IJMSERH

International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

Volume 14, Issue 1, January - March 2026

Impact Factor: 9.274

ijmserh@gmail.com

www.ijmserh.com

International Journal of Multidisciplinary and Scientific Emerging ResearcH (IJMSERH)

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 14, Issue 1, January-March 2026 ||

DOI: 10.15662/IJMSERH.2026.1401015

Energy-Efficient AI: Techniques for Green and Sustainable Machine Learning

Ms. Nikita Ravindra Rajurkar

Assistant Professor, Ranibai Agnihotri Institute of Computer Science & Information Technology, Wardha, Maharashtra, India

ABSTRACT: The rapid expansion of deep learning and large-scale artificial intelligence (AI) models has significantly increased computational demand and energy consumption, raising serious environmental and economic concerns. Training and deploying modern AI systems require extensive hardware resources, contributing to high carbon emissions and operational costs. This review provides a comprehensive analysis of energy-efficient techniques for green and sustainable machine learning. It presents a structured taxonomy covering model-level optimization (pruning, quantization, distillation), algorithm-level improvements, data-efficient strategies, hardware acceleration, and system-level energy management approaches. The study also examines energy measurement metrics, trade-offs between accuracy and power consumption, and emerging trends such as TinyML and edge intelligence. Furthermore, it identifies key research challenges, including standardized benchmarking and carbon-aware training. The review highlights future directions toward developing scalable, low-carbon AI ecosystems that balance performance, efficiency, and sustainability.

KEYWORDS: Energy-Efficient AI, Green AI, Sustainable Machine Learning, Model Compression, Edge Computing, Carbon Footprint, TinyML.

I. INTRODUCTION

Artificial Intelligence (AI), particularly deep learning, has achieved unprecedented breakthroughs in computer vision, natural language processing, and autonomous systems. However, these advancements have come at the cost of rapidly increasing computational demand and energy consumption. Recent studies highlight that training large-scale neural networks requires substantial GPU/TPU resources, leading to significant carbon emissions and environmental impact (1). As model sizes grow from millions to billions of parameters, concerns regarding sustainability, energy efficiency, and ecological responsibility have intensified. The concept of "Green AI" emphasizes reducing computational cost while maintaining competitive performance (2). Energy consumption in AI systems arises from both training and inference stages, with data movement and memory access often contributing more to power usage than arithmetic operations (3). Additionally, data center infrastructure, cooling mechanisms, and hardware inefficiencies further amplify the environmental footprint (4). Addressing these challenges requires optimization across multiple layers, including algorithm design, model architecture, hardware acceleration, and system-level scheduling.

Model compression techniques such as pruning, quantization, and knowledge distillation have emerged as effective strategies for reducing computational overhead without substantial accuracy degradation (5). Similarly, neural architecture search has evolved to incorporate energy-aware constraints to optimize performance–power trade-offs (6). Beyond model-level methods, system-level innovations such as specialized AI accelerators and edge-based inference platforms are advancing sustainable deployment paradigms (7).

Given the urgency of climate change and the exponential growth of AI workloads, a systematic review of energy-efficient techniques is essential. This paper synthesizes existing research, categorizes optimization strategies, evaluates trade-offs between accuracy and energy consumption, and identifies open research directions to support the development of environmentally sustainable AI systems.

II. REVIEW OF LITERATURE

Hinton, Vinyals, and Dean (2015) introduced knowledge distillation as a model compression technique where a smaller "student" network learns from a larger "teacher" model. The method transfers soft target probabilities, enabling compact models to retain high predictive accuracy with significantly reduced computational complexity. This approach directly

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 14, Issue 1, January-March 2026 ||

DOI: 10.15662/IJMSERH.2026.1401015

contributes to energy-efficient inference by lowering parameter count and reducing memory access overhead. Distillation is especially effective in deploying AI models on edge devices with constrained resources. (8)

