Synthesis: A Multidisciplinary Research Journal
Volume 4, Issue 1, January– March 2026, Pp: 46-57

Review Paper

Green Artificial Intelligence: A Critical Review of Energy-Efficient Machine Learning and Sustainable Computing Frameworks

[1*] M. SriRaghavendra, [2] Bobburi Divya Sree [3] Danduvari Diwakar

[1] *Associate Professor, Department of CSE (Data Science), Rajeev Gandhi Memorial College of Engineering & Technology, Nandyala, Andhra Pradesh, India. Email: sr.meeniga@gmail.com*
[2] *Student, Department of CSE (Data Science), Rajeev Gandhi Memorial College of Engineering & Technology, Nandyala, Andhra Pradesh, India, Email: bobburidhivyasree123@gmail.com*
[3] *Student, Department of CSE (Data Science), Rajeev Gandhi Memorial College of Engineering & Technology, Nandyala, Andhra Pradesh, India, Email: diwakarchintu768@gmail.com*

**Corresponding Author(s): sr.meeniga@gmail.com*

Article Info
Received: 18/11/2025
Revised: 26/02/2026
Accepted:13/03/2026
Published:31/03/2026

Abstract

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has significantly increased computational demands, leading to substantial energy consumption and environmental impact. Recent studies highlight that large-scale AI models contribute notably to carbon emissions due to intensive training and deployment processes. This has motivated the emergence of Green Artificial Intelligence (Green AI), which emphasizes the development of energy-efficient, resource-aware, and environmentally sustainable AI systems.This paper presents a comprehensive review of energy-efficient machine learning techniques and sustainable computing approaches across model, system, and hardware levels. The study systematically analyzes key methodologies, including model compression, pruning, quantization, knowledge distillation, efficient neural architectures, edge computing, and hardware-aware optimization strategies. In addition, carbon-aware evaluation frameworks and metrics for quantifying energy consumption and environmental impact are critically examined. A comparative analysis is conducted to evaluate these techniques based on energy efficiency, performance trade-offs, scalability, and implementation complexity. The findings indicate that no single approach is universally optimal, and effective Green AI solutions require a hybrid integration of multiple optimization strategies. Furthermore, the review identifies key challenges such as the lack of standardized benchmarks, hardware heterogeneity, and limited real-world validation. Overall, this work highlights the importance of integrating sustainability into AI system design and provides insights to guide future research toward developing scalable and environmentally responsible intelligent systems.

Keywords: Green Artificial Intelligence, Energy-Efficient Machine Learning, Sustainable Computing, Model Compression, Edge Computing, Carbon Footprint, Hardware-Aware Optimization, AI Sustainability

1. Introduction

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has led to unprecedented improvements across diverse application domains, including healthcare, transportation, finance, and smart infrastructure. However, the increasing complexity and scale of modern AI models, particularly deep neural

networks and transformer-based architectures, have resulted in significant computational and energy demands. Recent studies indicate that training large-scale models can consume substantial electrical power and generate considerable carbon emissions, raising critical concerns regarding the environmental sustainability of AI systems [1]–[3]. This growing energy footprint has motivated the emergence of Green Artificial Intelligence (Green AI), a paradigm that emphasizes the development of energy-efficient, resource-aware, and environmentally sustainable AI solutions.

Traditional AI research has primarily focused on maximizing predictive performance, often overlooking the associated computational cost and environmental impact. This performance-centric approach has led to the proliferation of over-parameterized models requiring extensive training cycles on high-performance hardware such as GPUs and TPUs [4]. Consequently, there is a pressing need to shift toward efficiency-aware design principles that balance model accuracy with energy consumption, computational efficiency, and carbon footprint [5]. Green AI addresses this challenge by promoting techniques that reduce energy usage during model training and inference while maintaining competitive performance.

In response to these concerns, a wide range of energy-efficient machine learning techniques have been proposed, including model compression, pruning, quantization, knowledge distillation, and lightweight architecture design [6], [7]. In parallel, system-level innovations such as specialized hardware accelerators, edge computing paradigms, and energy-efficient data center infrastructures have further contributed to reducing the environmental impact of AI deployments [8]. Additionally, the introduction of carbon-aware evaluation metrics and sustainability benchmarks has enabled researchers to quantify and compare the energy efficiency of different AI models more effectively.

Despite these advancements, the existing body of literature remains fragmented, with contributions dispersed across model-level optimization, hardware design, and system-level sustainability strategies. A comprehensive synthesis of these approaches is essential to provide a unified understanding of how energy efficiency can be systematically achieved in AI systems. Moreover, there is a lack of critical analysis regarding the trade-offs between model performance, computational cost, and environmental impact, which is crucial for guiding future research directions.

Motivated by these challenges, this review paper aims to provide a structured and critical analysis of Green AI methodologies, focusing on energy-efficient machine learning techniques and sustainable computing approaches.

The primary objectives of this study are as follows:

- To present a comprehensive overview of energy-efficient ML techniques, including model compression, optimization strategies, and lightweight architectures.
- To examine system-level and hardware-driven approaches for sustainable AI deployment, including edge computing and specialized accelerators.
- To analyze carbon-aware metrics and evaluation frameworks for quantifying AI energy consumption and environmental impact.
- To provide a comparative assessment of existing approaches, highlighting their strengths, limitations, and practical trade-offs.
- To identify key challenges and outline future research directions for advancing sustainable AI systems.

The remainder of this paper is organized as follows. Section II introduces the background and fundamental concepts of Green AI. Section III discusses energy-efficient machine learning techniques. Section IV presents system-level and hardware-based sustainability approaches. Section V explores carbon metrics and evaluation strategies. Section VI provides a comparative analysis of existing methods. Section VII outlines key challenges and future research directions. Finally, Section VIII concludes the paper.

2. Background on Green Artificial Intelligence

The concept of Green Artificial Intelligence (Green AI) has emerged as a response to the growing environmental concerns associated with large-scale AI systems. Unlike traditional AI paradigms that prioritize performance metrics such as accuracy and scalability, Green AI emphasizes the efficient utilization of computational resources, reduced energy consumption, and minimized carbon emissions throughout the AI lifecycle [9], [10]. This paradigm shift reflects a broader movement toward sustainable computing, where environmental impact is considered alongside model performance.

