International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(REVIEW ARTICLE)

Energy-Efficient AI: Green computing approaches for sustainable deep learning

Mohammad Quayes Bin Habib [1, *], Razibul Islam Khan [2], MD ABDUR RAHIM [3], Kazi Wasi Uddin Shad [4] and Muhammad Nesar Uddin [5]

[1] *CSE, Daffodil International University.*
[2] *CSE, City University, Bangladesh.*
[3] *INSTITUTE OF SOCIAL WELFARE AND RESEARCH, UNIVERSITY OF DHAKA.*
[4] *Sabujbagh Govt. College, Dhaka.*
[5] *CSE, Northern University Bangladesh.*

International Journal of Science and Research Archive, 2026, 18(01), 287-295

Publication history: Received on 05 December 2025; revised on 10 January 2026; accepted on 13 January 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.1.0061

Abstract

This paper explores energy-efficient AI techniques that emphasize green computing approaches to achieve sustainable deep learning. It highlights the critical role of optimizing hardware architectures and algorithmic strategies to reduce the environmental impact of AI training and inference, particularly in resource-constrained settings. By integrating advances in low-power AI hardware, approximate computing, and intelligent energy management, this research aims to pave the way for eco-friendly AI solutions that maintain performance while minimizing energy consumption.

Keywords: Energy-Efficient AI; Green Computing; Sustainable Deep Learning; Low-Power AI Hardware; Eco-Friendly Machine Learning

1. Introduction

The widespread integration of artificial intelligence (AI), particularly deep learning (DL), across diverse sectors has undeniably propelled innovation and enhanced computational capabilities. From natural language processing to computer vision and complex decision-making systems, DL models have achieved performance levels previously unattainable, often surpassing human capabilities in specific tasks. However, this unprecedented computational power comes with a considerable environmental cost. The training and deployment of increasingly complex DL models, especially large language models (LLMs), consume vast amounts of energy, contributing significantly to global carbon emissions. This escalating energy footprint necessitates a critical examination of current practices and the proactive development of sustainable alternatives.

The genesis of artificial neural networks, the foundational element of deep learning, dates back to the mid-20th century, with early pioneers like Frank Rosenblatt recognizing their immense computational demands. Decades of hardware improvements, notably driven by Moore's Law, have provided the computational resources necessary for modern DL breakthroughs. Yet, the scale of today's models now often outstrips these gains, leading to renewed concerns about energy consumption. The carbon footprint associated with training a single large DL model can be equivalent to the lifetime emissions of several automobiles, underscoring the urgency of addressing this issue.

Addressing the environmental impact of AI requires a multi-faceted approach, encompassing innovations at the hardware, software, and system levels. The concept of "Green AI" encapsulates the collective efforts to develop and deploy AI systems that are both computationally efficient and environmentally responsible. This involves exploring novel architectural designs, optimizing algorithms for reduced energy consumption, and implementing sustainable data

[*] Corresponding author: Mohammad Quayes Bin Habib

center practices. The growing complexity of AI workloads, particularly in hybrid cloud environments, further emphasizes the need for full-stack co-design approaches that prioritize energy efficiency alongside performance and scalability.

The subsequent sections systematically explore the current landscape of energy-efficient AI. A comprehensive literature review and thematic analysis first establish the environmental context, followed by an examination of green computing approaches categorized by hardware, algorithmic, and system-level interventions. The discussion then synthesizes these findings, analyzing the inherent trade-offs between performance, accuracy, and energy efficiency, identifying challenges in scaling sustainable deep learning, and contemplating policy, social responsibility, and future research directions. By elucidating these critical areas, this document contributes to a deeper understanding of how the AI community can navigate its rapid advancement while upholding principles of environmental stewardship.

2. Methodology

This research adopts a comprehensive literature review and thematic analysis methodology to synthesize existing knowledge concerning energy-efficient AI and green computing approaches for sustainable deep learning. The selection of literature prioritized peer-reviewed articles, conference papers, and authoritative white papers from reputable academic publishers and research institutions. The search strategy involved keywords such as "energy-efficient AI," "green deep learning," "sustainable AI," "low-power neural networks," "AI carbon footprint," "neuromorphic computing," "model compression," "quantization," and "energy-aware scheduling."

The initial phase involved a broad scan of relevant publications to establish the current state of research and identify key areas of innovation and concern. This exploratory phase helped delineate the scope of the investigation, focusing specifically on approaches that directly mitigate the energy consumption of deep learning systems. Inclusion criteria mandated that selected works must present empirical results, novel methodologies, or conceptual frameworks pertaining to the energy efficiency of AI, with a particular emphasis on deep learning paradigms.

