Towards Sustainable AI: A Comprehensive Framework for Green AI

Abdulaziz Tabbakh [1,7*], Lisan Al Amin [2], Mahbubul Islam [3], G M Iqbal Mahmud [4], Imranul Kabir Chowdhury [5], Md Saddam Hossain Mukta [6]

[1]Computer Enigneering Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.
[2] University of Maryland Baltimore County, Maryland, USA.
[3]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh.
[4] Department of Industrial Engineering and Management, Khulna University of Engineering Technology, Khulna 9203, Bangladesh.
[5] Missouri University of Science and Technology, Missouri, USA.
[6] School of Engineering Sciences, LUT University, Lappeenranta, Finland.
[7]Interdisciplinary Research Center for Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia.

*Corresponding author(s). E-mail(s): atabakh@kfupm.edu.sa; Contributing authors: salamin1@umbc.edu; mislam201011@mscse.uiu.ac.bd; gmiqbalm@gmail.com; ichowdhury@mst.edu; saddam.mukta@lut.fi;

Abstract

The rapid advancement of Artificial Intelligence (AI) has brought significant benefits across various domains, yet it has also led to increased energy consumption and environmental impact. This paper positions Green AI as a crucial direction for future research and development. It proposes a comprehensive framework for understanding, implementing, and advancing sustainable AI practices. We provide an overview of Green AI, highlighting its significance and current state regarding AI's energy consumption and environmental impact. The paper explores sustainable AI techniques, such as model optimization methods, and the development of efficient algorithms. Additionally, we review energy-efficient hardware alternatives like Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs), and discuss strategies for designing and operating energy-efficient data centers. Case studies in Natural Language Processing (NLP) and Computer Vision illustrate successful implementations of Green AI practices. Through these efforts, we aim to balance the performance and resource efficiency of AI technologies, aligning them with global sustainability goals.

Keywords: Artificial Intelligence, GPU, Sustainable Computing

1 Introduction

Artificial Intelligence (AI) has emerged as a transformative force across a myriad of industries, driving unprecedented advancements and delivering significant benefits. From healthcare and finance to transportation and entertainment, AI technologies are revolutionizing the way we live, work, and interact with the world [1]. Machine learning algorithms are enabling predictive analytics, natural language processing is enhancing human-computer interactions, and computer vision is powering advancements in autonomous systems [2, 3]. These rapid advancements in AI are not only fostering innovation but are also addressing complex challenges, improving efficiencies, and creating new opportunities for growth and development.

Despite these substantial benefits, the rapid proliferation and scaling of AI technologies come with significant environmental costs. Training sophisticated AI models, particularly deep learning models, requires immense computational power, which translates into substantial energy consumption [4]. For instance, the training of large-scale models like GPT-3 or AlphaGo involves thousands of petaflop/s-days, consuming megawatts of power and resulting in a considerable carbon footprint [5–7]. datacenters housing AI infrastructure are major energy consumers, contributing to greenhouse gas emissions and exacerbating the environmental impact [8]. As AI continues to evolve and expand, its energy demands are projected to grow, posing a critical challenge to sustainability.

The environmental impact of AI underscores the urgent need for adopting sustainable practices in AI research and development. Sustainable AI, often referred to as green AI [9], aims to mitigate the environmental footprint of AI technologies by enhancing their energy efficiency and promoting the use of eco-friendly resources. This involves developing energy-efficient algorithms, optimizing model architectures, and leveraging renewable energy sources for datacenter operations. Embracing green AI not only aligns with global sustainability goals but also ensures the long-term viability of AI advancements by reducing their ecological impact.

