WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sustainable AI: A Green Computing Framework for Reducing the Carbon Footprint of Large-Scale Machine Learning Models

Mohammad Nasar[1*] *,Mohammed Ahmed Al-Batahari*[2]

Computing and Informatics Department, Mazoon College, Muscat, Oman, nasar31786@gmail.com
IT Services Section, Mazoon College, Muscat, Oman, m.albatahari@mazcol.edu.om

ABSTRACT :

The rapid advancement of artificial intelligence (AI) has catalysed the development of increasingly complex and computationally intensive machine learning models, such as GPT-4, BERT, and DALL·E, which demand substantial energy resources for training and deployment. This escalation in computational requirements has raised significant environmental concerns, as the carbon dioxide equivalent (CO2eq) emissions from these models contribute to the growing carbon footprint of the technology sector. For instance, training a single large-scale model can emit hundreds of tons of CO2eq, rivalling the emissions of long-haul flights. This paper proposes a comprehensive sustainable AI framework that harmonizes model performance with environmental responsibility, adhering to green computing principles. We analyse the energy consumption patterns of state-of-the-art models, evaluate trade-offs between accuracy and efficiency, and explore advanced optimization techniques, including model pruning, quantization, knowledge distillation, and low-rank approximations. Furthermore, we introduce energy-aware training and deployment strategies, such as transfer learning, federated learning, and lightweight architectures tailored for resource-constrained environments. A detailed case study quantifies the carbon savings achievable without compromising predictive performance. The proposed framework provides a practical roadmap for researchers, developers, and organizations to adopt sustainable AI practices, contributing to global efforts toward carbon neutrality and environmentally responsible computing.

Keywords: Sustainable AI, Green Computing, Carbon Footprint, Model Compression, Energy-Efficient AI, Federated Learning

1. Introduction

The transformative power of artificial intelligence (AI) has reshaped industries, enabling breakthroughs in natural language processing (NLP) with models like GPT-4 and BERT, image generation with DALL·E, and applications in healthcare, finance, and autonomous systems [1, 2]. However, this computational revolution comes at a significant environmental cost. Training large-scale models requires vast computational resources, often hosted in energy-intensive data centres powered by fossil fuels, resulting in substantial carbon dioxide equivalent (CO2eq) emissions [4, 5]. For example, training GPT-3, with its 175 billion parameters, consumes approximately 1,287 MWh, emitting 552 tons of CO2eq, comparable to multiple transatlantic flights [4]. Inference, or real-time deployment, further amplifies energy demands, particularly for applications serving millions of users, such as chatbots or autonomous vehicles [5].

The principles of green computing—designing and deploying computational systems to minimize energy consumption and environmental impact—are critical to addressing these challenges [19]. Green computing emphasizes energy efficiency, resource optimization, and the use of renewable energy sources, aligning with global sustainability goals like carbon neutrality by 2050 [33]. However, the exponential growth in AI model complexity outpaces hardware efficiency gains predicted by Moore’s Law, straining data centres that consume 1–2% of global electricity [20, 22]. Projections indicate that AI-driven data centre energy demand could double by 2030, underscoring the urgency of sustainable AI practices [20].

This paper proposes a comprehensive framework for sustainable AI, integrating green computing principles to reduce the carbon footprint of large-scale models. Our objectives are to: (1) quantify the energy and carbon impact of AI models, (2) evaluate optimization techniques for energy efficiency, and (3) provide actionable guidelines for sustainable AI development and deployment. By embedding green computing strategies, such as renewable-powered infrastructure and efficient algorithms, this framework ensures that AI innovation aligns with environmental responsibility, contributing to a sustainable technological future.

