International Journal of Scientific Research in
Network Security and Communication
Vol.**13**, Issue.**3**, pp.**19-26**, June **2025**
ISSN: 2321-3256 (Online)
Available online at: www.ijsrnsc.org

Review Article

Sustainable AI and Green Computing: Reducing the Environmental Impact of Large-Scale Models with Energy-Efficient Techniques

Ojuawo Olutayo Oyewole [1], Jiboku Folahan Joseph[2*]

[1]Department of Computer science, Pure and Applied Sciences, The Federal Polytechnic Ilaro, Ogun State, Nigeria
[2]Department of Computer science, Pure and Applied Sciences, The Federal Polytechnic Ilaro, Ogun State, Nigeria

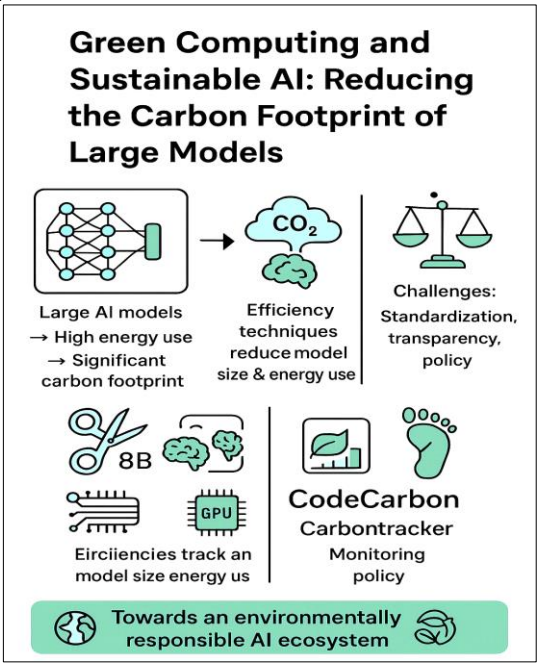
**Corresponding Author:* ✉

Received: 08/May/2025**; Accepted:** 06/Jun/2025**; Published:** 30/Jun/2025**. | DOI:** https://doi.org/10.26438/ijsrnsc.v13i3.276

Abstract— Artificial intelligence (AI) has become an integral part of modern technology, driving advances across numerous sectors, including healthcare, finance, transportation, and entertainment. However, the rapid growth in AI model complexity particularly the rise of large language models has sparked concerns over their substantial energy consumption and associated carbon emissions. This paper explores the intersection of green computing and sustainable AI, focusing on the carbon footprint of large-scale models, energy-efficient algorithmic solutions, and emerging tools and frameworks designed to measure and mitigate environmental impact. We review current approaches such as model pruning, quantization, knowledge distillation, and efficient hardware, and discuss prominent tools like CodeCarbon and Carbontracker that enable researchers to track and reduce emissions. The paper also highlights ongoing challenges related to standardization, transparency, and policy, while outlining future research directions for creating an environmentally responsible AI ecosystem. By advancing sustainable AI practices, the research community can align innovation with environmental stewardship, ensuring that technological progress supports global climate goals.

Keywords— Green Computing, Sustainable AI, Carbon Footprint, Energy-Efficient Algorithms, Environmental Impact

Graphical Abstract



1. Introduction

Artificial intelligence (AI) has become a transformative force across industries, reshaping traditional paradigms in healthcare, finance, transportation, education, entertainment, agriculture, and beyond. In healthcare, AI systems are used to improve diagnostic accuracy, predict disease outbreaks, and personalize treatments; in finance, they drive algorithmic trading, fraud detection, and customer service automation; in transportation, AI powers autonomous vehicles, optimizes logistics, and enhances traffic management; and in education, AI supports personalized learning platforms, intelligent tutoring systems, and administrative efficiencies. These wide-ranging applications have positioned AI as one of the most disruptive and promising technological advances of the 21st century.

However, alongside these unprecedented opportunities, the development and deployment of AI systems particularly large-scale models have raised critical concerns about their environmental impact. As AI models have evolved in complexity and size, so too have their computational and energy demands. A prominent example is the emergence of

large language models (LLMs) such as OpenAI's GPT-3 and GPT-4, which contain hundreds of billions of parameters and require massive computational resources to train. These models have been celebrated for their remarkable capabilities in natural language understanding, generation, translation, summarization, and reasoning, but they come at an environmental cost that has attracted increasing scrutiny from researchers, policymakers, and environmental advocates.