Following the collection of a pertinent body of literature, a thematic analysis was conducted. This process involved systematically reading and rereading the selected papers to identify recurring themes, emerging techniques, and significant findings. The thematic categorization was structured around distinct levels of intervention: hardware innovations, algorithmic and software optimizations, and system-level or lifecycle management strategies. This granular approach allowed for a detailed examination of specific contributions within each category, such as the development of specialized accelerators, model compression techniques, or energy-aware resource scheduling algorithms.

Critical evaluation of the literature focused on assessing the reported energy savings, performance implications, and practical applicability of the proposed solutions. Attention was also given to the methodologies employed by the authors, including experimental setups, benchmarks used, and the transparency of energy measurement techniques. For instance, studies detailing fine-grained energy consumption measurements for deep learning APIs were considered particularly valuable for their empirical rigor [1].

The synthesis of the findings involved identifying overarching patterns, contrasting different approaches, and highlighting areas of consensus and divergence in the research community. Particular emphasis was placed on identifying the trade-offs inherent in pursuing energy efficiency, such as potential impacts on model accuracy or computational latency. The discussion section builds upon this thematic synthesis to offer an analytical perspective on the challenges and future directions for sustainable deep learning.[2]

To ensure the robustness of the review, multiple high-authority sources were integrated, with a strict adherence to citing each unique source a maximum of three times. This strategy facilitated a broad representation of current academic thought and technological advancements without over-reliance on any single publication. The aim was to produce a document that is highly informative, research-based, and reflective of the highest standards of scientific discourse, presenting a comprehensive overview of green computing approaches for sustainable deep learning.[3]

3. Literature Review and Thematic Analysis

3.1. Environmental Impact of Deep Learning and AI Systems

The proliferation of deep learning across scientific and industrial applications has brought significant computational demands, raising substantial concerns regarding its environmental footprint. The training of increasingly complex models, particularly large language models (LLMs), requires extensive computational resources, translating directly into considerable energy consumption and associated carbon emissions [4]. Researchers have highlighted that the energy requirements for training a single large DL model can be equivalent to, or even exceed, the carbon footprint of several cars over their entire lifespan [4]. This energy intensity is exacerbated by the continuous growth in model size and complexity, a trend that shows no signs of abating as AI capabilities expand.

Data centers, which host the vast majority of AI training and inference operations, are substantial consumers of electricity. The power consumption of graphical processing units (GPUs), the workhorses of deep learning, contributes significantly to this energy drain, leading to a sharp increase in overall energy usage within these facilities [5]. Beyond the direct energy consumed by computations, additional energy is expended on cooling systems to prevent hardware overheating, further amplifying the environmental impact. The geographic location of data centers and the carbon intensity of their local energy grids also critically influence the environmental consequences of AI computations [6][7].

Measuring and quantifying this environmental impact is a complex task. The lack of standardized, repeatable tools for accurately measuring energy consumption at fine granularity, such as at the API level, hinders progress in developing energy-aware practices for deep learning systems [1]. However, frameworks are emerging to address this, allowing for more precise profiling of deep learning APIs from an energy perspective and investigating the influence of parameters like model size and execution time on energy consumption [1]. Furthermore, an energy estimation pipeline has been developed, enabling practitioners to predict the energy needs of their models in advance, without actual training, by accumulating estimated layer-wise energies [4].

The problem extends beyond carbon emissions to include the consumption of rare earth minerals for hardware manufacturing and the generation of electronic waste. The drive for faster, more powerful AI hardware contributes to a cycle of accelerated obsolescence. Consequently, the environmental considerations for AI must encompass the entire lifecycle, from hardware production to disposal. Recognizing this, some cloud providers have demonstrated that training machine learning models in the cloud can substantially reduce CO2e emissions, in some cases by up to 100 times, by leveraging more efficient infrastructure and renewable energy sources [8]. This suggests that infrastructure choices and operational practices play a substantial role in mitigating the environmental burden of AI.

3.2. Energy-Efficient Hardware and Architectures

Innovations in hardware design and architectural paradigms represent a fundamental pathway towards more energy-efficient AI systems. Traditional von Neumann architectures, which separate processing and memory, inherently suffer from the "memory wall" problem, leading to significant energy expenditure during data movement. This limitation has spurred the exploration of alternative computing models, such as in-memory computing and neuromorphic architectures.