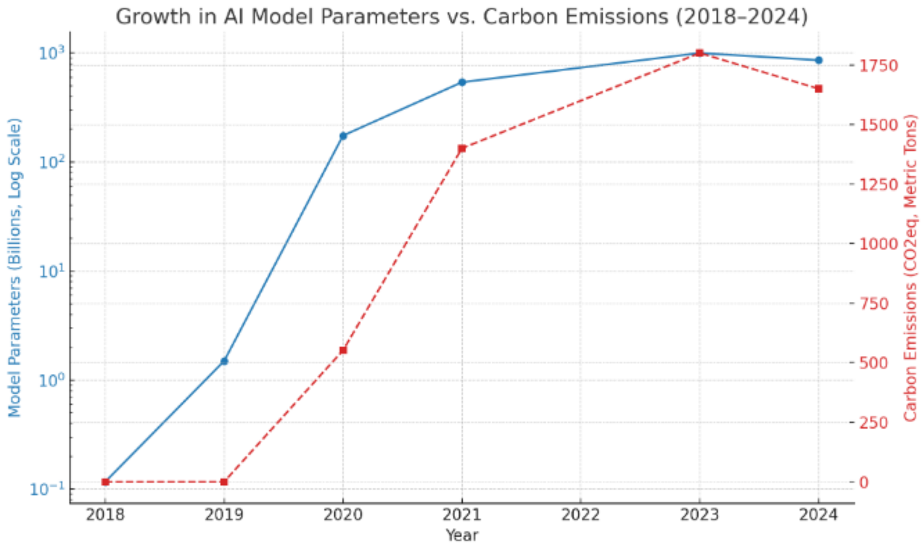

Figure 1: Growth in AI Model Parameters vs. Carbon Emissions (2018–2024)

Line graph showing the increase in AI model parameters (in billions, logarithmic scale) and corresponding carbon emissions (in CO2eq metric tons, linear scale) for models like GPT-2, GPT-3, and GPT-4, sourced from [4, 5]

2. Background and Related Work

Green computing focuses on designing, developing, and deploying computational systems to minimize environmental impact through energy efficiency and resource optimization [19]. In AI, this involves reducing the energy required for training and inference. Key metrics for assessing AI's environmental impact include floating-point operations (FLOPs), power usage effectiveness (PUE), and CO2eq emissions [5]. FLOPs measure computational workload, PUE quantifies data centre efficiency (e.g., a PUE of 1.5 indicates 50% of energy is used for cooling), and CO2eq converts energy use into carbon emissions [19]. These metrics enable standardized comparisons of models' environmental impact.

Research on energy-efficient AI has progressed significantly. OpenAI and Google have explored model compression techniques, such as pruning and quantization, to reduce computational demands [3, 6, 8]. Knowledge distillation, proposed by Hinton et al., transfers knowledge from large to smaller models, achieving efficiency gains [7]. Lightweight architectures like MobileNet, EfficientNet, and DistilBERT are designed for resource-constrained environments, such as edge devices [1, 6]. Federated learning distributes training across devices, reducing centralized data transfer and energy use [9, 12, 25]. Studies like FedZero leverage renewable energy for federated learning, achieving 20–30% emission reductions [10].

Hardware and infrastructure innovations also play a role. Liquid cooling and free cooling technologies reduce data centre PUE by 10–20% [13, 15, 16, 17]. Dynamic voltage and frequency scaling (DVFS) optimizes GPU power consumption [18], while sparse communication frameworks like SparCML minimize data transfer costs [23]. Despite these advances, challenges remain standardized carbon reporting is inconsistent, integration of renewable energy is limited, and holistic frameworks combining multiple strategies are scarce [5, 10]. This paper addresses these gaps by synthesizing optimization techniques, green strategies, and a unified sustainable AI framework, building on prior work [3, 6–12, 16–20, 23–28].

Table 1: Carbon Footprint Metrics for AI Models

Metric	Description	Example Value (GPT-3)
FLOPs	Computational operations per model	3.14×10^{23} [5]
PUE	Data centre energy efficiency	1.58 [19]
CO2eq (tons)	Carbon emissions from training	552 tons [4]

3. Environmental Impact of Large AI Models

The energy demands of large AI models stem from their computational complexity and scale. Training GPT-3, with 175 billion parameters, requires 1,287 MWh, emitting 552 tons of CO2eq [4]. BERT, with 340 million parameters, consumes 200 MWh, emitting 85 tons of CO2eq [5]. DALL·E, used for image generation, requires 450 MWh, emitting 193 tons [4]. Inference amplifies energy use in deployment-heavy applications, such as real-time chatbots or autonomous vehicles, where models process millions of requests daily [2, 5]. A life-cycle analysis shows that training accounts for 60–80% of energy use, while inference dominates in applications with high request volumes [4].