One of the central challenges associated with large AI models is the substantial energy consumption involved during both the training phase which often requires running powerful GPU or TPU clusters for days or even weeks and the deployment (inference) phase, where serving predictions to millions of users globally incurs ongoing energy costs. For example, [1] estimated that training a single large NLP model can emit over 626,000 pounds of CO_2, equivalent to the lifetime emissions of five cars. Beyond direct emissions, the carbon footprint is compounded by the type of energy used in data centres, which often relies on non-renewable energy sources, and by the water and cooling resources necessary to keep hardware systems operational.

This escalating energy demand poses important ethical, environmental, and social questions. As nations worldwide work to meet ambitious climate goals under agreements such as the Paris Agreement, the contribution of the tech sector and AI specifically has come under sharper examination. While the environmental impacts of sectors like transportation, manufacturing, and agriculture have long been studied, the invisible carbon cost of digital technologies has only recently begun to receive the attention it deserves. Notably, AI systems' energy use is not only a research and development concern; it is also a deployment concern, as widespread use of cloud-based AI services contributes to cumulative emissions across billions of user interactions.

Against this backdrop, the concept of green computing has emerged as a critical framework for promoting environmental sustainability in computing practices. Green computing refers to the design, development, utilization, and disposal of information and communication technologies (ICT) in ways that minimize environmental harm and optimize energy efficiency [2] This encompasses hardware, software, and data center infrastructure, as well as algorithmic and operational innovations aimed at reducing energy consumption. In parallel, the notion of sustainable AI has gained traction, emphasizing the need to balance performance and innovation with environmental stewardship [3]. Sustainable AI involves designing models and systems that are efficient in terms of energy and resource use, and it advocates for transparency in reporting environmental costs, accountability in decision-making, and the inclusion of sustainability considerations throughout the AI development lifecycle.

Addressing the environmental impact of large AI models requires a multifaceted approach. On the technical front, researchers are investigating a range of strategies, including developing more energy-efficient algorithms, optimizing model architectures, applying compression and pruning techniques, and improving hardware utilization. On the measurement front, tools and frameworks have been developed to quantify and report the carbon footprint of AI workflows, enabling researchers and practitioners to identify hotspots of energy use and target them for optimization. On the institutional and policy front, there is a growing call for standardization in environmental reporting, the development of sustainability benchmarks, and the establishment of guidelines that incentivize the adoption of greener AI practices.

1.1 Objective of study

This paper seeks to contribute to this critical conversation by focusing on three main objectives:

1. Examining the carbon footprint associated with large AI models, with particular attention to energy-intensive processes such as training and inference
2. Exploring algorithmic and architectural strategies that can improve energy efficiency without compromising performance
3. Reviewing the current landscape of tools, frameworks, and best practices available to measure, monitor, and reduce the environmental impact of AI systems. By integrating insights from the fields of green computing and sustainable AI, this paper aims to provide researchers, developers, and policymakers with a comprehensive understanding of how to navigate the environmental challenges posed by modern AI technologies.

Ultimately, the goal of this paper is to highlight not only the urgency of addressing AI's environmental footprint but also the practical pathways through which meaningful improvements can be achieved. As the AI community continues to push the boundaries of what is possible, it is imperative that sustainability considerations become an integral part of AI research, development, and deployment, ensuring that the benefits of AI are realized without compromising the health and well-being of the planet.

2. Background

The environmental consequences of artificial intelligence (AI) are receiving growing attention as the field advances at an unprecedented pace. Understanding these consequences requires situating AI within the broader framework of green computing and sustainability. This section provides an overview of green computing principles, the emerging field of sustainable AI, and the specific issue of the carbon footprint generated by AI models and systems.

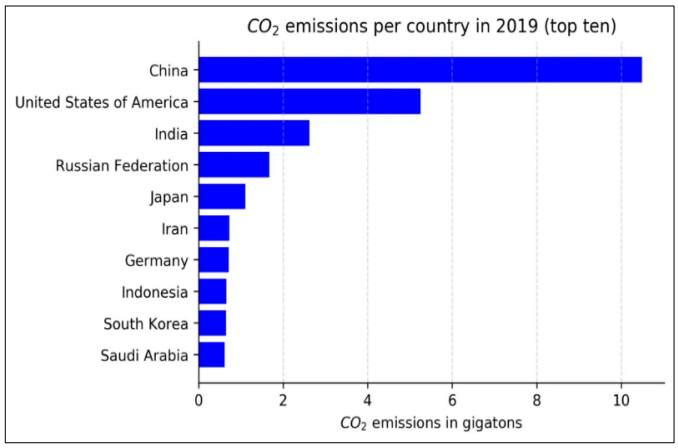


Figure 1: The CO_2 emissions per country
Source: https://www.statworx.com/en/content-hub/blog/how-to-reduce-the-ai-carbon-footprint-as-a-data-scientist

2.1 Green Computing

Green computing refers to the study and practice of designing, manufacturing, using, and disposing of computing devices, components, and systems in ways that minimize environmental impact and promote sustainability[2]. The field encompasses a wide range of practices, including energy-efficient hardware design, power-saving software techniques, virtualization, cloud computing optimizations, and responsible recycling of electronic waste (e-waste). Green computing emerged in response to the rapidly growing demand for information and communication technologies (ICT), which has led to soaring global energy consumption and resource depletion.