The rapid expansion of deep learning models, particularly in natural language processing and computer vision, has significantly increased computational requirements. State-of-the-art models often involve billions of parameters and require extensive training on high-performance computing infrastructures, leading to substantial energy usage and environmental cost [11]. Studies have shown that the carbon footprint of training a single large-scale model can be comparable to that of multiple households over extended periods, highlighting the urgency of adopting energy-aware AI practices [12]. Consequently, Green AI advocates for the development of models that achieve competitive performance with lower computational overhead.

Green AI can be broadly characterized by three fundamental dimensions: model efficiency, system-level optimization, and carbon awareness. Model efficiency focuses on reducing the complexity of machine learning models through techniques such as parameter reduction and architectural optimization. System-level optimization involves leveraging energy-efficient hardware, distributed computing frameworks, and edge-based processing to

minimize energy consumption during deployment. Carbon awareness introduces mechanisms to measure and manage the environmental impact of AI systems, enabling more informed decision-making during model development and deployment [13].

Another important aspect of Green AI is the shift toward cost-aware and resource-aware evaluation frameworks. Traditional evaluation metrics primarily assess predictive accuracy without considering the computational resources required to achieve such performance. In contrast, Green AI promotes the integration of energy consumption, training time, and carbon emissions into the evaluation process, thereby encouraging the development of more sustainable AI models [14]. This transition is essential for aligning AI research with global sustainability goals and reducing the ecological footprint of large-scale AI deployments.

Despite its growing importance, Green AI remains an evolving field with several open challenges. The lack of standardized benchmarks for measuring energy efficiency, the difficulty in accurately estimating carbon emissions across diverse hardware environments, and the trade-offs between performance and sustainability continue to hinder widespread adoption. Addressing these challenges requires a comprehensive understanding of both algorithmic and system-level innovations, which motivates the need for a structured review of existing techniques and approaches in this domain.

3. Energy-Efficient Machine Learning Techniques

The increasing computational demands of modern machine learning models have necessitated the development of energy-efficient techniques that reduce resource consumption while preserving predictive performance. These techniques primarily operate at the model level, focusing on optimizing network architecture, reducing parameter redundancy, and improving training efficiency. Among the most widely adopted approaches are model compression, quantization, knowledge distillation, and efficient architecture design, each contributing to the reduction of computational cost and energy consumption [15], [16].

Fig. 1 illustrates the overall Green AI framework, where data-driven learning is optimized through model-level and system-level techniques, while energy consumption and carbon emissions are continuously monitored to ensure sustainable AI deployment.

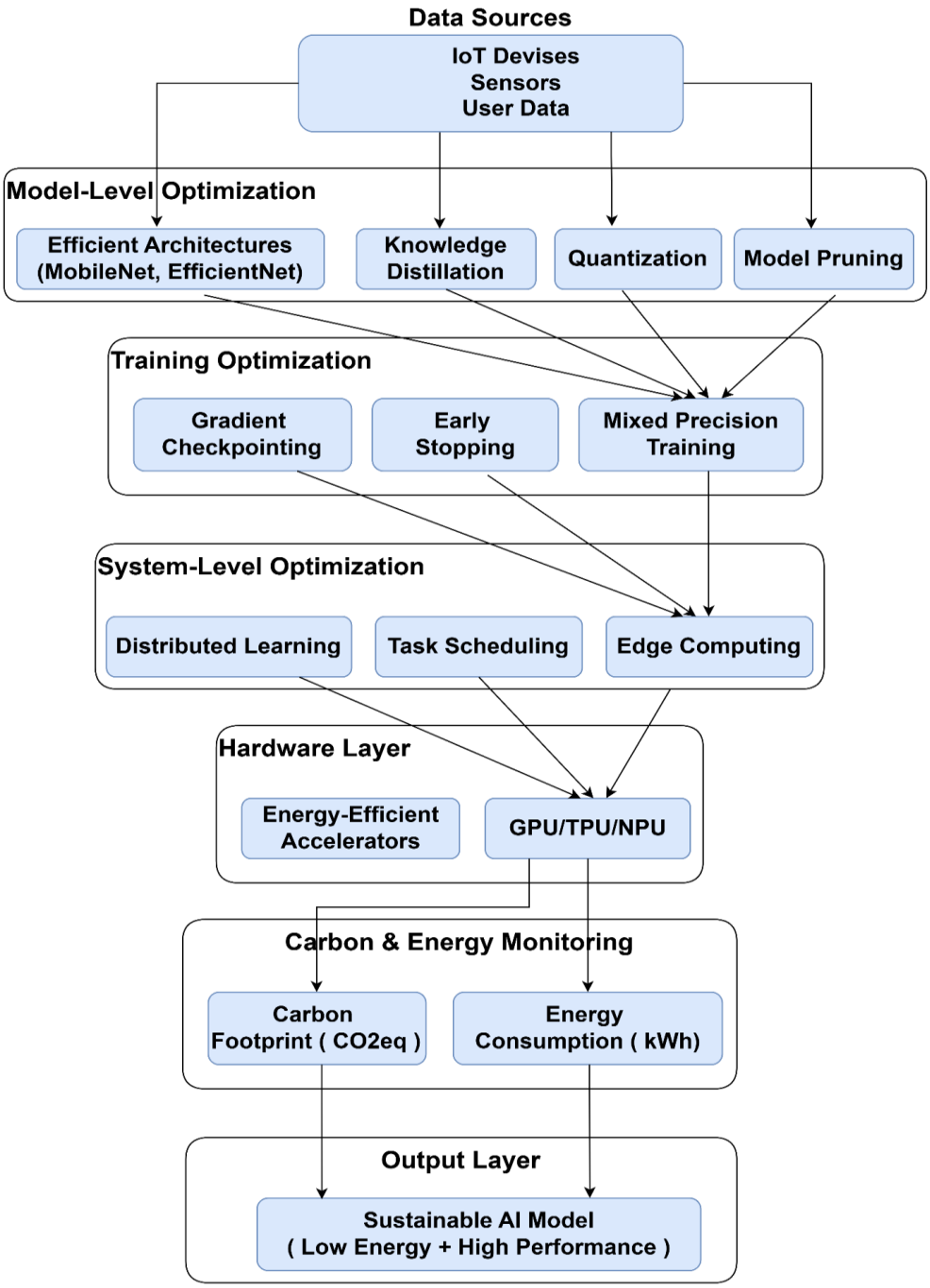

Fig. 1. Green AI system architecture integrating model-level, system-level, and hardware-level optimization for sustainable machine learning.

