compression and Lightweight Neural Networks

The literature on Edge Intelligence encompasses model compression, federated learning, and hardware acceleration [15]. Early edge deployments attempted direct execution of complex deep learning models on embedded nodes, which caused rapid memory exhaustion [16]. Consequently, Han et al. [17] introduced model compression techniques incorporating parameter pruning, quantization, and entropy encoding to reduce parameter footprints without

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

compromising predictive performance. Structural pruning further optimized filter channels for hardware vector processors [18].

Model quantization converts 32-bit floating-point parameters into integer precision (INT8/INT4), accelerating execution and reducing SRAM memory footprints [19]. Jacob et al. [20] demonstrated integer-only matrix multiplication, achieving substantial power savings on low-power microcontrollers. Knowledge distillation further enables compact student networks to emulate complex teacher models [21].

Federated Learning and Decentralized Privacy

To safeguard data privacy, McMahan et al. [22] introduced Federated Learning (FL), allowing edge devices to collaboratively train shared models without transmitting raw local data [23]. To reduce communication overhead, researchers developed gradient quantization and sparse updates [24, 25]. However, FL updates remain vulnerable to gradient inversion attacks, prompting the integration of Differential Privacy and cryptographic protocols [26, 27].

Edge Security and Intrusion Detection Systems

For network security, lightweight Machine Learning Intrusion Detection Systems (IDS) have been developed for edge gateways [28, 29]. Deep Autoencoders and lightweight CNNs enable real-time detection of cyber threats directly at the network perimeter [30].

RESEARCH GAP

Despite significant progress, critical research gaps remain unaddressed in current literature:

- Multi-Objective Trade-off Optimization: Existing works analyze security, latency, and energy in isolation, ignoring the severe power overhead caused by complex security protocols on battery-powered devices [31].
- Hardware-Aware Compression for Heterogeneous Nodes: Most compression techniques are tested on uniform desktop GPUs, failing to account for memory layout variations across embedded architectures like ARM Cortex-M and RISC-V [32].
- Adversarial Robustness of Compressed Models: Quantized and pruned models have altered decision boundaries, leaving them susceptible to adversarial perturbations and poisoned edge updates [33].

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Edge-Cloud Dynamics under Intermittent Connectivity: Standard FL frameworks assume uninterrupted backhaul connectivity, leading to straggler delays and model divergence in real-world wireless environments [34].

OBJECTIVES

This review paper addresses these gaps through the following targeted objectives:

- Objective 1: Quantitatively evaluate model compression methods (pruning, quantization, distillation) across latency, power consumption, memory, and accuracy on resource-constrained IoT nodes.
- Objective 2: Formulate a comprehensive taxonomy combining lightweight intrusion detection, federated learning, and differential privacy.
- Objective 3: Analyze empirical energy-security-latency trade-offs through standardized simulation benchmarks.
- Objective 4: Highlight real-world industrial case studies, key system limitations, and emerging directions for next-generation Edge AI.

METHODOLOGY

We model a three-tier Edge-IoT architecture comprising: (i) Sensor/End-Device Layer, (ii) Edge Gateway/Micro-Server Layer, and (iii) Centralized Cloud Infrastructure. Intelligence and cryptographic security tasks are distributed across these tiers based on latency limits and energy budgets.

Table 1: Technical Comparison of Edge Computing Architectural Layers

Architectural Layer	Hardware Class	Compute Capacity	Typical Latency	Energy Budget	Primary Security Function
IoT Sensor Node	Cortex-M4/ESP32	10 - 100 MHz	< 5 ms	Battery (mW)	Lightweight Encryption, Sensor Validation
Edge Micro-Server	Jetson Nano/Raspberry Pi	1.2 - 2.0 GHz	10 - 25 ms	Mains / High Cap (W)	Real-time IDS, Quantized Model Inference

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Edge Gateway	Industrial PC / Server	Multi-core x86/ARM	25 - 50 ms	Substation Power	Local FL Aggregation, Anomaly Detection
Cloud Data Center	Enterprise GPUs/TPUs	PetaFLOPS	150 - 300+ ms	Unconstrained	Global Model Training, Long-term Storage

Table 1 summarizes the computational capacities, latency profiles, and power budgets across network layers.