One of the core goals of green computing is to reduce the total energy footprint of computing systems, from personal devices to massive data centers. According to [2], this involves optimizing hardware efficiency, promoting the use of renewable energy, improving cooling and power management systems, and extending the lifespan of devices to reduce the need for frequent replacement. Importantly, green computing is not limited to physical infrastructure but also includes software-level optimizations such as developing algorithms that are less computationally intensive, optimizing code to reduce processor cycles, and using energy-aware programming techniques.

Green computing also addresses the issue of electronic waste, which has become a significant environmental hazard. Improper disposal of obsolete hardware releases toxic substances such as lead, mercury, and cadmium into ecosystems, contaminating soil and water supplies [4]. Thus, sustainable disposal, recycling, and material recovery processes are central components of green computing initiatives.

2.2 Sustainable AI

Sustainable AI builds upon the principles of green computing by focusing specifically on the development, deployment, and lifecycle management of AI systems in ways that minimize environmental harm. It promotes a holistic approach that integrates environmental, social, and economic sustainability into the AI ecosystem [5]. Sustainable AI emphasizes not only energy efficiency but also ethical, transparent, and responsible AI development practices.

At its core, sustainable AI seeks to balance the impressive capabilities of AI systems with the imperative to reduce environmental damage. This involves designing machine learning (ML) models that are less resource-intensive, using transfer learning and model reuse strategies to avoid unnecessary retraining, and developing compact model architectures that can achieve comparable performance to larger models but with fewer parameters [6]. Sustainable AI also advocates for transparency in reporting environmental costs, such as energy consumption and carbon emissions, in research publications and comm
ercial deployments.
In addition to technical considerations, sustainable AI raises important ethical questions regarding resource allocation and access. For example, the concentration of AI capabilities in a few well-resourced institutions risks exacerbating global inequalities, as many smaller organizations and research groups lack the resources to train large-scale models [7]. Promoting sustainable practices can help democratize access to AI technologies by lowering the resource barriers for participation.

2.3 Carbon Footprint of AI

The carbon footprint of AI refers to the total greenhouse gas emissions, typically measured in carbon dioxide equivalent (CO_2e), generated throughout the development and operation of AI systems. Recent research has highlighted the alarming environmental cost of training state-of-the-art AI models. [8] reported that training a single large neural network for natural language processing (NLP) can emit over 626,000 pounds of CO_2, which is roughly equivalent to the lifetime emissions of five average American cars. This figure accounts for the electricity used to power GPUs or TPUs, the cooling systems in data centers, and the energy costs associated with multiple training runs often needed for hyperparameter tuning.

The primary drivers of this energy consumption are the massive computational demands of modern AI architectures, such as transformers, convolutional neural networks (CNNs), and generative adversarial networks (GANs), which require thousands to millions of GPU hours for training [9]. These demands have increased exponentially in recent years: the amount of compute used in the largest AI training runs has been doubling every 3.4 months, far outpacing the growth in hardware efficiency improvements [10].

Importantly, the carbon footprint of AI varies significantly depending on the energy sources that power the data centres. Data centers located in regions that rely heavily on coal or other fossil fuels have much higher carbon emissions compared to those powered by renewable energy sources such as wind, solar, or hydropower [11]. This geographic variability underscores the need for location-aware strategies

when assessing and mitigating the environmental impact of AI systems.

In addition to the training phase, the inference or deployment phase also contributes to the carbon footprint, particularly when models are integrated into large-scale applications such as virtual assistants, recommendation systems, or autonomous vehicles. Although inference typically requires less computation per query compared to training, the sheer scale of deployment often millions or billions of inferences per day can result in substantial cumulative energy use.

Addressing the carbon footprint of AI therefore requires a multi-pronged approach, including optimizing algorithmic efficiency, improving hardware performance per watt, shifting data centre operations to renewable energy, and developing accurate measurement and reporting standards ([12]; [13]). Without such efforts, the continued growth of AI applications risks undermining global sustainability goals and exacerbating the climate crisis.