3.1 Model Compression and Pruning

Model compression aims to reduce the size and complexity of deep neural networks by eliminating redundant parameters and operations. One of the most effective compression strategies is model pruning, which removes less significant weights, neurons, or entire layers from a trained network. Pruning techniques can be broadly categorized into unstructured pruning, which removes individual weights, and structured pruning, which eliminates entire filters or channels to improve hardware efficiency [17], [18].

Recent studies demonstrate that pruning can significantly reduce model size and inference latency while maintaining comparable accuracy levels. For instance, magnitude-based pruning selectively removes weights with minimal contribution to model output, resulting in reduced computational overhead and improved energy efficiency. Furthermore, hybrid approaches that combine pruning with other optimization strategies have shown enhanced performance gains. A joint pruning–quantization framework has been shown to achieve substantial energy reduction while maintaining accuracy, highlighting the effectiveness of integrated optimization techniques.

3.2 Quantization Techniques

Quantization reduces the precision of model parameters and computations, typically converting high-precision floating-point representations into lower-bit formats such as 8-bit or even binary representations. This reduction in precision leads to significant savings in memory usage, computational complexity, and energy consumption [19].

Modern quantization approaches include post-training quantization and quantization-aware training, both of which aim to preserve model accuracy while minimizing precision. Quantization-aware training, in particular, incorporates low-precision constraints during model training, enabling the network to adapt to reduced numerical representation. Recent advancements have demonstrated that combining quantization with pruning can further enhance efficiency, achieving substantial reductions in both energy consumption and hardware resource utilization.

3.3 Knowledge Distillation

Knowledge distillation is a model compression technique in which a smaller student model is trained to replicate the behavior of a larger, more complex teacher model. This approach enables the deployment of lightweight models with significantly reduced computational requirements while retaining competitive performance [20].

In energy-constrained environments, such as edge devices and embedded systems, knowledge distillation plays a crucial role in enabling efficient inference. By transferring knowledge from high-capacity models to compact architectures, distillation reduces both training and inference costs. Recent studies highlight that distillation-based approaches can achieve substantial energy savings without compromising model generalization, making them highly suitable for sustainable AI applications [21].

3.4 Efficient Neural Network Architectures

Another critical direction in energy-efficient machine learning is the design of lightweight neural network architectures. Models such as MobileNet, EfficientNet, and ShuffleNet are specifically engineered to reduce computational complexity through techniques such as depthwise separable convolutions, neural architecture search, and parameter scaling [22].

These architectures are optimized for deployment on resource-constrained devices, enabling real-time inference with minimal energy consumption. Additionally, recent advances in hardware-aware model design incorporate constraints related to memory bandwidth, latency, and energy efficiency during the model development process. Such approaches ensure that the resulting architectures are not only accurate but also optimized for practical deployment scenarios.

3.5 Training Optimization Strategies

In addition to architectural improvements, several training optimization techniques have been proposed to reduce energy consumption during the learning process. These include early stopping, mixed-precision training, and gradient checkpointing, which collectively reduce computational overhead and memory usage.

Mixed-precision training, for example, leverages lower-precision arithmetic during training to accelerate computation and reduce energy consumption, while maintaining model accuracy through dynamic loss scaling. Similarly, gradient checkpointing reduces memory requirements by selectively recomputing intermediate activations, thereby enabling efficient training of large models on limited hardware resources.

Overall, energy-efficient machine learning techniques provide a comprehensive set of solutions for reducing the computational and environmental impact of AI systems. By combining model compression, efficient architectures, and optimized training strategies, it is possible to achieve a balance between performance and sustainability. However, the trade-offs between accuracy, efficiency, and deployment constraints remain a critical area of ongoing research, necessitating further investigation into integrated and adaptive optimization frameworks.

4. System-Level and Hardware-Based Sustainable AI Approaches

While model-level optimizations significantly reduce computational complexity, achieving truly sustainable AI systems requires innovations at the system and hardware levels. These approaches focus on optimizing the infrastructure on which AI models are trained and deployed, including edge computing, hardware accelerators, energy-efficient data centers, and resource-aware scheduling frameworks. By addressing energy consumption beyond algorithmic design, system-level strategies play a crucial role in enabling scalable and environmentally sustainable AI deployments.

4.1 Edge Computing for Energy Efficiency

Edge computing has emerged as a key paradigm for reducing the energy footprint of AI systems by shifting computation closer to data sources. Instead of transmitting large volumes of data to centralized cloud servers, edge AI enables local processing on resource-constrained devices, thereby reducing communication overhead, latency, and energy consumption [23], [24].

Recent studies highlight that edge-based AI systems can significantly improve energy efficiency by minimizing data transmission and enabling real-time decision-making. For example, energy-aware task scheduling and offloading strategies in edge environments dynamically allocate computational workloads based on device capabilities and energy constraints, leading to optimized resource utilization [25].

Furthermore, the integration of AI with Internet of Things (IoT) ecosystems has accelerated the adoption of edge computing, particularly in smart cities, healthcare, and industrial automation. In such environments, energy-efficient inference on edge devices is essential to ensure scalability and sustainability.

4.2 Specialized Hardware Accelerators

The development of AI-specific hardware accelerators, such as GPUs, TPUs, and Neural Processing Units (NPUs),

has significantly improved the energy efficiency of machine learning workloads. These accelerators are designed to optimize parallel computation and reduce data movement, which is a major contributor to energy consumption in AI systems [26].

Recent advancements in hardware design have focused on dataflow-aware architectures and domain-specific accelerators, which adapt computation patterns to minimize memory access and energy usage. For instance, flexible neural processing units (NPUs) enable adaptive dataflows that reduce redundant computations and improve overall energy efficiency [27].