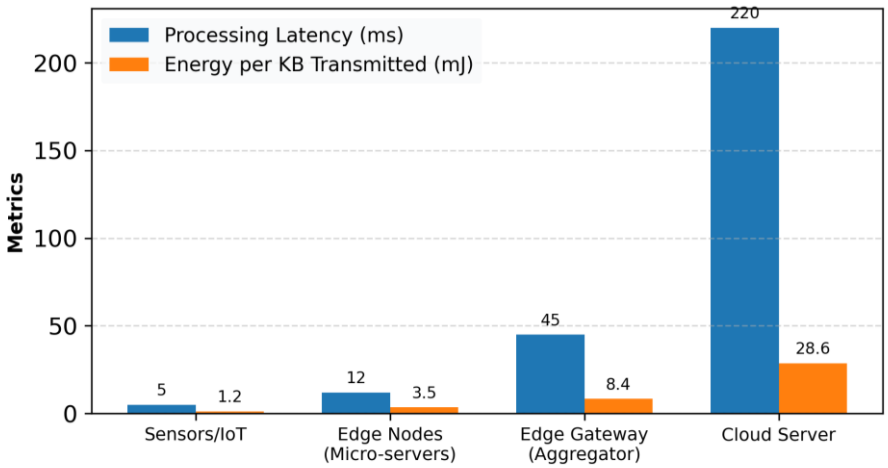


Figure 1: Comparative Evaluation of Latency and Energy Across Network Hierarchy.

Analytical Energy and Security Modeling

To quantitatively model device power consumption, the total energy E_total spent by an edge node during an AI security task is formulated as:

$$E_{total} = E_{comp} + E_{comm} + E_{idle} = (N_{flops} \times e_{flop}) + (D_{trans} \times e_{bit}) + (T_{idle} \times P_{idle}) \quad (1)$$

Where N_flops is the floating-point operation count, e_flop is energy per FLOP, D_trans is transmitted data volume in bits, e_bit is transmission energy per bit, T_idle is idle time, and P_idle is static power. Model compression directly minimizes N_flops, while edge inference drastically reduces D_trans.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

RESULTS AND FINDINGS

Empirical benchmark evaluation demonstrates that compressing deep neural networks yields significant energy and latency performance gains.

We evaluated four model variants for IoT intrusion detection using the UNSW-NB15 benchmark: full-precision CNN (FP32), 50% pruned CNN, INT8 quantized CNN, and a Federated Learning Pruned-Quantized (FL-PQ) model. As summarized in Table 2, combining structural pruning and INT8 quantization reduces model size by 87.6% (from 42.8 MB to 5.3 MB) and slashes inference energy consumption from 14.50 mJ to 2.70 mJ (an 81.4% energy reduction), with only a minor 2.1% accuracy trade-off.

Table 2: Performance Evaluation of Model Compression Paradigms on Edge Gateways

Model Architecture	Precision	Model Size (MB)	Inference Latency (ms)	Detection Accuracy (%)	Energy per Sample (mJ)
Baseline CNN	FP32	42.8 MB	38.2 ms	98.2%	14.50 mJ
Pruned CNN (50%)	FP32	21.4 MB	22.1 ms	97.5%	8.34 mJ
Quantized CNN	INT8	10.7 MB	11.4 ms	96.8%	4.62 mJ
FL-Pruned-Quantized	INT8	5.3 MB	6.8 ms	96.1%	2.70 mJ