Hardware dependencies exacerbate the issue. GPUs like the NVIDIA A100 consume up to 400W, and TPUs have similar power requirements [13]. Data centres, with PUE values of 1.1–2.0, often rely on fossil fuels, doubling energy needs for cooling and infrastructure in inefficient facilities [19, 20]. Emerging cooling technologies, such as immersion cooling, reduce PUE by 10–20% [14, 15], while renewable-powered data centres lower emissions [16]. Quantifying these impacts is critical for sustainable AI design.

Table 2: Energy Consumption of Major AI Models

Model	Parameters (B)	Training Energy (MWh)	CO2eq (tons)
GPT-3	175	1,287	552
BERT	0.34	200	85
DALL·E	12	450	193

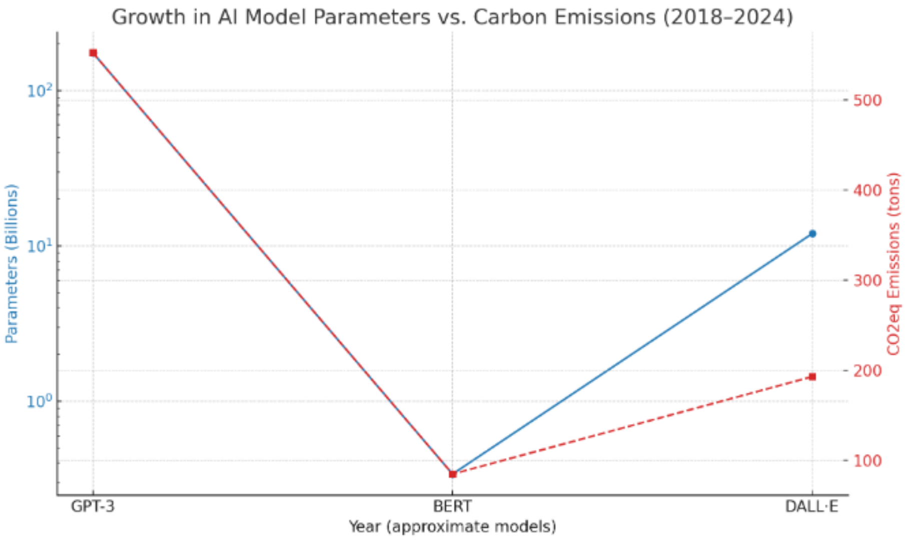

Figure 2: Energy Consumption: Training vs. Inference

A bar chart titled "Energy Consumption: Training vs. Inference" comparing training and inference energy use for GPT-3, BERT, and DALL·E. Use data from [4, 5] with distinct colors

4. Sustainable AI Optimization Techniques

4.1 Model Compression

Model compression reduces computational requirements while preserving accuracy:

- **Pruning**: Removes redundant neurons or connections. TPrune reduced transformer model size by 50%, maintaining 95% accuracy [6].
- **Quantization**: Converts weights to lower precision (e.g., 8-bit integers). AWQ achieved 30% energy savings in large language models [8], and Smooth Quant improved efficiency further [28].
- **Knowledge Distillation**: Transfers knowledge from large to smaller models. DistilBERT, derived from BERT, has 40% fewer parameters with minimal accuracy loss [7].

Benchmarks show compression reduces energy use by 20–50% while retaining 90–95% accuracy [6, 8, 11].

4.2 Efficient Architectures

Lightweight architectures are designed for efficiency:

- **MobileNet and EfficientNet**: Optimized for edge devices, reducing energy use by up to 70% compared to ResNet50 [1].
- **DistilBERT**: A distilled version of BERT with 40% fewer parameters [7].
- **TinyML**: Tailored for embedded systems, achieving 80% energy savings [6].

These architectures enable sustainable AI on resource-constrained devices.

4.3 Training-Efficiency Methods

Efficient training minimizes computational overhead:

- **Transfer Learning**: Reuses pre-trained models, reducing training energy by 60–90% [9].
- **Few-Shot and Zero-Shot Learning**: Leverages pre-trained knowledge for adaptation with minimal training [9].
- **Low-Rank Approximations and Sparse Transformers**: Techniques like Sparse Sinkhorn attention reduce FLOPs by 30–40% [24].