Additionally, emerging paradigms such as neuromorphic computing and photonic accelerators offer promising directions for ultra-low-power AI systems by mimicking biological neural processes or leveraging optical computation mechanisms. These technologies aim to significantly reduce energy consumption while maintaining high computational throughput.

4.3 Energy-Efficient Data Centers and Cloud Infrastructure

Large-scale AI model training is predominantly conducted in cloud-based data centers, which are major contributors to global energy consumption. To address this challenge, significant efforts have been made to design energy-efficient data centers through optimized cooling systems, renewable energy integration, and workload-aware resource management [28].

Modern data centers employ techniques such as dynamic resource scaling, virtualization, and carbon-aware scheduling, which allocate computational resources based on energy availability and environmental impact. These strategies enable more efficient utilization of hardware resources and reduce overall energy consumption without compromising performance [29].

Moreover, cloud providers are increasingly incorporating renewable energy sources into their infrastructure, aligning AI operations with sustainability goals. Such initiatives play a critical role in reducing the carbon footprint of large-scale AI deployments.

4.4 Resource-Aware Scheduling and Distributed Systems

Efficient resource management is another critical component of sustainable AI systems. Resource-aware scheduling algorithms optimize the allocation of computational tasks across distributed systems, considering factors such as energy consumption, latency, and workload distribution [30].

In distributed and federated learning environments, energy efficiency is achieved by minimizing communication overhead and optimizing local computation. Techniques such as task offloading, load balancing, and energy-aware orchestration enable adaptive system behavior that reduces energy usage while maintaining system performance.

Additionally, approximate computing techniques have been explored to further enhance energy efficiency by allowing controlled reductions in computational precision, thereby reducing energy consumption without significantly affecting output quality.

Overall, system-level and hardware-based approaches complement model-level optimizations by addressing energy efficiency across the entire AI pipeline. The integration of edge computing, specialized hardware, and intelligent resource management strategies enables the development of scalable and sustainable AI systems. However, challenges related to hardware heterogeneity, deployment complexity, and trade-offs between performance and energy efficiency remain key areas for future research.

5. Carbon Metrics and Evaluation in Green AI

The increasing environmental impact of artificial intelligence systems has led to the development of carbon-aware evaluation frameworks that quantify energy consumption and greenhouse gas emissions associated with machine learning models. Unlike traditional performance metrics that focus solely on accuracy, Green AI introduces multi-dimensional evaluation criteria, incorporating energy usage, carbon footprint, and computational efficiency into the assessment process [31], [32]. This paradigm shift is essential for enabling transparent and sustainable AI development.

5.1 Carbon Footprint Estimation in AI Systems

Carbon footprint estimation is a fundamental component of Green AI, typically expressed in terms of CO_2 equivalent emissions (CO_2eq) generated during model training and inference. The total carbon footprint of an AI system depends on several factors, including computational workload, hardware efficiency, energy source, and geographic location of data centers [33].

Recent studies have demonstrated that large-scale deep learning models can generate substantial emissions, sometimes reaching hundreds of tons of CO_2eq during training, thereby emphasizing the need for accurate measurement methodologies. These findings highlight the importance of incorporating carbon estimation into the AI development lifecycle to better understand environmental impact.

Several tools and frameworks have been proposed to estimate AI-related carbon emissions, such as energy tracking libraries and lifecycle assessment models. These tools measure energy consumption at different stages of computation and convert it into carbon emissions based on regional energy mixes, providing a standardized approach for sustainability assessment [34].

5.2 Energy Consumption Metrics

Energy consumption is another critical metric used to evaluate the efficiency of machine learning models. It is typically measured in terms of kilowatt-hours (kWh) consumed during training and inference phases. Energy usage varies significantly depending on model complexity, dataset size, and hardware configuration [35].

To facilitate fair comparisons, researchers have proposed normalized metrics such as energy per training sample and energy per inference operation, which provide insights into the efficiency of different models under varying workloads. These metrics enable benchmarking across models and encourage the development of energy-efficient architectures. Moreover, recent works emphasize the importance of reporting both training energy and inference energy, as the latter becomes increasingly significant in large-scale deployment scenarios.

5.3 Carbon-Aware and Efficiency-Aware Benchmarks

The integration of carbon and energy metrics into benchmarking frameworks has led to the emergence of carbon-aware AI benchmarks, which evaluate models based on both performance and environmental impact. These benchmarks aim to standardize evaluation practices and promote transparency in reporting [36].

Green AI advocates for the inclusion of efficiency-performance trade-offs, where models are assessed not only based on accuracy but also on computational cost and environmental footprint. This approach encourages the development of models that achieve optimal performance with minimal resource consumption. Additionally, lifecycle-based evaluation frameworks consider the environmental impact of AI systems across all stages, including training, deployment, and maintenance. Such holistic evaluation strategies provide a more comprehensive understanding of sustainability in AI systems.

5.4 Carbon-Aware Optimization and Scheduling

Recent advancements have explored carbon-aware optimization techniques, which dynamically adapt model training and deployment based on energy availability and carbon intensity of power grids. For instance, scheduling computational tasks during periods of low carbon intensity can significantly reduce emissions without affecting model performance [37]. Similarly, cloud-based AI systems are increasingly adopting carbon-aware workload scheduling, which prioritizes energy-efficient resources and renewable energy sources. These approaches align AI operations with sustainability goals and contribute to reducing the overall carbon footprint of large-scale deployments.

5.5 Challenges in Carbon Measurement and Standardization

Despite the progress in carbon-aware evaluation, several challenges remain. One of the primary issues is the lack of standardized methodologies for measuring and reporting energy consumption and carbon emissions. Variations in hardware configurations, software frameworks, and energy sources make it difficult to establish consistent benchmarks across studies [38].

Furthermore, existing estimation tools often rely on simplified assumptions that may not accurately capture real-world energy usage. The absence of unified reporting standards and transparent evaluation frameworks continues to hinder the comparability and reproducibility of Green AI research.[39][40]

Addressing these challenges requires the development of standardized metrics, improved measurement tools, and comprehensive evaluation protocols that consider the entire lifecycle of AI systems.