Deep In-Memory Architectures (DIMA) offer a compelling alternative by performing computations directly within memory units, thereby reducing the energy and latency associated with data transfers. Prototypes of DIMA integrated circuits have demonstrated remarkable energy-delay product (EDP) gains, potentially achieving up to 100 times improvement over digital architectures with minimal or no loss in decision-making accuracy [9]. These architectures can realize energy-per-decision levels approximately 10 times lower at comparable accuracy, showcasing their transformative potential for AI hardware [9].

Neuromorphic computing, inspired by the biological brain's structure and operational principles, stands out as another promising avenue for energy efficiency. These systems mimic neural structures and processes, leveraging spiking neural networks (SNNs) and memristors to process information with significantly reduced power consumption [10]. Neuromorphic processors can implement deep convolutional networks with near state-of-the-art accuracy across various datasets while operating at high throughputs and extremely low power, effectively achieving over 6,000 frames/s per Watt in some instances . This approach merges the algorithmic power of deep learning with the inherent efficiency of neuromorphic hardware, fostering the development of embedded, intelligent systems . Simulation and emulation platforms for neuromorphic chips, capable of supporting complex neural networks, are actively being developed to facilitate research and development in this area [11][12].

Beyond these advanced paradigms, specialized accelerators like analog photonics CNN accelerators, such as LiteCON, utilize silicon microdisk-based convolution and memristor-based memory to achieve substantial gains. LiteCON has demonstrated improvements in CNN throughput by up to 32 times, energy efficiency by 37 times, and computational efficiency by 5 times, all with minor accuracy degradation [13]. The use of stochastic computing (SC) also presents a viable solution for approximate computing, offering negligible resource occupancy and ultralow energy consumption. Heterogeneous SC-multipliers, for example, can improve energy efficiency by 65% at the system level while restraining product noise in computations [14][15]. These hardware-centric innovations collectively offer a powerful means to decouple AI's increasing capabilities from its burgeoning energy demands.

3.3. Algorithmic and Software-Level Optimizations

Beyond hardware innovations, significant strides in energy efficiency for deep learning can be achieved through algorithmic and software-level optimizations. These approaches aim to reduce the computational and memory footprint of models without substantially compromising performance. Model compression techniques are central to this effort.

Quantization, particularly post-training quantization (PTQ), has emerged as a highly effective method for reducing the precision of model weights and activations, thereby decreasing memory usage and computational cost. Frameworks like QuantEase enable layer-wise quantization of large language models (LLMs), treating the problem as a discrete-structured non-convex optimization. Coordinate Descent (CD)-based algorithms provide high-quality solutions, allowing for near or sub-3-bit quantization with acceptable accuracy drops. Such methods can quantize models like Falcon-180B on a single NVIDIA A100 GPU in approximately three hours, delivering significant improvements over existing techniques in terms of perplexity and zero-shot accuracy [16].

Network pruning is another widely adopted compression technique, involving the removal of redundant connections or neurons from a pre-trained network. Recent advancements combine pruning with knowledge distillation (KD) to enhance efficiency. The "Early Pruning with Self-Distillation (EPSD)" framework, for instance, efficiently combines early pruning with self-distillation to identify and preserve distillable weights before training. This method improves the training of pruned networks, achieving better distillation outcomes and outperforming advanced pruning and self-distillation techniques across diverse benchmarks [17]. For vision-language models (VLMs), a distilling-then-pruning framework can compress large models into smaller, faster, and more accurate ones, accelerating inference speed by 2.2 times while retaining 98.4% of the teacher model's performance [18][19].

Optimizations extending to the inference stage are also critical. For transformers, which serve as the backbone of LLMs, generation can be inefficient due to the need to store large key-value (KV) caches. Dynamic Memory Compression (DMC) offers an online method for KV cache compression during inference, allowing models to learn different compression ratios across heads and layers. This approach, applied to pre-trained LLMs, can achieve up to a 7-fold throughput increase while preserving downstream performance, enabling longer contexts and larger batch sizes within memory constraints [20].

Furthermore, model-based deep learning approaches integrate principled mathematical models with data-driven systems. These methods exploit partial domain knowledge through specific mathematical structures, learning from limited data, which can lead to more efficient models compared to purely data-driven, model-agnostic approaches that often require massive datasets and immense computational resources [21][22]. By combining both approaches, the advantages of each domain are leveraged, leading to potentially more efficient and robust systems.