Jacob et al. (2018) proposed integer-only quantization for deep neural networks, enabling inference using low-precision arithmetic (INT8). Their framework demonstrated substantial reductions in model size and latency while preserving accuracy. Quantization decreases energy consumption by minimizing memory bandwidth and arithmetic intensity, making it highly suitable for mobile and embedded AI applications. (9)

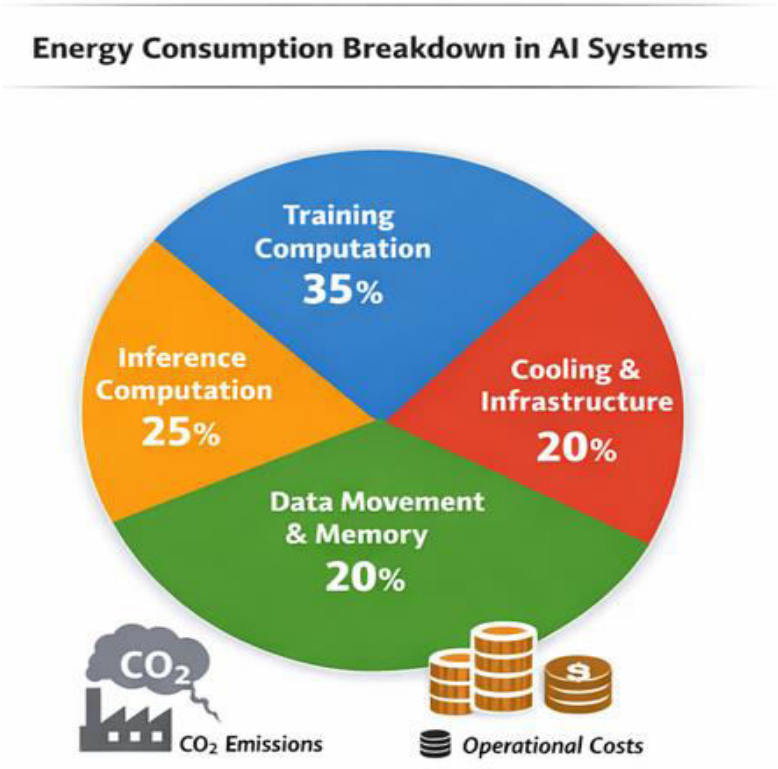
Howard et al. (2017) developed MobileNets, a class of lightweight convolutional neural networks using depthwise separable convolutions. This architectural innovation drastically reduces computation (measured in FLOPs) and model parameters compared to standard CNNs. MobileNets enable real-time, low-power vision applications, representing a foundational advancement in energy-aware model design. (10)

Rastegari et al. (2016) introduced XNOR-Net, a binary neural network that replaces floating-point operations with binary arithmetic. This approach significantly decreases memory usage and computational cost, leading to considerable energy savings. Binary networks demonstrate that extreme quantization can make AI viable on hardware with strict power constraints. (11)

Zoph and Le (2017) proposed Neural Architecture Search (NAS), automating model design using reinforcement learning. Although initial NAS methods were computationally expensive, subsequent energy-aware NAS adaptations optimized architectures based on performance–efficiency trade-offs. Their work laid the groundwork for automated energy-optimized model discovery. (12)

Lane et al. (2015) analyzed the feasibility of deploying deep learning models directly on smartphones. The study evaluated energy consumption, latency, and accuracy trade-offs in mobile sensing applications. Their findings emphasized the importance of on-device inference and adaptive computation strategies to balance battery usage and real-time performance. (13)