Overall, carbon metrics and evaluation frameworks play a crucial role in advancing Green AI by enabling the quantification and comparison of environmental impact across different models and systems. By integrating energy and carbon considerations into model evaluation, researchers can promote the development of sustainable AI solutions that balance performance with environmental responsibility.

6. Comparative Analysis of Green AI Techniques

A comprehensive comparison of existing Green AI techniques is essential to understand the trade-offs between model performance, energy efficiency, and deployment feasibility. While individual approaches such as pruning, quantization, and edge computing demonstrate significant improvements in isolation, their effectiveness varies depending on the application domain, hardware environment, and system constraints [41][42][43].

Recent studies emphasize that Green AI should be evaluated as a multi-objective optimization problem, where accuracy, energy consumption, and carbon emissions must be jointly considered rather than optimized independently. This necessitates a structured comparison of techniques across multiple dimensions, including efficiency gains, scalability, and practical limitations [44].

6.1 Comparative Evaluation Framework

To systematically analyze Green AI techniques, the following evaluation dimensions are considered:

- **Energy Efficiency Gain**: Reduction in energy consumption (training/inference)
- **Model Performance Impact**: Accuracy trade-off after optimization
- **Hardware Dependency**: Requirement for specialized hardware
- **Scalability**: Suitability for large-scale deployment
- **Implementation Complexity**: Ease of integration into existing pipelines

These dimensions align with recent Green AI evaluation frameworks that advocate joint reporting of performance and environmental impact.

6.2 Comparative Analysis Table

Technique	Level	Energy Saving	Performance Impact	Hardware Dependency	Scalability	Key Limitation
Model Pruning	Model-level	High	Slight accuracy loss	Low	High	Requires fine-tuning

Quantization	Model-level	High	Minimal (if optimized)	Medium	High	Precision degradation
Knowledge Distillation	Model-level	Moderate–High	Minimal	Low	High	Depends on teacher model
Efficient Architectures	Model-level	High	Maintained	Medium	High	Design complexity
Edge Computing	System-level	High	Maintained	Medium–High	Medium	Limited device capacity
Hardware Accelerators	Hardware-level	Very High	Maintained	High	High	Cost and availability
Data Center Optimization	System-level	Moderate–High	No impact	High	Very High	Infrastructure dependency
Carbon-Aware Scheduling	System-level	Moderate	No impact	Medium	High	Requires energy data integration

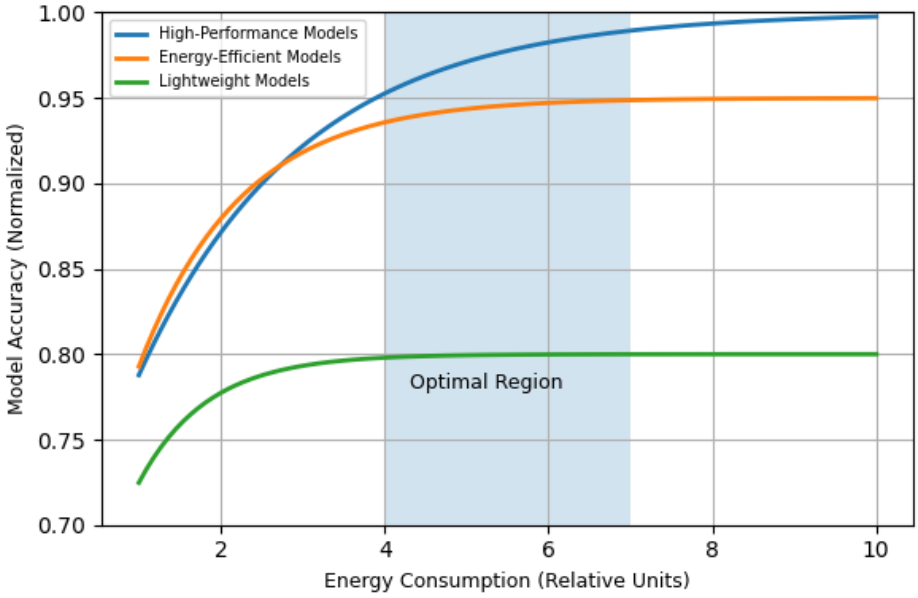

Fig.2. Trade-off between model accuracy and energy consumption in Green AI, illustrating the balance between high-performance models and energy-efficient optimization techniques

As shown in Fig. 2, there exists an inherent trade-off between model accuracy and energy consumption, where high-performance models achieve superior accuracy at the cost of increased computational resources, while energy-efficient techniques aim to operate within an optimal region that balances performance and sustainability.

6.3 Key Observations and Insights

6.3.1 No Single Technique is Universally Optimal

The comparative analysis reveals that no single Green AI technique consistently outperforms others across all scenarios. Model-level approaches such as pruning and quantization offer broad applicability and ease of integration, whereas system-level strategies such as edge computing and data center optimization provide greater impact in large-scale deployments. This indicates that the effectiveness of a given technique is highly dependent on application requirements, hardware constraints, and deployment environments. Consequently, a hybrid approach that combines multiple optimization strategies is often necessary to achieve optimal energy efficiency.

6.3.2 Trade-off between Efficiency and Accuracy

A fundamental challenge in Green AI is the trade-off between energy efficiency and model performance. Aggressive optimization techniques, such as high-level pruning or low-bit quantization, can significantly reduce energy consumption but may introduce accuracy degradation. Conversely, conservative optimization maintains performance but yields limited energy savings. This trade-off necessitates the development of adaptive optimization frameworks that balance efficiency and accuracy based on application-specific requirements.

6.3.3 Importance of Hardware-Aware Design

The analysis highlights that the effectiveness of energy-efficient techniques is closely tied to the underlying hardware infrastructure. Techniques such as quantization and neural architecture optimization often achieve maximum efficiency only when aligned with hardware capabilities, such as memory bandwidth and parallel processing support. As a result, Green AI solutions must consider hardware–software co-design, where model architectures are developed in conjunction with target hardware platforms to achieve optimal energy efficiency.