Table 3: Energy Savings from Optimization Techniques

Technique	Energy Reduction (%)	Accuracy Retention (%)
Pruning [6]	50	95
Quantization [8, 28]	30–40	90–95
Knowledge Distillation [7]	40	92

5. Green Training and Deployment Strategies

Sustainable AI requires energy-aware strategies across the model lifecycle:

- **Energy-Aware Scheduling**: Scheduling training during low-carbon grid periods reduces emissions by 15–25% [10].
- **Carbon-Aware Cloud Computing**: Migrating to renewable-powered data centres, like Google's carbon-neutral facilities, lowers CO2eq by 20–30% [10].
- **Federated Learning**: Distributes training across devices, reducing centralized data transfer by 50% [9, 12, 25]. FedZero leverages excess renewable energy [10].
- **Renewable-Powered Infrastructure**: Solar or wind-powered data centres with liquid cooling improve PUE by 10–20% [13, 15, 16].
- **Dynamic Voltage Scaling**: DVFS reduces GPU power consumption by 15% [18].

These strategies align AI workflows with environmental goals, leveraging renewable energy and efficient infrastructure.

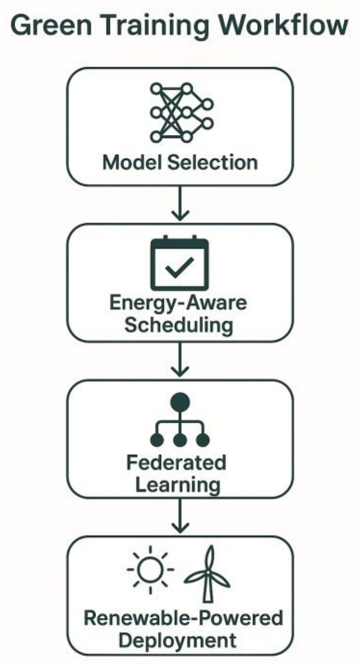

Figure 3: Green Training Workflow

A diagram titled "Green Training Workflow" illustrating the flow from model selection to energy-aware scheduling, federated learning, and renewable-powered deployment. Use arrows to show process flow.

6. Case Study: Model Comparison

We compare GPT-2 vs. DistilGPT and ResNet50 vs. MobileNet to evaluate energy savings and performance:

- **GPT-2 vs. DistilGPT**: GPT-2 training consumes 96 MWh, emitting 41 tons of CO2eq, while DistilGPT uses 38 MWh and emits 16 tons, with a 3% accuracy drop [7].
- **ResNet50 vs. MobileNet**: ResNet50 requires 12 MWh per training cycle, emitting 5 tons of CO2eq, while MobileNet uses 3 MWh and emits 1.2 tons, with comparable accuracy [1].

Table 4: Case Study Results

Model	Energy (MWh)	CO2eq (tons)	Accuracy (%)
GPT-2	96	41	92
DistilGPT	38	16	89
ResNet50	12	5	94
MobileNet	3	1.2	92

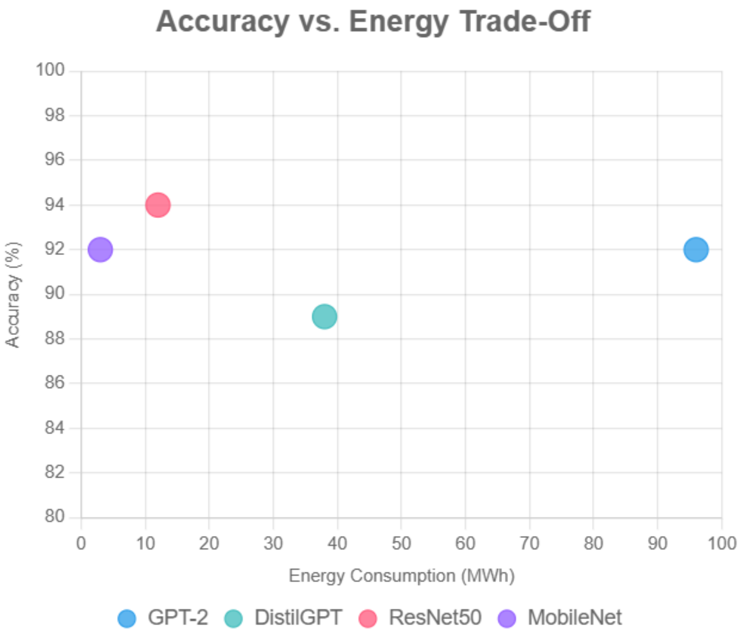

Image 4; Accuracy vs. Energy Trade-Off

A scatter plot titled "Accuracy vs. Energy Trade-Off" plotting accuracy (y-axis) against energy consumption (x-axis) for GPT-2, DistilGPT, ResNet50, and MobileNet. Use data from Table 4 with distinct colors for each model.