III. ENERGY CONSUMPTION IN AI SYSTEMS



| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

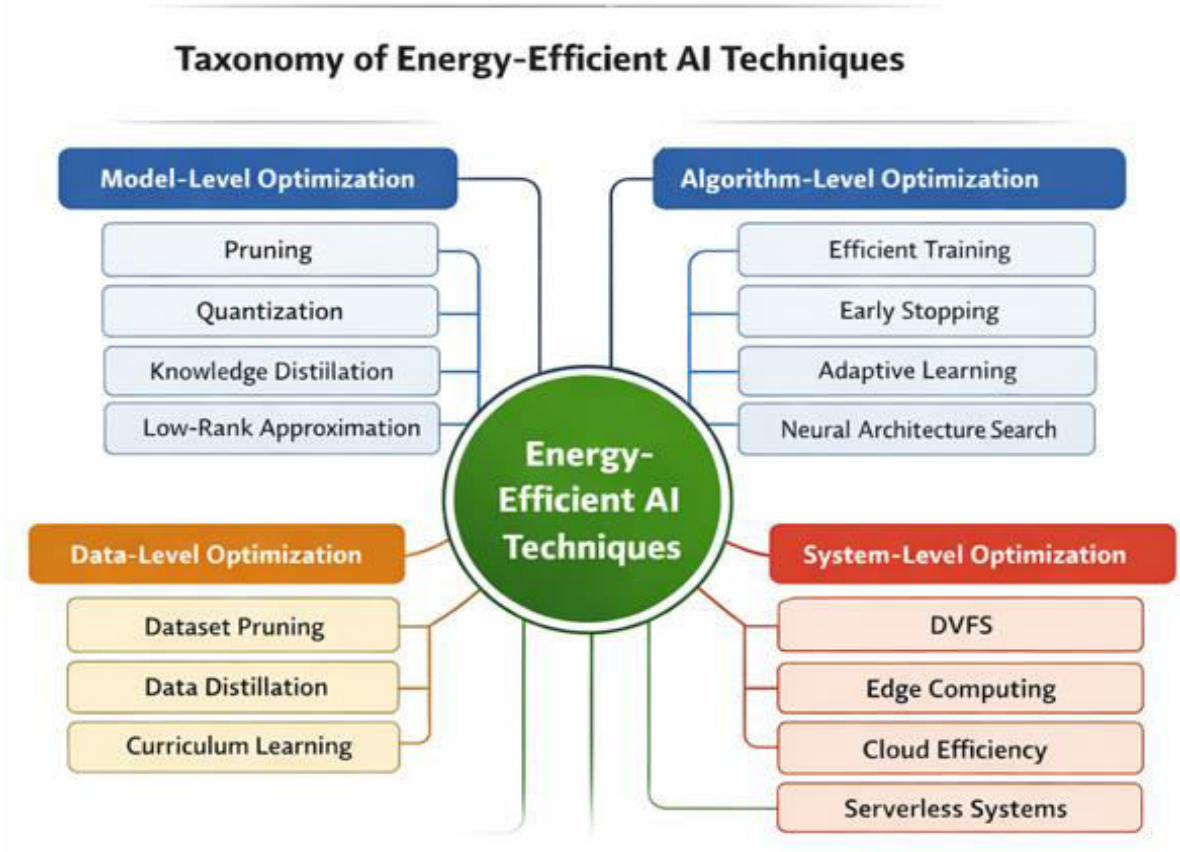
|| Volume 14, Issue 1, January-March 2026 ||

DOI: 10.15662/IJMSERH.2026.1401015

Energy consumption in AI systems has become a critical concern due to the rapid growth of deep learning models and large-scale computational workloads. The primary energy demand arises from two major phases: training and inference. Training deep neural networks, especially large transformer-based models, requires extensive matrix multiplications executed on high-performance GPUs or TPUs over prolonged periods. This process consumes substantial electrical power and generates significant heat, increasing cooling requirements in data centers. Inference, although less intensive than training, can cumulatively consume large amounts of energy when deployed at scale across millions of devices or cloud requests. A major contributor to energy usage is data movement between memory hierarchies rather than arithmetic computation itself. Accessing off-chip memory (DRAM) consumes significantly more energy than on-chip operations, making memory-efficient architectures essential for sustainability. Additionally, hardware inefficiencies, underutilized computational resources, and redundant model parameters further increase power consumption. Data center infrastructure compounds this issue, as energy is required not only for computation but also for cooling, networking, and storage.