6.3.4 System-Level Approaches Offer Greater Real-World Impact

While model-level optimizations reduce computational complexity, system-level approaches provide a more comprehensive reduction in energy consumption across the AI lifecycle. Techniques such as edge computing, energy-efficient data centers, and resource-aware scheduling address energy usage during both training and deployment phases. These approaches are particularly effective in real-world scenarios where large-scale data processing and continuous inference are required, making them critical for sustainable AI deployment.

6.3.5 Emerging Shift toward Carbon-Aware AI Systems

A notable trend observed in recent research is the shift toward carbon-aware AI, where environmental impact is explicitly incorporated into model evaluation and deployment strategies. This includes the use of carbon footprint metrics, energy-aware benchmarking, and carbon-aware scheduling techniques. Such developments indicate a transition from traditional performance-driven AI toward sustainability-driven AI, where environmental considerations play a central role in system design and evaluation.

7. Challenges and Future Research Directions

Despite the rapid progress in Green Artificial Intelligence, several critical challenges continue to hinder its widespread adoption and practical deployment. Addressing these challenges is essential for achieving truly sustainable AI systems and aligning technological advancements with global environmental goals [45].

7.1 Key Challenges in Green AI

7.1.1 Lack of Standardized Evaluation Frameworks

One of the primary challenges in Green AI is the absence of universally accepted standards for measuring energy consumption and carbon emissions. Existing studies often employ different methodologies, metrics, and experimental setups, making it difficult to compare results across research works. This lack of standardization limits reproducibility and hinders the development of consistent benchmarking frameworks [46].

7.1.2 Hardware and Infrastructure Heterogeneity

The performance and energy efficiency of AI systems are highly dependent on underlying hardware configurations. Variations in GPUs, TPUs, edge devices, and data center infrastructures introduce inconsistencies in energy measurements and optimization outcomes. Moreover, the environmental impact of AI extends beyond computation to include hardware manufacturing, resource extraction, and electronic waste, which are often overlooked in current studies.

7.1.3 Trade-off between Performance and Sustainability

A persistent challenge in Green AI is balancing model accuracy with energy efficiency. While optimization techniques such as pruning and quantization reduce computational cost, they may degrade model performance if not carefully implemented [47]. This trade-off complicates the adoption of energy-efficient models in critical applications where accuracy is paramount.

7.1.4 Limited Real-World Deployment and Validation

Most Green AI techniques are evaluated in controlled experimental environments, with limited validation in real-world deployment scenarios. The lack of large-scale empirical studies makes it difficult to assess the practical effectiveness and scalability of proposed methods. Furthermore, industry adoption remains limited due to insufficient awareness and lack of practical implementation guidelines [48].

7.1.5 Insufficient Integration of Sustainability into AI Lifecycle

Current research primarily focuses on optimizing individual components such as models or hardware, rather than considering the entire AI lifecycle. Sustainable AI requires a holistic approach that includes data collection, model training, deployment, maintenance, and end-of-life considerations. The absence of lifecycle-based frameworks limits the overall effectiveness of Green AI initiatives [49][50].

7.2 Future Research Directions

7.2.1 Development of Standardized Green AI Benchmarks

Future research should focus on establishing standardized evaluation frameworks that integrate energy consumption, carbon emissions, and performance metrics. The development of universally accepted benchmarks will enable fair comparison across models and promote transparency in reporting.

7.2.2 Hardware–Software Co-Design for Sustainability

There is a growing need for integrated design approaches that jointly optimize machine learning algorithms and hardware architectures. Hardware-aware model design, combined with energy-efficient accelerators, can significantly improve system-level efficiency and enable scalable Green AI solutions.

7.2.3 Carbon-Aware and Adaptive AI Systems

Future AI systems should incorporate carbon-awareness into their operational frameworks. This includes dynamic scheduling based on energy availability, adaptive model selection, and real-time optimization strategies that minimize environmental impact without compromising performance.

7.2.4 Lifecycle-Based Sustainability Modeling

Developing comprehensive lifecycle assessment models for AI systems is essential to capture the full environmental impact, including hardware production, operational energy usage, and disposal. Such models will provide a more accurate representation of sustainability and guide responsible AI development.

7.2.5 Integration of Green AI with Policy and Governance

The alignment of AI development with environmental regulations and sustainability policies is crucial for large-scale adoption. Future research should explore governance frameworks, regulatory standards, and industry practices that promote responsible and sustainable AI deployment.

7.2.6 Interdisciplinary Research and Collaboration

Green AI is inherently interdisciplinary, requiring collaboration across computer science, environmental science, energy systems, and policy domains. Future research should emphasize cross-domain integration to address complex sustainability challenges and develop holistic AI solutions.

7.3 Summary

In summary, while Green AI presents promising solutions for reducing the environmental impact of artificial intelligence, significant challenges remain in standardization, scalability, and real-world adoption. Addressing these challenges through interdisciplinary research, standardized frameworks, and system-level innovations will be critical for advancing sustainable AI systems in the future.

8. Conclusion

This review presented a comprehensive analysis of Green Artificial Intelligence (Green AI), focusing on energy-efficient machine learning techniques and sustainable computing approaches across model, system, and hardware levels. The study highlighted that the rapid growth of AI systems has introduced significant energy consumption and environmental concerns, necessitating a shift toward sustainability-aware design and evaluation frameworks. Recent advancements demonstrate that techniques such as model compression, quantization, efficient architectures, and hardware-aware optimization can substantially reduce computational cost while maintaining competitive performance.

The comparative analysis revealed that no single technique is universally optimal, and effective Green AI solutions require a hybrid integration of model-level, system-level, and infrastructure-level optimizations. Furthermore, the emergence of carbon-aware metrics and evaluation frameworks has enabled more transparent assessment of environmental impact, although challenges related to standardization and reproducibility remain. The study also emphasized the importance of lifecycle-based sustainability, where environmental considerations extend beyond training to include deployment, maintenance, and hardware implications.

Despite notable progress, several open challenges persist, including the lack of unified benchmarking standards, hardware heterogeneity, and the trade-off between performance and energy efficiency. Addressing these issues requires interdisciplinary collaboration and the development of standardized frameworks that integrate energy, carbon, and performance metrics into a unified evaluation paradigm.