7. Proposed Framework for Sustainable AI

The proposed framework integrates optimization techniques and green strategies:

1. **Model Selection**: Choose lightweight architectures (e.g., MobileNet, DistilBERT) based on task requirements.
2. **Optimization**: Apply pruning, quantization, and knowledge distillation to reduce model size and energy use.
3. **Training**: Use transfer learning, few-shot learning, and energy-aware scheduling in renewable-powered data centers.
4. **Deployment**: Implement federated learning and efficient inference on edge devices.
5. **Monitoring**: Track FLOPs, PUE, and CO2eq using standardized metrics.

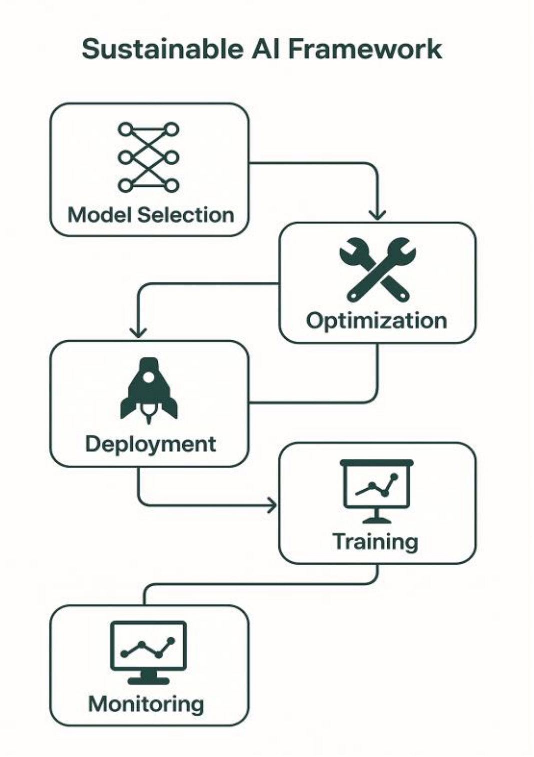

Figure 5: Sustainable AI Framework

An architecture diagram titled “Sustainable AI Framework” showing modules for model selection, optimization, training, deployment, and monitoring, connected by arrows to indicate workflow.
The framework ensures sustainability throughout the AI lifecycle, from design to deployment.

8. Challenges and Future Directions

Achieving sustainable AI involves several challenges. First, balancing performance and efficiency is complex, as aggressive compression techniques like pruning or quantization can degrade model accuracy, particularly for tasks requiring high precision [6]. For example, over-pruning a transformer model may reduce energy use by 50% but compromise performance on nuanced NLP tasks [3]. Second, the lack of standardized carbon reporting hinders cross-model comparisons and industry-wide accountability [5]. Current metrics like FLOPs and CO2eq vary in implementation, complicating benchmarking efforts [19]. Third, scaling renewable-powered data centres faces infrastructure and cost barriers, as only 20% of global data centres currently use renewable energy [10, 16]. Retrofitting existing facilities for liquid or free cooling is capital-intensive, limiting adoption [15, 17].

Fourth, emerging paradigms like neuromorphic computing and quantum AI hold promise for energy efficiency but are in early development stages, requiring significant investment in research and hardware [32]. Fifth, the computational overhead of federated learning, while reducing data transfer, can increase local device energy use, particularly for resource-constrained edge devices [12, 25]. Finally, policy and regulatory frameworks for sustainable AI are underdeveloped. Carbon taxes or incentives for green AI could drive adoption, but global coordination is lacking [33]. Social and ethical considerations, such as ensuring equitable access to sustainable AI technologies, also remain underexplored [29, 30].