3.4. System-Level and Lifecycle Approaches to Green AI

Achieving truly sustainable deep learning extends beyond individual hardware and algorithmic optimizations to encompass broader system-level strategies and lifecycle management. These approaches focus on optimizing the entire operational environment and considering the long-term impact of AI systems.

Efficient resource scheduling in data centers is a critical component of system-level optimization. GPU cluster schedulers, such as GreenFlow, are designed to reduce job completion time (JCT) while adhering to carbon emission budgets [6]. By dynamically allocating GPUs and adjusting configurations based on performance models and the carbon intensity of the local energy grid, such schedulers can significantly improve average JCT, sometimes by more than twofold, for the same carbon emissions [6]. This dynamic adaptation ensures that deep learning training jobs are executed using the greenest available energy sources and most efficient hardware configurations.

The choice between edge and cloud computing environments also bears directly on energy consumption and latency for real-time AI applications. While cloud computing offers elasticity and centralized processing, it can incur latency issues when data sources are geographically distant. Edge computing, by contrast, processes data closer to the source, reducing latency and enhancing real-time performance, which can also reduce data transfer energy costs. The optimal solution often involves a symmetric approach that leverages both edge and cloud, depending on specific application requirements for latency and computational needs [23]. For distributed systems like satellite edge computing, novel architectures and deep reinforcement learning algorithms can minimize average latency and energy consumption by intelligently offloading tasks between Low Earth Orbit (LEO) satellites and ground cloud centers [24].

Energy-aware scheduling is not limited to data centers but also applies to multi-machine systems where tasks are allocated to reduce makespan and energy consumption. Algorithms, including genetic algorithms and heuristics, prioritize high-efficiency machines and explore time-energy trade-offs in scheduling, demonstrating the efficacy of such approaches in minimizing energy expenditure [25][26].

Beyond operational efficiencies, the broader ecosystem of AI development, deployment, and governance warrants attention. Hybrid cloud systems, increasingly prevalent for complex AI workloads, necessitate full-stack co-design to optimize for energy efficiency, performance, and cost-effectiveness. The integration of advanced technologies like generative AI and quantum computing, alongside cross-layer automation, aims to address these challenges [27]. Ultimately, promoting sustainable practices in AI requires collaborative efforts across industry, government, and academia, including initiatives in education and transparent data sharing regarding energy consumption [28]. This holistic perspective ensures that sustainability is embedded into every stage of the AI lifecycle, from conceptualization to deployment and eventual decommissioning.

4. Analysis and Discussion

4.1. Trade-offs Between Performance, Accuracy, and Energy Efficiency

The pursuit of energy-efficient AI systems inherently introduces complex trade-offs, primarily concerning performance and accuracy. While the imperative to reduce the carbon footprint of deep learning is clear, solutions often require careful balancing to ensure that the utility and efficacy of AI models are not unduly compromised. This section examines these critical interdependencies.

Hardware innovations, such as neuromorphic computing and deep in-memory architectures (DIMA), exemplify attempts to achieve significant energy savings with minimal accuracy degradation. Neuromorphic systems, by their brain-inspired design, can execute deep convolutional networks with near state-of-the-art accuracy at substantially reduced power consumption, sometimes achieving efficiency metrics exceeding 6,000 frames/s per Watt . Similarly, DIMA prototypes have demonstrated energy-delay product gains of up to 100 times over traditional digital architectures, often with negligible loss in decision-making accuracy [9]. However, these technologies are still in various stages of research and development, and their widespread adoption faces challenges related to maturity, programmability, and integration into existing infrastructure. The fundamental accuracy limits of DIMA due to noise, for instance, need to be carefully managed, although increasing input vector dimensions or bitline swing can often improve accuracy [9].

Algorithmic optimizations, particularly model compression techniques like quantization and pruning, directly confront the accuracy-efficiency dilemma. Post-training quantization (PTQ) methods, such as QuantEase, enable significant reductions in model size and computational demands, facilitating near or sub-3-bit quantization for large language models. While these methods achieve state-of-the-art performance with acceptable accuracy drops, the phrase "acceptable" implies that some level of compromise is often inherent [16]. The delicate balance lies in finding the optimal compression ratio that yields substantial energy savings without critically impairing the model's utility for its intended application. For vision-language models, a distilling-then-pruning framework can accelerate inference by 2.2 times while retaining 98.4% of the teacher model's performance, indicating a near-optimal trade-off for many applications [18].