Energy consumption is typically evaluated using metrics such as power draw (watts), total energy usage (kilowatt-hours), floating-point operations (FLOPs), and carbon emissions. However, FLOPs alone do not accurately represent real-world energy usage, as hardware architecture and workload characteristics strongly influence power efficiency. Therefore, comprehensive measurement frameworks are necessary to understand and optimize the environmental footprint of AI systems.

IV. TAXONOMY OF ENERGY-EFFICIENT AI TECHNIQUES



1. Model-Level Optimization: Model-level techniques focus on reducing the size and computational complexity of neural networks without significantly degrading accuracy. Common approaches include pruning (removing redundant weights), quantization (reducing numerical precision), knowledge distillation (training smaller student models), and low-rank factorization. Sparse neural networks also minimize unnecessary connections. These methods reduce memory usage, arithmetic operations, and inference latency. As a result, they directly lower energy consumption during both training and deployment phases.

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 14, Issue 1, January-March 2026 ||

DOI: 10.15662/IJMSERH.2026.1401015

2. Algorithm-Level Optimization: Algorithm-level optimization improves the learning process to reduce computational overhead. Techniques such as adaptive learning rate scheduling, early stopping, gradient checkpointing, and efficient backpropagation reduce unnecessary iterations. Energy-aware loss functions and optimization objectives are also emerging. Neural Architecture Search (NAS) can incorporate energy constraints during model discovery. These strategies minimize training time and hardware utilization, leading to measurable energy savings.

3. Data-Level Optimization: Data-centric techniques aim to reduce redundant or low-value training samples. Methods such as dataset pruning, dataset distillation, active learning, and curriculum learning help train models using fewer but more informative samples. Reducing dataset size lowers training cycles and computational cost. Efficient data preprocessing pipelines also minimize storage and transfer energy. These approaches are particularly valuable for large-scale AI systems.

4. Hardware-Level Optimization: Hardware-level methods leverage specialized accelerators to improve power efficiency. GPUs, TPUs, ASICs, FPGAs, and neuromorphic chips are designed to optimize parallel processing and minimize energy per operation. In-memory computing reduces data movement overhead. Low-power edge AI chips enable efficient on-device inference. Hardware–software co-design further enhances computational efficiency and reduces total system energy consumption.

5. System-Level Optimization: System-level techniques address energy efficiency across distributed AI infrastructures. Examples include dynamic voltage and frequency scaling (DVFS), workload consolidation, and energy-aware task scheduling in cloud data centers. Edge–cloud collaborative inference reduces unnecessary data transmission. Serverless computing and container orchestration improve resource utilization. These strategies ensure optimized performance while maintaining sustainability across the entire AI deployment lifecycle.

V. GREEN AI FRAMEWORKS AND TOOLS

Green AI frameworks and tools are designed to measure, monitor, and optimize the energy consumption of machine learning systems throughout their lifecycle. These tools enable researchers and practitioners to quantify power usage, estimate carbon emissions, and make informed decisions about model design and deployment. Energy profiling libraries integrate with popular deep learning frameworks to track GPU/CPU utilization, memory access patterns, and runtime energy consumption. Carbon tracking tools further estimate CO_2 emissions by incorporating regional electricity carbon intensity data. Such transparency encourages accountability and promotes environmentally responsible AI development. In addition to monitoring tools, hardware-aware machine learning frameworks optimize models based on device-specific constraints, enabling efficient deployment on edge devices, mobile platforms, and low-power accelerators. Automated machine learning (AutoML) systems increasingly incorporate energy metrics into architecture search and hyperparameter tuning processes. Cloud providers also offer energy-efficient infrastructure options, including optimized virtual machines and renewable-powered data centers. Collectively, these frameworks support a shift from performance-centric AI development toward sustainability-aware practices, facilitating scalable, low-carbon artificial intelligence systems.