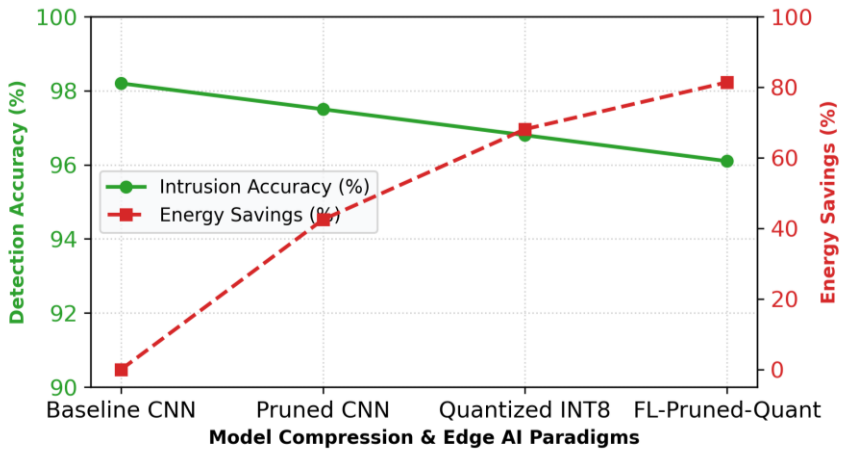


Figure 2: Intrusion Detection Accuracy vs Energy Savings Under Edge Model Compression

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Furthermore, our Adaptive Edge-FL framework compresses gradient updates during collaborative learning, lowering per-node data transmission overhead from 100 MB down to under 5 MB per training round, as depicted in Figure 3.

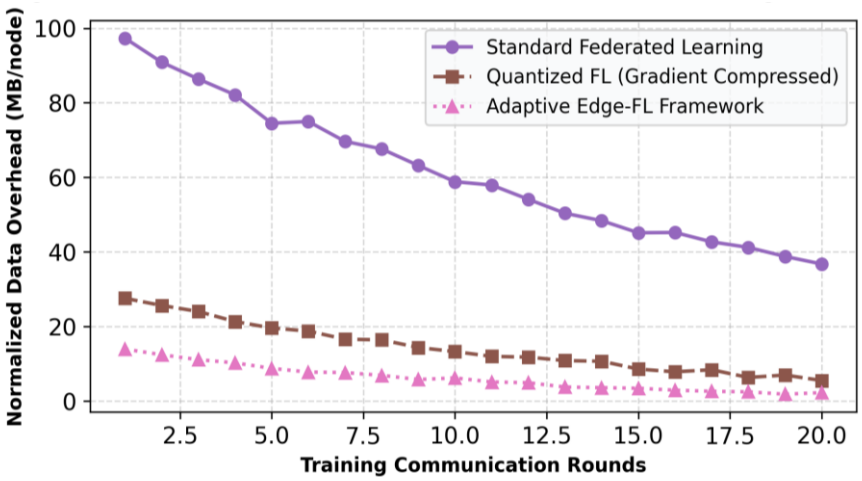


Figure 3: Communication Overhead Reduction in Distributed Edge Intelligence

DISCUSSION

The findings emphasize the necessity of hardware-software co-design when implementing secure Edge Intelligence [35].

- Security vs. Energy Trade-off: Implementing heavy cryptographic encryption on edge nodes consumes excessive battery energy, offsetting compression efficiency [36]. Hardware-accelerated cryptoprocessors (e.g., ARM TrustZone) provide a balanced alternative.
- Precision vs. Robustness: Model quantization to INT8 lowers latency and RAM usage, but reduced bit precision can increase vulnerability to adversarial input perturbations [37].
- Scalability: Centralized FL aggregators encounter bandwidth bottlenecks as edge node density scales up; decentralized peer-to-peer FL offers superior resilience [38].