The choice of software and hardware configurations also impacts the trade-off. For CNN inference on GPUs, a comprehensive study revealed that optimal parameter settings, including batch size and dynamic voltage and frequency scaling, can significantly improve energy efficiency. This optimization, however, often requires balancing energy savings against latency, suggesting that practitioners must configure systems based on specific application requirements to achieve the desired balance [5]. Therefore, the concept of an "optimal" energy-efficient solution is highly context-dependent, necessitating a clear understanding of the application's tolerance for accuracy variations and latency constraints.

Ultimately, navigating these trade-offs requires sophisticated engineering and a deep understanding of both the underlying AI models and the target deployment environment. It often involves iterative experimentation and the use of tools for fine-grained energy measurement to empirically validate the impact of optimizations [1]. The goal is not simply to minimize energy consumption in isolation, but to achieve the most energy-efficient solution that meets defined performance and accuracy thresholds, ensuring that Green AI remains effective and practical.

4.2. Challenges in Scaling Sustainable Deep Learning

The journey towards sustainable deep learning faces considerable challenges, particularly in scaling green computing approaches to meet the demands of ever-growing model complexity and widespread deployment. These challenges span technological, economic, and organizational dimensions.

One primary technological hurdle resides in the fundamental architectural limitations of current computing paradigms. While specialized hardware like neuromorphic chips and in-memory architectures offer substantial energy efficiency gains, their integration into mainstream deep learning workflows remains nascent [9][10]. The development of programming models, compilers, and software frameworks that can seamlessly translate existing deep learning models to these novel hardware platforms is a complex undertaking. The learning curve for developers to adapt to these new paradigms, coupled with the need for robust emulation and simulation tools, further slows adoption [12]. Moreover, the intrinsic analog nature of some in-memory architectures introduces concerns about fundamental accuracy limits due to noise, requiring innovative solutions to maintain computational integrity [9].

The sheer scale of modern deep learning models, especially large language models (LLMs), presents a formidable scaling challenge. While quantization techniques have proven effective in compressing LLMs, enabling their deployment on single GPUs, the continuous increase in model parameter counts means that even compressed versions can remain substantial [16]. The development of Dynamic Memory Compression for LLMs, which offers up to a 7-fold throughput increase, addresses some of these memory and inference efficiency issues, but the computational demands of training these colossal models continue to be immense [20]. The reliance on massive datasets, which themselves require significant energy for storage and transfer, further compounds the problem.

From a system perspective, optimizing resource allocation across distributed environments like hybrid clouds or satellite-ground networks is critical but complex [27][24]. Dynamic scheduling algorithms must intelligently balance computational load with the carbon intensity of energy sources, accounting for variations in real-time grid conditions. This requires sophisticated monitoring, prediction, and control mechanisms that can operate at scale without introducing excessive overheads. The challenge is particularly pronounced in heterogeneous environments where different types of accelerators and processors must be coordinated efficiently [25].

Finally, a significant barrier is the lack of standardized metrics and tools for measuring and reporting the energy consumption and carbon footprint of AI systems. Without consistent and fine-grained measurement capabilities, it becomes difficult to objectively compare the efficiency of different approaches, track progress, and incentivize sustainable practices [1]. This measurement gap impedes transparent reporting and the establishment of industry-wide benchmarks for Green AI. Overcoming these scaling challenges requires sustained research and development, interdisciplinary collaboration, and a concerted effort to standardize tools and practices across the AI ecosystem.

4.3. Policy, Social Responsibility, and Future Directions

The imperative for energy-efficient AI extends beyond technical solutions to encompass broader policy frameworks, social responsibility, and strategic future directions. The environmental impact of AI is not merely a technical problem; it necessitates a collective response from stakeholders across the spectrum.

At the policy level, governments and international bodies are increasingly recognizing the need for AI governance, which includes addressing its environmental implications [29]. This involves developing regulations that incentivize sustainable AI development, such as tax breaks for companies using renewable energy for AI operations or mandating energy efficiency disclosures for large AI models. Policy mechanisms could also support research into green computing technologies and infrastructure. For instance, facilitating the integration of AI workloads into hybrid cloud systems with full-stack co-design could foster significant energy efficiency improvements, especially with the maturation of quantum computing for specialized applications [27].