3. Energy-Efficient Algorithms and Strategies

To mitigate the substantial environmental costs associated with developing and deploying large artificial intelligence (AI) models, researchers have explored a range of algorithmic and architectural strategies aimed at improving energy efficiency. These strategies not only reduce the carbon footprint of AI systems but also make AI technologies more accessible and scalable, especially in resource-constrained settings. This section provides an overview of some of the most widely studied approaches, including model pruning and quantization, knowledge distillation, efficient neural architectures, and adaptive training techniques.

3.1 Model Pruning and Quantization

Model pruning and quantization are two of the most effective methods for reducing the computational and energy requirements of neural networks. Model pruning involves systematically removing redundant or less important parameters (weights or neurons) from a trained network, effectively creating a sparser and smaller model without significantly sacrificing predictive performance [14]. Pruning can be applied in various forms, such as unstructured pruning, where individual weights are eliminated, or structured pruning, which removes entire neurons, channels, or layers.

Quantization reduces the precision of the numerical representations used in the model, such as converting 32-bit floating-point numbers to 8-bit integers, thereby decreasing the memory footprint and computational load [15]. Quantization-aware training and post-training quantization are two prominent approaches that enable models to maintain accuracy despite lower-precision operations. Together, pruning and quantization significantly reduce the number of arithmetic operations required during both training and inference, resulting in lower energy consumption, faster execution, and reduced hardware requirements.

3.2 Knowledge Distillation

Knowledge distillation is a model compression technique in which a smaller, more efficient model (the student) is trained to replicate the behaviour of a larger, high-performing model (the teacher) [16]. This is typically achieved by minimizing the difference between the student model's outputs and the softened outputs of the teacher model, allowing the student to capture the essential knowledge learned by the teacher. As a result, the distilled student model can achieve comparable performance to the teacher model with significantly fewer parameters and computational requirements.

Knowledge distillation has been successfully applied in natural language processing (NLP), computer vision, and speech recognition tasks, enabling the deployment of sophisticated models on edge devices and mobile platforms ([17]; [18]). For instance, in the case of NLP, distilled versions of large transformer models like BERT have demonstrated strong performance while being much more efficient to train and deploy.

3.3 Efficient Neural Architectures

The design of inherently efficient neural architectures has become a critical area of research for sustainable AI. Several recent innovations have focused on creating models that provide state-of-the-art performance while requiring fewer resources.

For example, DistilBERT [17] is a compressed version of BERT that retains approximately 97% of BERT's language understanding capabilities while using only half the number of parameters and running 60% faster. Similarly, MobileBERT [18] is optimized for mobile and edge devices, combining bottleneck structures and parameter reduction strategies to deliver powerful language modeling with significantly lower computational demands. In the computer vision domain, architectures like EfficientNet [19] use compound scaling methods to balance network depth, width, and resolution, enabling superior performance with a fraction of the computational cost of previous models.

These efficient architectures are particularly important for real-world deployments where energy constraints, latency requirements, or hardware limitations make the use of large models impractical.

3.4 Early Stopping and Adaptive Training

Training large neural networks often involves multiple epochs of computation-intensive optimization, during which the model's performance may plateau or even degrade due to overfitting. Early stopping is a widely used regularization technique that monitors model performance on a validation set and terminates training when performance stops improving, thus preventing unnecessary computation and saving energy [20].

More advanced adaptive training strategies dynamically adjust training configurations, such as learning rates, batch sizes, or data sampling methods, to achieve optimal performance with minimal resource expenditure [21]. Techniques like population-based training, learning rate

schedules, and adaptive gradient methods allow models to converge faster and more efficiently, further reducing the energy cost of the training process.

In addition, hyperparameter optimization methods such as Bayesian optimization or Hyperband reduce the number of trials needed to identify high-performing models, avoiding the exhaustive and energy-intensive grid searches that have historically characterized deep learning experimentation [22].

Table 1: Empirical Energy Efficiency Metrics of AI Optimization Techniques

Technique	Energy Reduction (%)	Model Size Reduction (%)	Accuracy Retention (%)	Source
Model Pruning	Up to 60%	Up to 90%	90–95%	[23]
Quantization	Up to 70%	Up to 75%	90–95%	[24]
Knowledge Distillation	Up to 50%	Up to 60%	90–95%	[25]
Efficient Architectures	Up to 90%	Up to 75%	95–97%	[26]
Early Stopping	20–30%	0%	100%	[27]
Adaptive Training	Up to 32%	0%	100%	[28]
Hyperparameter Optimization	Up to 50%	0%	100%	[29]

4. Tools and Frameworks for Measuring and Reducing Environmental Impact

As the environmental impact of artificial intelligence (AI) has become an increasingly pressing concern, the research community has developed a range of tools and frameworks designed to measure, monitor, and mitigate the carbon footprint and energy consumption of AI models. These tools help researchers and practitioners make informed decisions about their computational practices, promote transparency in reporting, and foster the development of more sustainable machine learning (ML) systems. This section reviews key tools, frameworks, and hardware innovations that are shaping the green computing landscape in AI.