This research aims to address the growing environmental concerns associated with the rapid advancement of AI technologies. Unlike previous studies focusing solely on improving model performance, this work integrates ecological sustainability as a core objective. The study differs from others by proposing an actionable framework that balances AI efficiency with ecological impact, motivated by the need to align AI development with global sustainability goals. Throughout this paper, 'framework' is used to denote the comprehensive structure for Green AI practices including model development and implementation, while 'methodology' is related to a systematic procedure for implementation. The term 'algorithm' refers to computational technique that drive AI models, whereas 'model' signifies the particular AI systems that are formulated based on these algorithms. The paper is organized as follows: Section 2 defines Green AI, discusses its significance, and provides an overview of AI's current energy consumption and environmental impact. Then, section 3 explores various methods for enhancing energy efficiency in AI, including model optimization and efficient algorithm development. Section 4 details the evolution and capabilities of GPUs, highlighting their significant energy consumption and environmental impact followed by section 5 which reviews energy-efficient hardware options like Tensor Processing Units (TPUs) and Field Programmable Gate Arrays (FPGAs), highlighting their benefits and applications. In section 6, a detailed implementations of Green AI practices in Natural Language Processing (NLP) and Computer Vision (CV) are highlighted and section 7 presents strategies for designing and operating datacenters that prioritize energy efficiency and renewable energy sources. Section 8 delineates the constraints and identifies potential avenues for amelioration in the present investigation, then section 10 conclude the paper by synthesizing the principal findings, advocating for the implementation of Green AI practices, and elucidates the prospective advantages of sustainable AI methodologies for future advancements.

2 Green AI: An Overview

In this section, we define Green AI, discuss its significance, and provide an overview of AI's current energy consumption and environmental impact.

2.1 Relevance of Green AI

Green AI refers to the development and deployment of artificial intelligence technologies that minimize environmental impact by reducing energy consumption and carbon emissions. The specific criteria for Green AI include energy efficiency (e.g., watt-hours per model training) and a reduced carbon footprint[9]. This involves optimizing AI algorithms, using energy-efficient hardware, and leveraging renewable energy sources to power AI operations.

The relevance of green AI lies in its alignment with global sustainability goals, such as those outlined in the United Nations' Sustainable Development Goals (SDGs) [10]. As AI technologies become more prevalent, their environmental footprint grows, making it imperative to integrate sustainability into AI research and development. Green AI helps in mitigating climate change and ensures the long-term viability and ethical responsibility of AI advancements [11].

2.2 Energy Consumption and Environmental Impact of AI

The energy consumption associated with AI technologies, particularly deep learning models, is substantial. Training large-scale AI models like GPT-3 and AlphaGo requires immense computational power, resulting in significant energy use. For instance, training GPT-3 required 1,287 MWh of electricity, equivalent to the annual energy consumption of over 120 U.S. homes [12, 13]. This translates to substantial carbon emissions, contributing to the overall carbon footprint of AI technologies.

A study by [14] highlighted that training a single AI model can emit as much carbon as five cars over their lifetimes. Moreover, datacenters housing AI infrastructure are major consumers of electricity. According to the International Energy Agency (IEA), datacenters worldwide consumed about 200 TWh of electricity in 2018, accounting for 1% of global electricity use and contributing to 0.3% of global CO2 emissions [15].

2.3 Importance of Integrating Sustainability in AI

Integrating sustainability in AI is critical for several reasons:
Environmental Protection: Reducing the energy consumption and carbon footprint of AI technologies helps mitigate their impact on climate change. Sustainable AI practices can significantly lower greenhouse gas emissions and conserve natural resources.
Economic Efficiency: Energy-efficient AI systems can reduce operational costs for businesses by lowering electricity bills and cooling requirements for datacenters. This economic benefit is particularly relevant as energy prices fluctuate and regulatory pressures increase.
Ethical Responsibility: As AI becomes integral to various aspects of society, ensuring that its development is environmentally responsible is an ethical imperative. Sustainable AI practices align with broader societal values of environmental stewardship and intergenerational equity.
Regulatory Compliance: With increasing regulatory focus on sustainability, adopting green AI practices can help organizations comply with environmental regulations and avoid potential penalties. It also prepares them for future regulations that may impose stricter limits on energy use and carbon emissions.
Reputation and Competitiveness: Companies that prioritize sustainability can enhance their reputation and competitiveness. Consumers and stakeholders are increasingly favoring organizations that demonstrate a commitment to environmental responsibility.

3 Sustainable AI Techniques

In this section, we explore various methods for enhancing energy efficiency in AI, including model optimization and efficient algorithm development.