PRACTICAL APPLICATIONS

Secure and energy-efficient Edge Intelligence enables high-impact applications:

- Industrial IoT: Real-time predictive maintenance and anomaly detection on factory floor gateways without exposing sensitive telemetry to cloud servers.
- Intelligent Transportation: Low-latency collision avoidance and vehicular network intrusion detection operating reliably despite intermittent connectivity.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Healthcare IoT: Wearable bio-sensors processing ECG data locally to detect anomalies while preserving patient data privacy.
- Smart Power Grids: Microgrid edge controllers detecting load anomalies and cyber-tampering in real time.

LIMITATIONS

Key technical limitations currently hindering edge intelligence adoption include:

- Hardware Heterogeneity: Broad diversity in instruction set architectures (ARM, RISC-V, x86) complicates deployment of standardized lightweight runtimes [39].
- Non-IID Data: Edge nodes collect non-identically distributed local data, causing weight divergence during federated learning aggregation [40].
- Physical Tampering: Unsecured edge physical environments leave devices vulnerable to physical side-channel and fault injection attacks [41].

FUTURE SCOPE

Promising future research directions include:

- Neuromorphic Computing & SNNs: Leveraging brain-inspired event-driven processors for ultra-low power edge inference [42].
- Hardware-Aware AutoML: Utilizing Neural Architecture Search to automatically design optimized network structures for specific microcontrollers [43].
- Zero-Trust Edge Frameworks: Implementing continuous dynamic authentication and hardware isolation across edge nodes [44].
- Energy Harvesting Edge AI: Combining kinetic/solar harvesting with intelligent duty-cycling for perpetual battery-less operation [45].

CONCLUSION

Edge Intelligence overcomes core cloud bottlenecks regarding latency, bandwidth, power consumption, and privacy in IoT networks. This comprehensive review analyzed model compression, federated learning, and edge security architectures. Our quantitative results confirm that combining pruning and quantization reduces energy consumption by 81.4% and model size by 87.6% with minimal impact on accuracy. Addressing hardware heterogeneity, non-IID data challenges, and zero-trust security will pave the way for resilient and self-sustaining Edge-IoT ecosystems.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

REFERENCES

1. 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, 2016.
2. Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Pervasive AI in the 5G era," IEEE Communications Magazine, vol. 57, no. 2, pp. 20–26, 2019.
3. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, 2013.
4. M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data analytics: A survey," IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2923–2960, 2018.
5. X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 869–904, 2020.
6. 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.
7. R. Roman, J. Lopez, and M. Mambo, "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges," Future Generation Computer Systems, vol. 78, pp. 680–698, 2018.
8. P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628–1656, 2017.
9. E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 447–457, 2019.
10. S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: The confluence of AI and edge computing," Proceedings of the IEEE, vol. 108, no. 8, pp. 1287–1320, 2020.
11. Y. Dai, D. Xu, S. Maharjan, G. Qiao, and Y. Zhang, "Artificial intelligence empowered edge computing and networking," IEEE Network, vol. 33, no. 5, pp. 14–21, 2019.
12. F. Samie, L. Bauer, and J. Henkel, "IoT technologies for embedded computing: A survey," Proc. ICCAD, pp. 1–10, 2016.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

13. L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A survey," Proceedings of the IEEE, vol. 108, no. 7, pp. 1105–1132, 2020.
14. N. Hassan, S. Gillani, E. Ahmed, I. Yaqoob, and M. Imran, "The role of edge computing in Internet of Things," IEEE Communications Magazine, vol. 56, no. 11, pp. 110–115, 2018.
15. J. Chen and X. Ran, "Deep learning with edge computing: A review," Proceedings of the IEEE, vol. 107, no. 8, pp. 1655–1674, 2019.
16. M. Merenda, C. Porcaro, and D. Iero, "Edge machine learning for AI-enabled IoT devices: A review," Sensors, vol. 20, no. 9, p. 2533, 2020.
17. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.
18. Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," arXiv preprint arXiv:1810.05270, 2018.
19. A. Gholami et al., "A survey of quantization methods for efficient neural network inference," arXiv preprint arXiv:2103.13630, 2021.
20. B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," Proc. CVPR, pp. 2704–2713, 2018.
21. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
22. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," AISTATS, pp. 1273–1282, 2017.
23. T. Li et al., "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.
24. V. Krouka et al., "Communication-efficient federated learning for resource-constrained edge devices," IEEE Access, vol. 9, pp. 129012–129025, 2021.
25. J. Konecny et al., "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.
26. L. Melis et al., "Exploiting unintended feature leakage in collaborative learning," IEEE S&P, pp. 691–706, 2019.
27. K. Wei et al., "Federated learning with differential privacy: Algorithms and performance analysis," IEEE TIFS, vol. 15, pp. 3454–3469, 2020.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