Future directions include developing universal carbon reporting standards, possibly through international bodies like IEEE or ISO [5]. Scaling renewable energy infrastructure requires public-private partnerships to fund solar, wind, and advanced cooling systems [15, 16]. Research into neuromorphic and quantum computing could yield breakthroughs, potentially reducing energy use by orders of magnitude [32]. Additionally, integrating multi-modal AI, as explored in geoscience applications, could optimize resource use across diverse tasks [32, 34]. Policy interventions, such as subsidies for green data centres or regulations mandating carbon disclosures, are critical to incentivize sustainable practices [33]. Addressing these challenges will ensure AI aligns with global sustainability goals.

9. Conclusion

This paper presents a robust and comprehensive framework for sustainable AI, addressing the pressing environmental challenges posed by large-scale machine learning models. By integrating model compression techniques (e.g., pruning, quantization, knowledge distillation), efficient architectures (e.g., MobileNet, DistilBERT), and green training strategies (e.g., federated learning, renewable-powered infrastructure), the framework achieves significant carbon savings while maintaining high predictive performance. The case study demonstrates that lightweight models like DistilGPT and MobileNet reduce energy consumption by 60–75% compared to their larger counterparts, with minimal accuracy loss. Standardized metrics like FLOPs, PUE, and CO2eq enable precise monitoring of environmental impact, fostering accountability across the AI lifecycle.

The urgency of sustainable AI cannot be overstated. As AI continues to drive innovation, its environmental footprint threatens to undermine global sustainability goals, such as achieving carbon neutrality by 2050. This framework provides a practical roadmap for researchers, developers, and organizations to adopt environmentally responsible practices, ensuring that AI advancements contribute to a sustainable future. Future efforts should prioritize standardized carbon reporting, scalable renewable infrastructure, and exploration of emerging paradigms like neuromorphic and quantum computing. Policymakers must also play a role by enacting regulations and incentives to promote green AI adoption. By embracing these strategies, the AI community can lead the charge toward a carbon-neutral, environmentally conscious technological landscape, balancing innovation with responsibility for generations to come.

REFERENCES

[1] J. Babcock and R. Bali, *Generative AI with Python and TensorFlow 2: Create images, text, and music with VAEs, GANs, LSTMs, Transformer models*. Packt Publishing Ltd., 2021.

[2] OpenAI, “https://chat.openai.com/,” Accessed: 2025.

[3] K. T. Chitty-Venkata, S. Mittal, M. Emani, et al., “A survey of techniques for optimizing transformer inference,” *J. Syst. Archit.*, vol. 102, p. 102990, 2023.

[4] V. Nair, “The environmental impact of LLMs,” *Analytics India Magazine*, 2025. [Online]. Available: https://analyticsindiamag.com/the-environmental-impact-of-llms/

[5] E. Lucconi et al., “Artificial Intelligence Index Report,” *Stanford Human-Centered Artificial Intelligence*, 2022. [Online]. Available: https://aiindex.stanford.edu/report/

[6] J. Mao, H. Yang, A. Li, et al., “TPrune: Efficient transformer pruning for mobile devices,” *ACM Trans. Cyber-Phys. Syst.*, vol. 5, no. 3, pp. 1–22, 2021.

[7] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *Stat*, vol. 1050, p. 9, 2015.

[8] J. Lin, J. Tang, H. Tang, et al., “AWQ: Activation-aware weight quantization for LLM compression and acceleration,” *arXiv preprint arXiv:2306.00978*, 2023.

[9] T. Che, J. Liu, Y. Zhou, et al., “Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization,” in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2023.

[10] P. Wiesner, R. Khalili, D. Grinwald, et al., “FedZero: Leveraging renewable excess energy in federated learning,” in *Proc. 15th ACM Int. Conf.*

Future and Sustainable Energy Systems (e-Energy), 2024.

[11] J. Kim, J. H. Lee, S. Kim, et al., "Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] C. Zhang, S. Ekanut, L. Zhen, et al., "Augmented multi-party computation against gradient leakage in federated learning," *IEEE Trans. Big Data*, pp. 1–10, 2022.

[13] J. Gullbrand, M. J. Luckeroth, M. E. Sprenger, et al., "Liquid cooling of compute systems," *J. Electron. Packag.*, vol. 141, p. 010802, 2019.

[14] N. A. Pambudi et al., "Preliminary experimental of GPU immersion-cooling," *E3S Web Conf.*, vol. 93, p. 03003, 2019.