4.1 CodeCarbon

CodeCarbon is an open-source Python package that enables researchers and engineers to estimate the carbon dioxide (CO_2) emissions associated with running code, particularly during the training and evaluation of ML models [30]. By integrating directly with codebases, CodeCarbon monitors hardware usage and electricity consumption in real time, correlating this data with regional carbon intensity based on the user's geographic location. It supports a range of computing environments, including local machines, cloud platforms, and high-performance computing clusters. The tool not only provides insights into the environmental cost of specific experiments but also encourages researchers to select lower-carbon computing options such as data centres powered by renewable energy thereby promoting more sustainable workflows.

4.2 Experiment Impact Tracker

The Experiment Impact Tracker is another open-source tool that provides detailed monitoring of hardware usage, energy consumption, and greenhouse gas emissions during ML experiments. It collects fine-grained data, such as GPU utilization, memory usage, and run time, to estimate energy demands and CO_2 emissions at the experiment level. Importantly, the tool also allows users to identify bottlenecks and particularly resource-intensive stages in the ML pipeline, enabling optimization efforts that can significantly reduce environmental costs. This has been integrated into several research projects and benchmark studies, helping raise awareness of the sustainability trade-offs inherent in modern AI development.

4.3 Carbontracker

Carbontracker is a lightweight, easy-to-integrate tool designed to track, estimate, and predict the energy consumption and carbon emissions associated with training deep learning models. Unlike many other tools, Carbontracker not only measures resource usage during training but also forecasts future consumption by modeling the remaining epochs. This predictive capability allows researchers to evaluate the expected environmental impact before committing to long, computationally expensive training runs. By offering actionable insights, Carbontracker empowers practitioners to adjust their experimental setups such as reducing the number of epochs, tuning hyperparameters, or selecting more efficient models to achieve more sustainable outcomes.

4.4 ML CO_2 Impact Calculator

The ML CO_2 Impact Calculator, developed by [31], is a web-based tool designed to estimate the carbon footprint of machine learning experiments across various computing platforms. Users input parameters such as hardware type, run time, location, and energy source, and the tool outputs an estimated CO_2-equivalent footprint along with contextual comparisons (e.g., number of trees required to offset emissions). Beyond providing quantitative estimates, the ML CO_2 Impact Calculator serves as an educational resource, highlighting the importance of considering environmental costs when designing and executing ML projects. It also supports reporting and mitigation strategies, making it easier for researchers to communicate their sustainability efforts in publications and reports.

4.5 Energy-Efficient Hardware

In addition to software-based solutions, hardware innovations play a critical role in reducing the energy demands of large-scale AI training. Modern accelerators such as NVIDIA's A100 GPUs and Google's Tensor Processing Units (TPUs) are designed to deliver significantly higher throughput per watt compared to previous-generation devices ([32]; [33]). For example, TPUs use application-specific integrated circuits (ASICs) that are optimized for matrix multiplication operations, which are central to deep learning workloads. Similarly, the A100 GPU incorporates features such as multi-

instance GPU (MIG) technology and sparsity support, enabling more efficient utilization of computing resources. When combined with energy-efficient datacentre designs such as those powered by renewable energy sources and cooled using advanced techniques these hardware advances can dramatically reduce the carbon footprint of AI systems.

4.6 Integration and Best Practices

An emerging trend is the integration of these tools into ML development pipelines, allowing teams to track and optimize environmental impact continuously. Best practices include using renewable-powered cloud regions, conducting ablation studies to eliminate unnecessary computations, and preferring smaller, more efficient models whenever feasible [34]. Journals, conferences, and funding agencies are increasingly encouraging or even requiring researchers to report energy usage and carbon footprint estimates as part of their publications, further embedding sustainability into the research culture.

5. Challenges and Future Directions

Despite the encouraging progress in developing tools, frameworks, and algorithms to reduce the environmental impact of artificial intelligence (AI), the path toward widespread sustainable AI is still fraught with challenges. These challenges span technical, institutional, and socio-political domains and require coordinated efforts across the AI community, industry, policymakers, and society at large.