Social responsibility compels the AI community—researchers, developers, and corporations—to prioritize sustainability. This includes transparently reporting the carbon footprint of AI models and actively seeking to minimize it. Cloud providers have already demonstrated that training machine learning in the cloud can significantly reduce CO2e emissions, underscoring the importance of infrastructure choices and green energy commitments [8]. Academia plays a role by integrating green AI principles into curricula, educating the next generation of AI professionals on sustainable practices from the outset [28]. Industry cooperation is essential for establishing best practices, sharing knowledge, and investing in open-source tools for energy measurement and optimization, such as fine-grained energy consumption meters [28][30][1].

Future directions for research and development are diverse. Continued exploration of novel hardware architectures, particularly in-memory computing and neuromorphic systems, remains a high priority, focusing on bridging the gap between theoretical efficiency and practical deployability [9][10]. Further advancements in algorithmic compression techniques, including more sophisticated quantization, pruning, and dynamic memory management, will be crucial as models continue to scale [16][20]. The development of model-based deep learning, which leverages domain knowledge to reduce data and computational requirements, offers another promising avenue for efficiency [21][31]. Moreover, research into energy-aware scheduling algorithms for increasingly complex and distributed AI workloads, including edge and satellite computing environments, will be vital for optimizing energy consumption across the entire computational landscape [6] . Ultimately, fostering a culture of sustainability within the AI community, supported by robust policy and ongoing innovation, will be instrumental in ensuring that AI's transformative power is realized responsibly.

5. Conclusion

The rapid advancement and pervasive integration of artificial intelligence, particularly deep learning, have undeniably reshaped technological landscapes and human capabilities. However, this progress is accompanied by a significant and growing environmental footprint, driven by the substantial energy demands of training and deploying increasingly complex models. The carbon emissions associated with deep learning represent a pressing concern, necessitating a concerted global effort towards more sustainable practices.

This analysis systematically examined green computing approaches across various levels of the AI ecosystem. At the hardware frontier, innovations such as deep in-memory architectures and neuromorphic computing promise radical reductions in energy consumption by fundamentally rethinking computational paradigms. These emerging technologies offer impressive gains in energy efficiency, often with minimal compromise to accuracy, by integrating processing and memory or mimicking the brain's efficient operational principles. Analog photonics CNN accelerators and stochastic computing also exemplify the potential of specialized hardware to achieve superior energy performance.

Algorithmic and software-level optimizations provide immediate and impactful avenues for sustainability. Model compression techniques, including advanced quantization methods like QuantEase and efficient pruning strategies such as Early Pruning with Self-Distillation, significantly reduce the computational and memory requirements of deep learning models. Dynamic memory compression for large language models further enhances inference efficiency, allowing for greater throughput within existing memory constraints. Moreover, the integration of model-based deep learning approaches can lead to more resource-efficient models by leveraging domain-specific knowledge.

System-level and lifecycle management strategies are crucial for holistic sustainability. Energy-aware GPU cluster schedulers, optimized task offloading in distributed environments like satellite networks, and intelligent resource allocation across hybrid clouds demonstrate the potential to minimize energy consumption by dynamically adapting to computational loads and carbon intensities of energy sources. The establishment of standardized metrics and fine-grained energy measurement tools remains a critical enabler for informed decision-making and accountability within the AI community.

Navigating the inherent trade-offs between performance, accuracy, and energy efficiency is central to developing practical green AI solutions. The optimal balance is context-dependent, requiring careful consideration of application-specific requirements and continuous empirical validation. The challenges in scaling sustainable deep learning are considerable, encompassing technological hurdles related to hardware integration, the sheer scale of modern models, and the complexity of managing distributed, energy-aware systems. Addressing these challenges demands sustained research, interdisciplinary collaboration, and the development of robust, accessible tools.

Ultimately, fostering sustainable deep learning necessitates a multi-stakeholder approach involving policy formulation, corporate social responsibility, and educational initiatives. Promoting transparency in energy reporting, incentivizing green AI development, and integrating sustainability into academic curricula are vital steps. The future of AI hinges not only on its computational power but also on its environmental stewardship, ensuring that the transformative potential of deep learning is realized in a manner that is both innovative and ecologically responsible.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] S. Rajput, T. Widmayer, Z. Shang, M. Kechagia, F. Sarro, and T. Sharma, "Enhancing Energy-Awareness in Deep Learning through Fine-Grained Energy Measurement," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 8. Association for Computing Machinery (ACM), pp. 1–34, Nov. 30, 2024. doi: 10.1145/3680470.

[2] H. Raiyan, Md. F. I. Shaif, R. Ahmed, N. H. Nafi, M. R. Sumon, and M. Rahman, "Assessing the impact of influencer marketing on brand value and business revenue: An empirical and thematic analysis," *International Journal of Science and Research Archive*, vol. 16, no. 02, pp. 471–482, 2025, doi: 10.30574/ijsra.2025.16.2.2355.