VI. CHALLENGES AND OPEN RESEARCH ISSUES

1. Lack of Standardized Energy Benchmarking: There is no universally accepted benchmark for measuring energy consumption in AI systems. Variations in hardware, workloads, and reporting metrics make cross-study comparisons difficult and inconsistent.
2. Inaccurate Energy Measurement Methods: Many studies rely on indirect metrics such as FLOPs instead of real power measurements. This leads to discrepancies between theoretical efficiency and actual energy consumption.
3. Reproducibility of Energy Experiments: Energy results are highly dependent on hardware configuration and environmental conditions. Limited transparency in reporting experimental setups affects reproducibility.
4. Accuracy–Energy Trade-off: Reducing model size or precision can degrade predictive performance. Achieving an optimal balance between computational efficiency and model accuracy remains a critical challenge.
5. Carbon-Aware Training Strategies: Most AI training pipelines do not adapt to carbon intensity variations in power grids. Developing time-aware or location-aware training strategies is still an emerging area.
6. Sustainable Dataset Management: Large-scale datasets require significant storage and processing energy. Efficient data curation and lifecycle management are underexplored research directions.
7. Hardware–Software Co-Design Complexity: Optimizing AI systems across both hardware and software layers requires interdisciplinary expertise. Designing unified frameworks for co-optimization remains challenging.

| ISSN: 2349-6037 | www.ijmserh.com ||Impact Factor: 9.274||Quarterly, Peer Reviewed & Refereed Journal|

|| Volume 14, Issue 1, January-March 2026 ||

DOI: 10.15662/IJMSERH.2026.1401015

8. Scalability of Energy-Efficient Techniques: Techniques that work for small or edge models may not scale effectively to large foundation models. Ensuring efficiency at hyperscale levels is an open research problem.

VII. CONCLUSION

The rapid advancement of artificial intelligence has delivered transformative capabilities across industries, but it has also introduced significant energy and environmental challenges. As deep learning models continue to scale in size and complexity, the associated computational demand increases substantially, intensifying concerns regarding carbon emissions, operational costs, and sustainable deployment. This review systematically examined energy consumption sources in AI systems and categorized existing solutions into model-level, algorithm-level, data-level, hardware-level, and system-level optimization techniques. Each category contributes uniquely to reducing computational overhead while preserving performance. Green AI frameworks and carbon-aware tools further enable transparency in energy measurement and responsible development practices. However, persistent challenges such as the lack of standardized benchmarking, reproducibility issues, and accuracy–efficiency trade-offs highlight the need for continued research. Future AI systems must prioritize sustainability as a core design objective rather than an afterthought. Achieving scalable, low-carbon, and energy-aware intelligence will require interdisciplinary collaboration across machine learning, hardware architecture, and systems engineering to build a truly sustainable AI ecosystem.

REFERENCES

1. Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems, 28*, 1135–1143.
2. Horowitz, M. (2014). 1.1 Computing's energy problem (and what we can do about it). *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 10–14. https://doi.org/10.1109/ISSCC.2014.6757323
3. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
4. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831
5. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355
6. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE, 105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740
7. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114.
8. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
9. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
10. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2704–2713.
11. Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., Jiao, Y., Kawsar, F., & Mascolo, C. (2015). DeepX: A software accelerator for low-power deep learning inference on mobile devices. *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, 1–12.
12. Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). XNOR-Net: ImageNet classification using binary convolutional neural networks. *Proceedings of the European Conference on Computer Vision*, 525–542.
13. Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. *International Conference on Learning Representations*.

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

International Journal of Multidisciplinary and Scientific Emerging Research (IJMSERH)

Impact Factor: 9.274

ijmserh@gmail.com www.ijmserh.com