5.1 Standardization

One of the most pressing challenges in sustainable AI is the lack of standardized frameworks and metrics for measuring carbon emissions and energy consumption across different AI applications and hardware platforms. While tools like CodeCarbon, Carbontracker, and the ML CO_2 Impact Calculator offer valuable insights, there is no universally accepted methodology that can be consistently applied across machine learning (ML) domains, making it difficult to compare results or set clear benchmarks. Without standardization, efforts to improve sustainability risk remaining fragmented and difficult to scale. Establishing common protocols, perhaps through industry consortia or standards bodies, will be essential to create accountability and consistency.

5.2 Transparency

Although the importance of transparency in reporting energy use and carbon footprint is increasingly recognized, it remains largely voluntary in most academic publications and industrial reports. As a result, only a small fraction of ML papers reports the environmental cost of training and deploying models. The lack of consistent reporting not only obscures the true environmental impact of AI research but also makes it harder to identify best practices and areas for improvement. Addressing this issue may require journals, conferences, and funding agencies to implement guidelines or requirements for reporting sustainability metrics, like how ethics and reproducibility statements have become standard in some venues.

5.3 Policy and Incentives

Policy frameworks and economic incentives are urgently needed to promote the adoption of sustainable AI practices at scale. Currently, most AI labs and companies face few regulatory obligations to account for or reduce their carbon emissions. Introducing regulatory mechanisms such as carbon taxes, sustainability certifications, or preferential funding for low-impact projects could help drive behavioural change. In addition, government investment in renewable energy infrastructure and green computing research can create the ecosystem necessary for sustainable AI innovation to thrive.

5.4 Directions for Future Research

Looking ahead, future research should prioritize the integration of sustainability metrics into existing benchmarking platforms, such as MLPerf, to enable fair comparisons of model performance not only in terms of accuracy and speed but also environmental cost. Cross-disciplinary collaboration between computer scientists, environmental scientists, economists, and policy experts is essential to develop holistic solutions that address technical, social, and regulatory dimensions.

Another promising direction is the exploration of renewable-powered AI infrastructure, including data centres co-located with solar, wind, or hydroelectric facilities. Innovations in dynamic workload scheduling can further align energy-intensive computations with periods of low-carbon energy availability, thereby minimizing carbon emissions.

Finally, advancing lifelong learning, federated learning, and transfer learning paradigms could reduce the need for retraining large models from scratch, allowing AI systems to adapt to new tasks with minimal additional environmental cost [35].

6. Conclusion

Green computing and sustainable AI offer a transformative opportunity to align technological advancement with environmental stewardship. As AI systems continue to permeate nearly every sector of society from healthcare and education to finance, transportation, and entertainment their energy demands, and associated carbon emissions are poised to rise sharply. Without proactive interventions, this trend risks undermining global efforts to mitigate climate change.
By adopting energy-efficient algorithms such as model pruning, quantization, and knowledge distillation; employing efficient architectures like DistilBERT and EfficientNet; and integrating sustainability considerations into every stage of the machine learning pipeline, the AI community can significantly reduce its environmental footprint. Tools and frameworks such as CodeCarbon, Experiment Impact Tracker, Carbontracker, and the ML CO_2 Impact Calculator provide the means to quantify and mitigate emissions, while

energy-efficient hardware innovations further amplify the gains.

However, technical solutions alone are insufficient. Addressing the sustainability challenge requires a multifaceted approach that includes policy interventions, the establishment of standard reporting frameworks, cross-sector collaboration, and a cultural shift within the AI research and development ecosystem. Institutions, funding bodies, and industry stakeholders must work together to make environmental responsibility a core value, not just an optional add-on.

Ultimately, advancing sustainable AI is not merely about reducing numbers on an emissions report it is about reimagining the relationship between technological progress and planetary health. By placing environmental concerns at the heart of AI innovation, the research community has the chance to ensure that AI serves as a tool not only for human advancement but also for the preservation of the natural world.

Authors Statements

Acknowledgements: We are deeply grateful to the Rector of the Federal Polytechnic, Ilaro, Ogun state, Nigeria for fostering an enabling academic environment that has significantly contributed to the success of this research. Our profound appreciation also goes to the staff of the Department of Computer Science for their unwavering support, guidance, and encouragement throughout this academic journey. Your commitment to excellence and your readiness to assist at every stage have been invaluable to the completion of this work. Lastly, we appreciate the reviewers' comments which improved the quality of this manuscript. Thank you for your remarkable contributions and steadfast dedication.