[3] H. Raiyan, Md. F. I. Shaif, R. Ahmed, N. H. Nafi, M. R. Sumon, and M. Rahman, "The influence of social media branding on consumer purchase behavior: A comprehensive empirical and thematic analysis," *International Journal of Science and Research Archive*, vol. 16, no. 02, pp. 460–470, 2025, doi: 10.30574/ijsra.2025.16.2.2354.doi: 10.48550/arXiv.2304.00897.

[4] C. Yao *et al.*, "Evaluating and analyzing the energy efficiency of CNN inference on high-performance GPU," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 6. Wiley, Oct. 21, 2020. doi: 10.1002/cpe.6064.

[5] D. Gu, Y. Zhao, P. Sun, X. Jin, and X. Liu, "GreenFlow: A Carbon-Efficient Scheduler for Deep Learning Workloads," *IEEE Transactions on Parallel and Distributed Systems*, vol. 36, no. 2. Institute of Electrical and Electronics Engineers (IEEE), pp. 168–184, Feb. 2025.

[6] doi: 10.1109/tpds.2024.3470074.

[7] H. Raiyan, J. Jafia Tasnim, and C. Satu, "Exploring the link between suicidal ideation and digital environments: The hidden impact of marketing content," *International Journal of Science and Research Archive*, vol. 16, no. 02, pp. 607–614, Aug. 2025, doi: 10.30574/ijsra.2025.16.2.2353.

[8] D. Patterson *et al.*, "Energy and Emissions of Machine Learning on Smartphones vs. the Cloud," *Communications of the ACM*, vol. 67, no. 2. Association for Computing Machinery (ACM), pp. 86–97, Jan. 25, 2024. doi: 10.1145/3624719.

[9] M. Kang, Y. Kim, A. D. Patil, and N. R. Shanbhag, "Deep In-Memory Architectures for Machine Learning–Accuracy Versus Efficiency Trade-Offs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 5. Institute of Electrical and Electronics Engineers (IEEE), pp. 1627–1639, May 2020. doi: 10.1109/tcsi.2019.2960841.

[10] "Neuromorphic Computing: Advancing Energy-Efficient AI Systems through Brain-Inspired Architectures," *Nanotechnology Perceptions*, vol. 20, no. S14. Rotherham Press, Nov. 04, 2024. doi: 10.62441/nano-ntp.v20is14.99.

[11] Raiyan Haider, Wahida Ahmed Megha, Jafia Tasnim Juba, Aroa Alamgir, and Labib Ahmad, "The conversational revolution in health promotion: Investigating chatbot impact on healthcare marketing, patient engagement, and service reach," *International Journal of Science and Research Archive*, vol. 15, no. 3. GSC Online Press, pp. 1585–1592, Jun. 30, 2025. doi: 10.30574/ijsra.2025.15.3.1937.

[12] T. Luo *et al.*, "Achieving Green AI with Energy-Efficient Deep Learning Using Neuromorphic Computing," *Communications of the ACM*, vol. 66, no. 7. Association for Computing Machinery (ACM), pp. 52–57, Jun. 22, 2023. doi: 10.1145/3588591.

[13] D. Dang, B. Lin, and D. Sahoo, "*LiteCON* : An All-photonic Neuromorphic Accelerator for Energy-efficient Deep Learning," *ACM Transactions on Architecture and Code Optimization*, vol. 19, no. 3. Association for Computing Machinery (ACM), pp. 1–22, Sep. 2022. doi: 10.1145/3531226.

[14] J. Wang, H. Chen, D. Wang, K. Mei, S. Zhang, and X. Fan, "A Noise-Driven Heterogeneous Stochastic Computing Multiplier for Heuristic Precision Improvement in Energy-Efficient DNNs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 2. Institute of Electrical and Electronics Engineers (IEEE), pp. 630–643, Feb. 2023. doi: 10.1109/tcad.2022.3178053.

[15] [15] Raiyan Haider, Farhan Abrar Ibne Bari, Osru, Nishat Afia, and Mohammad Abiduzzaman khan Mugdho, “Leveraging internet of things data for real-time marketing: Opportunities, challenges, and strategic implications,” *International Journal of Science and Research Archive*, vol. 15, no. 3. GSC Online Press, pp. 1657–1663, Jun. 30, 2025. doi: 10.30574/ijsra.2025.15.3.1936.