In conclusion, Green AI represents a critical direction for the future of artificial intelligence, where sustainability is treated as a fundamental design objective rather than an afterthought. Advancing this paradigm will require coordinated efforts in algorithm design, hardware innovation, and policy development to ensure that AI systems remain both technologically powerful and environmentally responsible.

Author Contributions: M. SriRaghavendra conceptualized the study, designed the overall structure of the review, and led the writing of the manuscript, including the development of key sections such as energy-efficient machine learning techniques and comparative analysis. Bobburi Divya Sree contributed to the literature survey, data curation, and drafting of sections related to system-level and hardware-based sustainable computing approaches, as well as assisted in critical revisions of the manuscript. Danduvari Diwakar contributed to the analysis of carbon metrics and evaluation frameworks, prepared the challenges and future research directions, and supported manuscript editing and refinement. All authors reviewed, revised, and approved the final version of the manuscript for publication and agree to be accountable for all aspects of the work.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes

References

[1] V. Bolón-Canedo, “A review of green artificial intelligence: Towards a more sustainable machine learning,” Neurocomputing, vol. 575, 2024, doi: 10.1016/j.neucom.2024.127268.

[2] R. Różycki, “Energy-aware machine learning models—A review of techniques and environmental impact,” Energies, vol. 18, no. 11, p. 2810, 2025, doi: 10.3390/en18112810.

[3] F. Scala et al., “An efficient model training framework for green AI,” Machine Learning, 2025, doi: 10.1007/s10994-025-06907-w.

[4] S. Dash, “Green AI: Enhancing sustainability and energy efficiency in AI-integrated enterprise systems,” IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.0322000.

[5] M. Ali Imran, A. Flávia dos Reis, G. Brante, P. Valente Klaine, and R. Demo Souza, “Machine Learning in Energy Efficiency Optimization,” Machine Learning for Future Wireless Communications, pp. 105–117, Dec. 2019, doi: 10.1002/9781119562306.ch6.

[6] S. Khan et al., “Green AI techniques for reducing energy consumption in artificial intelligence systems: A systematic review,” Array, vol. 21, 2025, doi: 10.1016/j.array.2025.100279.

[7] R. Verdecchia, J. Sallou, and L. Cruz, “A systematic review of Green <scp>AI</scp>,” WIREs Data Mining and Knowledge Discovery, vol. 13, no. 4, Jun. 2023, doi: 10.1002/widm.1507.

[8] T. Yarally et al., "Uncovering energy-efficient practices in deep learning training: Preliminary steps towards Green AI," arXiv preprint arXiv:2303.13972, 2023.

[9] A. Tabbakh, M. R. Hashemi, and S. Ghorashi, "Towards sustainable AI: A comprehensive framework for Green AI systems," AI and Ethics, vol. 5, 2024, doi: 10.1007/s43621-024-00641-4.

[10] R. Różycki, "Energy-aware machine learning models—A review of techniques and environmental impact," Energies, vol. 18, no. 11, p. 2810, 2025, doi: 10.3390/en18112810.

[11] V. Bolón-Canedo, "A review of Green Artificial Intelligence: Towards a more sustainable machine learning," Neurocomputing, vol. 575, 2024, doi: 10.1016/j.neucom.2024.127268.

[12] S. Khan, A. Rehman, and M. A. Khan, "Green AI techniques for reducing energy consumption in artificial intelligence systems: A systematic review," Array, vol. 21, 2025, doi: 10.1016/j.array.2025.100279.

[13] A. Sakhamuru and S. Vasireddy, "A comprehensive review of state-of-the-art generative AI models in natural language processing: Architectures, innovations, applications, and future directions," Frontiers in Health Informatics, vol. 13, no. 3, pp. 9498–9506, 2024.

[14] Z. Fan, Z. Yan, and S. Wen, "Deep Learning and Artificial Intelligence in Sustainability: A Review of SDGs, Renewable Energy, and Environmental Health," Sustainability, vol. 15, no. 18, p. 13493, Sep. 2023, doi: 10.3390/su151813493.

[15] R. Różycki, D. A. Solarska, and G. Waligóra, "Energy-Aware Machine Learning Models—A Review of Recent Techniques and Perspectives," Energies, vol. 18, no. 11, p. 2810, May 2025, doi: 10.3390/en18112810.

[16] Y. Zhang et al., "PQ-PIM: A pruning–quantization joint optimization framework for energy-efficient DNN accelerators," Microprocessors and Microsystems, vol. 91, 2022, doi: 10.1016/j.micpro.2022.104561.

[17] Saidjon Kamolov, "Optimizing model pruning for energy-efficient deep learning," Annals of Mathematics and Computer Science, vol. 25, pp. 112–119, Dec. 2024, doi: 10.56947/amcs.v25.428.

[18] J. Tmamna et al., "Pruning Deep Neural Networks for Green Energy-Efficient Models: A Survey," Cognitive Computation, vol. 16, no. 6, pp. 2931–2952, Jul. 2024, doi: 10.1007/s12559-024-10313-0.

[19] U. Bibi et al., "Advances in Pruning and Quantization for Natural Language Processing," IEEE Access, vol. 12, pp. 139113–139128, 2024, doi: 10.1109/access.2024.3465631.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," Neural Information Processing Systems Workshop, updated applications in recent studies, doi: 10.48550/arXiv.1503.02531.

[21] K. Balaskas et al., "Hardware-aware DNN compression via diverse pruning and mixed-precision quantization," ACM Transactions on Embedded Computing Systems, 2023, doi: 10.1145/3581784.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Proceedings of the International Conference on Machine Learning, updated efficiency adaptations in recent deployments, doi: 10.48550/arXiv.1905.11946.

[23] Y. Himeur et al., "Edge AI for Internet of Energy: Challenges and perspectives," Energy Reports, vol. 10, pp. 123–145, 2024, doi: 10.1016/j.egyr.2023.12.358.

[24] S. Chen et al., "Efficient scheduling of energy-constrained tasks in edge computing networks," Future Generation Computer Systems, vol. 152, 2024, doi: 10.1016/j.future.2024.01.011.

[25] H. J. Damsgaard et al., "Adaptive approximate computing in edge AI and IoT applications: A review," Journal of Systems Architecture, vol. 150, p. 103114, May 2024, doi: 10.1016/j.sysarc.2024.103114.