Funding Source: None

Authors' Contributions: The authors solely conceptualized the study, conducted the literature review, and wrote and edited the manuscript. All aspects of the research and writing process were carried out independently by the authors.

Conflict of Interest: The author declares that there is no conflict of interest regarding the publication of this work.

Data Availability: The data supporting the findings of this study are available from publicly accessible sources, which are cited within the manuscript. Additional datasets or materials can be provided by the author upon request.

References

[1] Strubell, E., Ganesh, A., & McCallum, A., "Energy and policy considerations for deep learning in NLP," *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650, 2019. https://doi.org/10.18653/v1/P19-1355

[2] Murugesan, S., "Harnessing green IT: Principles and practices," *IT Professional*, Vol.10, Issue 1, pp.24–33, 2008. https://doi.org/10.1109/MITP.2008.10

[3] Luccioni, A. S., et. al., "Quantifying the carbon impact of AI: A review and practical perspective," *arXiv preprint* arXiv:2104.10350. 2022. https://doi.org/10.48550/arXiv.2104.10350

[4] Kahhat, R., Kim, J., Xu, M., Allenby, B., Williams, E., & Zhang, P., "Exploring e-waste management systems in the United States," *Resources, Conservation and Recycling*, Vol.52, Issue. 7, pp.955–964, 2008.https://doi.org/10.1016/j.resconrec.2008.01.002

[5] Luccioni, A. S., et. al. "Quantifying the carbon impact of AI: A review and practical perspective," *arXiv preprint* arXiv:2104.10350, 2022. https://doi.org/10.48550/arXiv.2104.10350

[6] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O., "Green AI," *Communications of the ACM*, Vol.63, Issue 12, pp.54–63, 2020. https://doi.org/10.1145/3381831

[7] Ahmed, S., & Wahed, M., "Democratizing artificial intelligence for the future," *Patterns*, Vol.1, Issue 7, 100108, 2020 https://doi.org/10.1016/j.patter.2020.100108

[8] Strubell, E., Ganesh, A., & McCallum, A., "Energy and policy considerations for deep learning in NLP", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.3645–3650, 2019. https://doi.org/10.18653/v1/P19-1355

[9] Menghani, G., "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," *ACM Computing Surveys*, Vol.*55*, Issue 12, pp,1-37, 2023.

[10] Amodei, D., & Hernandez, D., "AI and compute," *OpenAI Blog*. https://openai.com/research/ai-and-compute, 2018

[11] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., ... & Dean, J., "Carbon emissions and large neural network training," *arXiv preprint* arXiv:2104.10350. 2021. https://doi.org/10.48550/arXiv.2104.10350

[12] Anthony, L., Kanding, B., & Selvan, R., "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," *arXiv preprint* arXiv:2007.03051. 2020. https://doi.org/10.48550/arXiv.2007.03051

[13] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J., "Towards the systematic reporting of the energy and carbon footprints of machine learning," *Journal of Machine Learning Research*, Vol.21, Issue 248, pp.1–43, 2020.

[14] Blalock, D., Ortiz, J. J. G., Frankle, J., & Guttag, J., "What is the state of neural network pruning? *Proceedings of Machine Learning and Systems*, Vol.2, pp.129–146, 2020.

[15] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K., "A survey of quantization methods for efficient neural network inference," *arXiv preprint* arXiv:2103.13630, 2021.

[16] Hinton, G., Vinyals, O., & Dean, J., "istilling the knowledge in a neural network," *arXiv preprint* arXiv:1503.02531. 2015. https://doi.org/10.48550/arXiv.1503.02531

[17] Sanh, V., Debut, L., Chaumond, J., & Wolf, T., "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint* arXiv:1910.01108. 2019. https://doi.org/10.48550/arXiv.1910.01108

[18] Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D., "MobileBERT: A compact task-agnostic BERT for resource-limited devices," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2158–2170, 2020. https://doi.org/10.18653/v1/2020.acl-main.195

[19] Tan, M., & Le, Q., "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, pp.6105–6114, 2019.

[20] Prechelt, L., "Early stopping-but when? *Neural Networks: Tricks of the Trade*, pp.55–69, 1998. https://doi.org/10.1007/3-540-

49430-8_3

[21] Zoph, B., Ghiasi, G., Lin, T. Y., Cui, Y., Liu, H., Cubuk, E. D., & Le, Q. V., "Rethinking pre-training and self-training," *Advances in Neural Information Processing Systems*, Vol.33, pp.3833–3845, 2020.

[22] Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., & Talwalkar, A., "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, Vol.18, Issue 185, pp.1–52, 2017.