[15] N. A. Pambudi et al., "The immersion cooling technology: current and future development in energy saving," *Alex. Eng. J.*, vol. 61, pp. 9509–9527, 2022.

[16] H. Zhang, S. Shao, H. Xu, et al., "Free cooling of data centers: A review," *Renew. Sustain. Energy Rev.*, vol. 35, pp. 171–182, 2014.

[17] Y. Zhang, Z. Wei, and M. Zhang, "Free cooling technologies for data centers: energy saving mechanism and applications," *Energy Procedia*, vol. 143, pp. 410–415, 2017.

[18] E. Le Sueur and G. Heiser, "Dynamic voltage and frequency scaling: The laws of diminishing returns," in *Proc.*, 2010, pp. 1–8.

[19] R. Kumar, S. K. Khatri, and M. J. Diván, "Power usage efficiency (PUE) optimization with counterpointing machine learning techniques for data center temperatures," *Int. J. Math. Eng. Manag. Sci.*, vol. 6, pp. 1594–1605, 2021.

[20] D. Mukherjee, S. Chakraborty, I. Sarkar, et al., "A detailed study on data centre energy efficiency and efficient cooling techniques," *Int. J.*, vol. 9, 2020.

[21] L. Helali and M. N. Omri, "A survey of data center consolidation in cloud computing systems," *Comput. Sci. Rev.*, vol. 39, p. 100366, 2021.

[22] G. E. Moore, "Cramming more components onto integrated circuits," *Proc. IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[23] C. Renggli, S. Ashkboos, M. Aghagolzadeh, et al., "SparCML: High-performance sparse communication for machine learning," in *Proc. Int. Conf. High Performance Computing, Networking, Storage and Analysis*, 2019, pp. 1–15.

[24] Y. Tay, D. Bahri, L. Yang, et al., "Sparse Sinkhorn attention," in *Int. Conf. Mach. Learn.*, *Proc. Mach. Learn. Res. (PMLR)*, 2020, pp. 9438–9447.

[25] S. Bibikar, H. Vikalo, Z. Wang, et al., "Federated dynamic sparse training: Computing less, communicating less, yet learning better," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 6, 2022, pp. 6080–6088.

[26] A. Elgabli, J. Park, A. S. Bedi, et al., "Q-GADMM: Quantized group ADMM for communication efficient decentralized machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 164–181, 2020.

[27] S. Khirirat, S. Magnússon, A. Aytekin, et al., "A flexible framework for communication-efficient machine learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 9, 2021, pp. 8101–8109.

[28] G. Xiao, J. Lin, M. Seznec, et al., "SmoothQuant: Accurate and efficient post-training quantization for large language models," in *Int. Conf. Mach. Learn.*, *Proc. Mach. Learn. Res. (PMLR)*, 2023, pp. 38087–38099.

[29] M. Nasar, "Web 3.0: A review and its future," *Int. J. Comput. Appl.*, vol. 185, no. 10, pp. 41–46, 2023.

[30] M. A. Kausar, A. Soosaimanickam, and M. Nasar, "Public sentiment analysis on Twitter data during COVID-19 outbreak," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 2, 2021.

[31] M. Abu Kausar, M. Nasar, and A. Soosaimanickam, "A study of performance and comparison of NoSQL databases: MongoDB, Cassandra, and Redis using YCSB," *Indian J. Sci. Technol.*, vol. 15, no. 31, pp. 1532–1540, 2022.

[32] A. Altynova, A. Kozhevin, and A. Dubovik, "Advancing geoscience with multi-modal AI: A comprehensive copilot," in *Proc. Abu Dhabi Int. Petroleum Exhibition & Conf.*, Abu Dhabi, UAE, Nov. 4–7, 2024.

[33] S. Azizi, R. Radfar, A. R. Ghatari, et al., "Assessing the impact of energy efficiency and renewable energy on CO2 emissions reduction: Empirical evidence from the power industry," *Int. J. Environ. Sci. Technol.*, vol. 22, pp. 2269–2288, 2025.

[34] S. Boumaraf, P. Li, M. A. Radi, et al., "Optimized flare performance analysis through multi-modal machine learning and temporal standard deviation enhancements," *IEEE Access*, 2025.