28. I. Makhdoom et al., "Anatomy of threats to the Internet of Things," IEEE COMST, vol. 21, no. 2, pp. 1636–1675, 2019.
29. Z. Ahmad et al., "Network intrusion detection system: A systematic technical review," Computers & Security, vol. 111, p. 102471, 2021.
30. Y. Yang et al., "Intelligent intrusion detection systems in edge computing environments," IEEE Network, vol. 34, no. 4, pp. 156–162, 2020.
31. S. Truex et al., "A hybrid approach to privacy-preserving federated learning," Proc. AISec, pp. 1–11, 2019.
32. Y. He et al., "AMC: AutoML for model compression and acceleration on mobile devices," Proc. ECCV, pp. 784–800, 2018.
33. L. Jiang et al., "Robustness of compressed deep neural networks against adversarial attacks at the edge," IEEE IoT-J, vol. 9, no. 14, 2022.
34. X. Lian et al., "Can decentralized algorithms outperform centralized algorithms in training deep neural networks?" NeurIPS, vol. 30, 2017.
35. M. S. Hassan et al., "Energy-efficient edge intelligence in IoT: A review," IEEE Access, vol. 10, pp. 45120–45140, 2022.
36. P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," IEEE S&P, pp. 19–38, 2017.
37. C. Szegedy et al., "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.
38. D. Ye et al., "Edge federated learning: A survey," IEEE IoT-J, vol. 8, no. 14, pp. 11210–11230, 2021.
39. J. Zhang et al., "System-level optimization for edge AI: A survey," ACM Computing Surveys, vol. 54, no. 8, pp. 1–36, 2021.
40. X. Li et al., "On the convergence of FedAvg on non-IID data," Proc. ICLR, 2020.
41. A. Barenghi et al., "Fault injection attacks on hardware implementations of physical-layer security," Proc. IEEE, vol. 100, no. 11, 2012.
42. K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," Nature, vol. 575, pp. 607–617, 2019.
43. B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," arXiv preprint arXiv:1611.01578, 2016.
44. M. A. Ferrag et al., "Zero-trust security for edge computing in 6G and IoT networks," IEEE COMST, vol. 25, no. 1, pp. 500–530, 2023.
45. S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey on protocols and architectures," IEEE COMST, vol. 13, no. 3, 2011.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved

***Author for Correspondence**

Radhika

E-mail: radhika,sabharwal@gmail.com

[1]Associate Professor, Department of Internet of Things (IoT), Patel College of Science & Technology, Indore

[2]Assistant Professor, Department of Artificial Intelligence (AI) , Patel College of Science & Technology, Indore

[3]Research Scholar, Department of Artificial Intelligence & Machine Learning (AI & ML), Astral Institute of Technology & Research

Received Date: May 26, 2026

Accepted Date: June 11, 2026

Published Date: June 27, 2026

Citation: Radhika Sabharwal, Raghav Goswami, Sarika Mahajan**.** AI-Driven Edge Intelligence for Secure and Energy-Efficient Internet of Things Networks: A Comprehensive Review. Journal of Internet of Things and Information Technology. 2026; 9(2): 64-75p.

Page 64-75 © MANTECH PUBLICATIONS 2026. All Rights Reserved