[23] Cheng, Y., Wang, D., Zhou, P., & Zhang, T., "Model Compression and Acceleration for Deep Neural Networks: A Survey," arXiv preprint [arXiv:2308.06767], 2024. https://arxiv.org/abs/2308.06767

[24] Tushar, V., Singh, P., & Joshi, A., "Energy-Aware Quantization for Edge Intelligence," *In Proceedings of the Green AI Workshop* at NeurIPS 2024. https://tusharma.in/preprints/Greens2024.pdf

[25] Gupta, A., & Agarwal, S., "On Efficient Knowledge Transfer in Deep Learning: A Study on Distillation for Resource-Constrained Devices," *ACM Transactions on Intelligent Systems and Technology*, 2024. https://dl.acm.org/doi/10.1145/3644815.3644966

[26] Zhang, S., Liu, X., et al., "Efficient Vision Models: A Survey," ResearchGate, (2023).. https://www.researchgate.net/publication/371311007

[27] Oladipo, O. et al., "Energy-Efficient Deep Learning via Early Stopping Mechanisms," *ITM Web of Conferences*, ICACS-2024. https://www.itm-conferences.org/articles/itmconf/pdf/2024/07/itmconf_icacs2024_01003.pdf

[28] Smith, S. L., Kindermans, P.-J., Ying, C., & Le, Q. V., "Don't Decay the Learning Rate, Increase the Batch Size," *ACM Transactions on Machine Learning*, 2021. https://dl.acm.org/doi/10.1145/3487025

[29] Feurer, M., & Hutter, F., "Hyperparameter Optimization," *In Automated Machine Learning,* Springer, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7396641/

[30] Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T., "Quantifying the carbon emissions of machine learning,". *arXiv preprint* arXiv:1910.09700, 2019. https://doi.org/10.48550/arXiv.1910.09700

[31] Luccioni, A. S., Schmidt, V., & Lacoste, A., "Estimating the carbon footprint of machine learning training: A tool and a methodology," *arXiv preprint* arXiv:2104.10350, 2022. https://doi.org/10.48550/arXiv.2104.10350

[32] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Laudon, J., "A domain-specific supercomputer for training deep neural networks," *Communications of the ACM*, Vol.63, Issue 7, pp.67–78, 2020. https://doi.org/10.1145/3360307

[33] Mattson, P., Reddi, V. J., Cheng, C., Coleman, C., Kanter, D., Micikevicius, P., ... & Jouppi, N., "MLPerf: An industry standard benchmark suite for machine learning performance," *IEEE Micro*, Vol.40, Issue 2, pp.8–16, 2020. https://doi.org/10.1109/MM.2020.2977146

[34] Menghani, G., "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," *ACM Computing Surveys*, Vol.*55*, Issue 12, pp.1-37, 2023.

[35] Wang, J., Yu, H., Wen, Z., & Zhang, S., "Federated learning with matched averaging," *arXiv preprint* arXiv:2002.06440, 2019.

AUTHORS PROFILE

Ojuawo Olutayo Oyewole-1 is a Nigerian computer scientist and academic, currently serving as Lecturer in the Department of Computer Science at The Federal Polytechnic, Ilaro. He holds a BSc in Computer Science from the University of Agriculture, Abeokuta, and an MSc in Network and Information Security from Kingston University, London. With over a decade of teaching experience, Ojuawo specializes in areas such as network security, information security, database design, cybersecurity, and data communication. His teaching portfolio includes courses like database management systems, computer programming, and ICT. Ojuawo has contributed to various academic publications and conferences. Notably, he co-authored the paper "Internet of Things (IoTs): Nigeria Road to Development," which explores the potential of IoT technologies in advancing Nigeria's socio-economic landscape. He is also a member of the Nigeria Computer Society (NCS), reflecting his active engagement in the professional community.

Folahan Joseph Jiboku-2 is a Lecturer in the Department of Computer Science at The Federal Polytechnic, Ilaro, Ogun State, Nigeria. He holds a bachelor's degree in computer information systems from Babcock University and master's degree in information systems from Lead City University. Mr. Jiboku's research interests encompass Information Security, Cybersecurity, and Management Information Systems. He has contributed to various scholarly publications, including studies on user-centered design, augmented reality, and network intrusion detection systems In addition to his academic pursuits, Mr. Jiboku is recognized as a data science enthusiast and front-end developer, with a keen interest in machine learning technologies . He is affiliated with the School of Pure and Applied Sciences at The Federal Polytechnic, Ilaro, where he is actively involved in teaching and research activities.