3.1 Model Optimization

Model optimization plays a crucial role in reducing the energy consumption of AI systems [16]. Two key techniques in model optimization are pruning and quantization [17, 18].

Data Augmentation In addition to model optimization techniques, data augmentation plays a crucial role in enhancing model efficiency while reducing energy consumption. For example, convolution leverage provides an additional degree of freedom in optimizing models, as discussed by [19]. By augmenting datasets, we can reduce the need for extensive model training, thereby conserving computational resources.[20]

Pruning: Pruning involves removing parameters or neurons that have marginal effect on the output, from a neural network. This technique reduces the size of the model, thereby decreasing the computational resources required for training and inference and the storage needed to save the model [21]. Pruning can be done in several ways, Figure 1 depicts two types of model pruning example in Neural network and describe below:

1. Weight Pruning: refers to removing weights that contribute marginally to the model's output. This process involves identifying and eliminating small or zero-valued weights from the network [21]. The process can be formalized as:

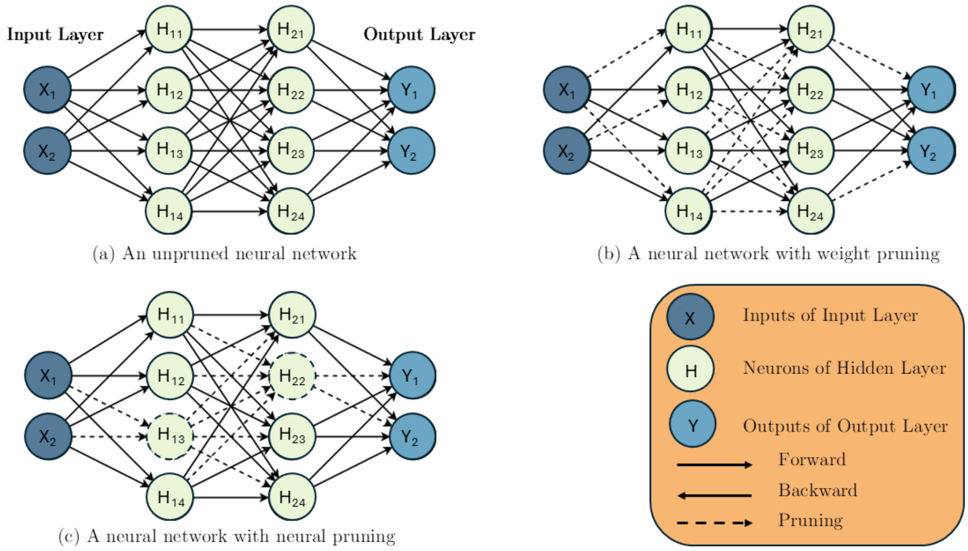
   $$\hat{W} = W \odot M \tag{1}$$

   where W is the original weight matrix, M is a binary mask (with elements 0 or 1), $\odot$ denotes element-wise multiplication, and $\hat{W}$ is the pruned weight matrix. The mask M is determined based on the significance of each weight, often using magnitude-based criteria [21]. Figure 1 (b) shows an example of the weight pruning. Every pruned or removed connection means that the parameter (weight) will not be trained or included in inference and will not be saved.
2. Neuron Pruning: This approach focuses on eliminating entire neurons or layer (with neurons in that layer) from network, particularly those with minimal impact on the model's overall performance. For a given layer l, neuron pruning can be represented as:

   $$\hat{y}_l = f(\hat{W}_l x_l + \hat{b}_l) \tag{2}$$

   where x_l is the input to layer l, $\hat{W}_l$ is the pruned weight matrix, $\hat{b}_l$ is the pruned bias vector, f is the activation function, and $\hat{y}_l$ is the output of the pruned layer. The dimensions of $\hat{W}_l$ and $\hat{b}_l$ are reduced compared to their unpruned counterparts [22]. Figure 1 (c) shows an example of a neural network with some pruned neurons. When a neuron is pruned, all weight associated with the connections that goes in and out of the neuron as well as the bais are not considered in the training and inference and are not stored resulting in reduction in the computational complexity and storage requirements.