[26] A. Raha et al., "FlexNPU: A dataflow-aware flexible deep learning accelerator for energy-efficient edge devices," Frontiers in High Performance Computing, 2025, doi: 10.3389/fhpcp.2025.1570210.

[27] Y. Xie et al., "An energy-aware generative AI edge inference framework," Electronics, vol. 14, no. 20, 2025, doi: 10.3390/electronics14204086.

[28] Abhishake Reddy Onteddu, Rahul Reddy Bandhela and RamMohan Reddy Kundavaram, "Enhancing E-Commerce Product Recommendations through Data Engineering and Machine Learning," Economic Sciences, vol. 20, no. 1, pp. 171–183, Mar. 2024, doi: 10.69889/vqgzw857.

[29] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," IEEE Communications Surveys & Tutorials, vol. 23, no. 1, pp. 451–479, 2021, doi: 10.1109/COMST.2020.3035902.

[30] A. Sakhamuru and S. Vasireddy, "AI-Enabled Cross-Layer QoS Routing Framework for Mission-Critical 5G/6G-Integrated MANETs and UAV Swarms," 2025 International Conference on Sustainable Communication Networks and Application (ICSCN), pp. 787–794, Oct. 2025, doi: 10.1109/icscn67106.2025.11308381.

[31] A. Sánchez-Mompó et al., "Green MLOps to Green GenOps: An empirical study of energy consumption and carbon footprint in machine learning operations," Information, vol. 16, no. 4, p. 281, 2025, doi: 10.3390/info16040281.

[32] W. Gao and J. Wang, "The Environmental Impact of Micro/Nanomachines: A Review," ACS Nano, vol. 8, no. 4, pp. 3170–3180, Mar. 2014, doi: 10.1021/nn500077a.

[33] L. Bouza et al., "How to estimate carbon footprint when training deep learning models? A guide and review," Environmental Research Letters, vol. 18, no. 12, 2023, doi: 10.1088/1748-9326/acf81c.

[34] V. Liu et al., "Green AI: Exploring carbon footprints, mitigation strategies, and future directions," Discover Sustainability, vol. 4, no. 49, 2024, doi: 10.1007/s44163-024-00149-w.

[35] L. Gaur et al., "Artificial intelligence for carbon emissions using system-of-systems approach," Environmental Impact Assessment Review, vol. 98, 2023, doi: 10.1016/j.eiar.2023.106964.
[36] S. Borraccia et al., "Green metrics for AI: A hybrid strategy for environmental sustainability," Array, vol. 29, 2026, doi: 10.1016/j.array.2025.100652.
[37] R. Jha et al., "Forecasting data center CO_2 emissions using AI models for sustainable computing," Frontiers in Sustainability, 2025, doi: 10.3389/frsus.2024.1507030.
[38] S. Khan et al., "Green AI techniques for reducing energy consumption in artificial intelligence systems: A systematic review," Array, vol. 21, 2025, doi: 10.1016/j.array.2025.100279.
[39] V. Bolón-Canedo et al., "A review of green artificial intelligence: Towards a more sustainable future," Neurocomputing, vol. 575, 2024, doi: 10.1016/j.neucom.2024.127268.
[40] Rahul Reddy Bandhela, RamMohan Reddy Kundavaram, Abhishake Reddy Onteddu, "Enhancing Digital Wallet Payments through Data Analytics: A Study on Fraud Prevention and Personalized User Experience", Journal of Computational Analysis and Applications (JoCAAA), vol. 33, no. 2, pp. 1523–1535, Mar. 2024.
[41] F. Scala, S. Flesca, and L. Pontieri, "Play it Straight: An Intelligent Data Pruning Technique for Green-AI," Discovery Science, pp. 69–85, 2025, doi: 10.1007/978-3-031-78977-9_5.
[42] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future," Neurocomputing, vol. 599, p. 128096, Sep. 2024, doi: 10.1016/j.neucom.2024.128096.
[43] Srinivasarao Goda, Pratap Pachipulusu, Sakhamuru Amulya, and Pathan Hussian Basha, "Secure Blockchain-Based Consumer Electronics Platform for Smart Homes with Efficient Access Control and Performance Evaluation", Synth. Multidiscip. Res. J., vol. 3, no. 4, pp. 54–65, Dec. 2025
[44] S. M. Hasan, T. Islam, M. Saifuzzaman, K. R. Ahmed, C.-H. Huang, and A. R. Shahid, "Carbon Emission Quantification of Machine Learning: A Review," IEEE Transactions on Sustainable Computing, vol. 10, no. 6, pp. 1085–1102, Nov. 2025, doi: 10.1109/tsusc.2025.3578834.
[45] Abhishake Reddy Onteddu, "Scalable and Secure Group Key Agreement for Cloud Data Sharing Using Combinatorial Block Design", jier, vol. 1, no. 3, Dec. 2021.
[46] M. Toderas, "Artificial Intelligence for Sustainability: A Systematic Review and Critical Analysis of AI Applications, Challenges, and Future Directions," Sustainability, vol. 17, no. 17, p. 8049, Sep. 2025, doi: 10.3390/su17178049.
[47] A. Singh, A. Kanaujia, V. K. Singh, and R. Vinuesa, "Artificial intelligence for <scp>Sustainable Development Goals</scp>: Bibliometric patterns and concept evolution trajectories," Sustainable Development, vol. 32, no. 1, pp. 724–754, Jul. 2023, doi: 10.1002/sd.2706.
[48] S. Bibi et al., "Artificial intelligence shaping a smarter and greener planet for sustainable development," Discover Sustainability, vol. 6, 2025, doi: 10.1007/s44163-025-00647-5.
[49] N. Martin and V. Sharma, "Green AI: Navigating sustainability and environmental impact from an Indian perspective," E3S Web of Conferences, vol. 632, p. 02010, 2025, doi: 10.1051/e3sconf/202563202010.
[50] Y. I. Alzoubi and A. Mishra, "Green artificial intelligence initiatives: Potentials and challenges," Journal of Cleaner Production, vol. 468, p. 143090, Aug. 2024, doi: 10.1016/j.jclepro.2024.143090.