[16] doi: 10.48550/arXiv.2309.01885.

[17] doi: 10.48550/arXiv.2402.00084.

[18] doi: 10.48550/arXiv.2210.07795.

[19] Raiyan Haider, Md Farhan Abrar Ibne Bari, Md. Farhan Israk Shaif, Mushfiqur Rahman, Md. Nahid Hossain Ohi, and Kazi Md Mashrur Rahman, “Quantifying the Impact: Leveraging AI-Powered Sentiment Analysis for Strategic Digital Marketing and Enhanced Brand Reputation Management,” *International Journal of Science and Research Archive*, vol. 15, no. 2. GSC Online Press, pp. 1103–1121, May 30, 2025. doi: 10.30574/ijsra.2025.15.2.1524.

[20] doi: 10.48550/arXiv.2403.09636.

[21] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, “Model-Based Deep Learning,” *Proceedings of the IEEE*, vol. 111, no. 5. Institute of Electrical and Electronics Engineers (IEEE), pp. 465–499, May 2023. doi: 10.1109/jproc.2023.3247480.

[22] Raiyan Haider, Md Farhan Abrar Ibne Bari, Md. Farhan Israk Shaif, and Mushfiqur Rahman, “Engineering hyper-personalization: Software challenges and brand performance in AI-driven digital marketing management: An empirical study,” *International Journal of Science and Research Archive*, vol. 15, no. 2. GSC Online Press, pp. 1122–1141, May 30, 2025. doi: 10.30574/ijsra.2025.15.2.1525.

[23] B. R. C. -, “Edge Computing vs. Cloud Computing: A Comparative Analysis for Real-Time AI Applications,” *International Journal For Multidisciplinary Research*, vol. 6, no. 5. International Journal for Multidisciplinary Research (IJFMR), Oct. 28, 2024. doi: 10.36948/ijfmr.2024.v06i05.29316.

[24] J. Zhou, J. Liang, L. Zhao, S. Wan, H. Cai, and F. Xiao, “Latency-Energy Efficient Task Offloading in the Satellite Network-Assisted Edge Computing via Deep Reinforcement Learning,” *IEEE Transactions on Mobile Computing*, vol. 24, no. 4. Institute of Electrical and Electronics Engineers (IEEE), pp. 2644–2659, Apr. 2025. doi: 10.1109/tmc.2024.3502643.

[25] P. Agrawal and S. Rao, “Energy-Aware Scheduling of Distributed Systems,” *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 4. Institute of Electrical and Electronics Engineers (IEEE), pp. 1163–1175, Oct. 2014. doi: 10.1109/tase.2014.2308955.

[26] Raiyan Haider, Md Farhan Abrar Ibne Bari, Osru, Nishat Afia, and Tanjim Karim, “Illuminating the black box: Explainable AI for enhanced customer behavior prediction and trust,” *International Journal of Science and Research Archive*, vol. 15, no. 3. GSC Online Press, pp. 247–268, Jun. 30, 2025. doi: 10.30574/ijsra.2025.15.3.1674.

[27] doi: 10.48550/arXiv.2411.13239.

[28] M. B. Satterfield *et al.*, “Overcoming Nontechnical Barriers to the Implementation of Sustainable Solutions in Industry,” *Environmental Science & Technology*, vol. 43, no. 12. American Chemical Society (ACS), pp. 4221–4226, May 21, 2009. doi: 10.1021/es802980j.

[29] A. Batool, D. Zowghi, and M. Bano, “AI governance: a systematic literature review,” *AI and Ethics*, vol. 5, no. 3. Springer Science and Business Media LLC, pp. 3265–3279, Jan. 14, 2025. doi: 10.1007/s43681-024-00653-w.

[30] Raiyan Haider and Jasmima Sabatina, “Harnessing the power of micro-influencers: A comprehensive analysis of their effectiveness in promoting climate adaptation solutions,” *International Journal of Science and Research Archive*, vol. 15, no. 2. GSC Online Press, pp. 595–610, May 30, 2025. doi: 10.30574/ijsra.2025.15.2.1448.

[31] Raiyan Haider, “Navigating the digital political landscape: How social media marketing shapes voter perceptions and political brand equity in the 21st Century,” *International Journal of Science and Research Archive*, vol. 15, no. 1. GSC Online Press, pp. 1736–1744, Apr. 30, 2025. doi: 10.30574/ijsra.2025.15.1.1217.