Pruning reduces the size of the model and speeds up computation, leading to lower energy consumption. A pruned model requires less memory and fewer operations, resulting in a reduction in power usage. Experimental results have shown that pruning can reduce the number of parameters in a neural network - and hence the storage needed - by up to 90% without significantly affecting accuracy. Experiments also show that the number of parameter in AlexNet can be reduced by almost 89% from 61 million to 6.7 million parameters without any loss in the accuracy while the number of parameters in VGG-16 CNN model can go as low as 10.3 million from 138



(a) An unpruned neural network

(b) A neural network with weight pruning

(c) A neural network with neural pruning

Fig. 1: Different types of model pruning example in Neural network.

million parameters in the the orginial model (almost 93% reduction) while maintaining the same accuracy [23]. However, Pruning, while reducing the number of model parameters, typically incurs a performance cost, often manifesting as a slight decrease in accuracy. Apply model pruning to AlexNet resulted in a 90% reduction in energy consumption while causing only a minor 0.5% drop in accuracy [23].

Quantization: Quantization involves reducing the precision of the numbers used to represent the model parameters. Instead of using 32-bit floating-point numbers, which are standard in many AI models, quantization uses lower precision formats such as 16-bit or 8-bit integers [18, 24]. This reduction in precision leads to:

1. Smaller Model Size: Lower precision requires fewer bits to store each parameter, thus reducing the overall model size.
2. Faster Computation: Lower precision arithmetic operations are faster to compute, which results in shorter training and inference times.
3. Energy Efficiency: Both the reduced memory footprint and the faster computations contribute to lower energy consumption.

Quantization can be applied during both training and inference [24]. Techniques such as post-training quantization and quantization-aware training help maintain the model's performance while achieving significant reductions in energy use.

3.2 Efficient Algorithms

The design of efficient algorithms is paramount for creating energy-efficient AI systems. Efficient algorithms minimize the computational complexity and optimize the use of resources, thus reducing energy consumption. Key strategies include:

Algorithmic Efficiency: Algorithmic efficiency focuses on optimizing the algorithms themselves to require fewer computational resources. This can be achieved through:

1. Optimized Data Structures: Using data structures that are more efficient in terms of memory and processing requirements.
2. Efficient Mathematical Operations: Reducing the complexity of mathematical operations involved in the algorithms.
3. Sparse Representations: Leveraging sparsity in data and computations to reduce the number of active computations [25, 26].

For example, algorithms such as efficient neural architecture search (NAS) [27] can automatically design neural network architectures that are not only high-performing but also resource-efficient. These optimized architectures can significantly lower the energy consumption of AI models.

Approximate Computing: Approximate computing [28] is a paradigm where computations are performed in a manner that trades off some accuracy for significant gains in efficiency. This is particularly useful in scenarios where perfect accuracy is not crucial. Techniques include:

1. Probabilistic Computing: Using probabilistic methods to approximate complex computations.
2. Adaptive Precision: Dynamically adjusting the precision of computations based on the current requirements of the task.

Approximate computing can lead to substantial energy savings by reducing the number and precision of required operations, which directly impact energy consumption.

Transfer Learning: Transfer learning leverages pre-trained models on similar tasks, reducing the need to train models from scratch [29]. By reusing existing models and fine-tuning them for specific tasks, transfer learning significantly reduces the computational resources required for training, thus lowering energy consumption. This approach is particularly effective in applications such as natural language processing and computer vision, where large datasets and extensive training are otherwise necessary.

In summary, sustainable AI techniques such as model optimization and the design of efficient algorithms are critical for reducing the environmental impact of AI technologies. Pruning and quantization help streamline models, while efficient algorithm design ensures minimal resource use. These practices contribute to the development of energy-efficient AI systems, aligning with broader sustainability goals.

4 Current State of GPUs and Their Role in AI

GPUs have evolved from specialized hardware for gaming and image processing into powerful accelerators, driving many modern AI and machine learning applications. Their highly parallel architecture allows GPUs to excel at the massive computational workloads required to train deep neural networks. This section details the evolution and capabilities of GPUs, highlighting their significant energy consumption and environmental impact. Evolution of GPU and concurrent technology development summarized in Figure 2.

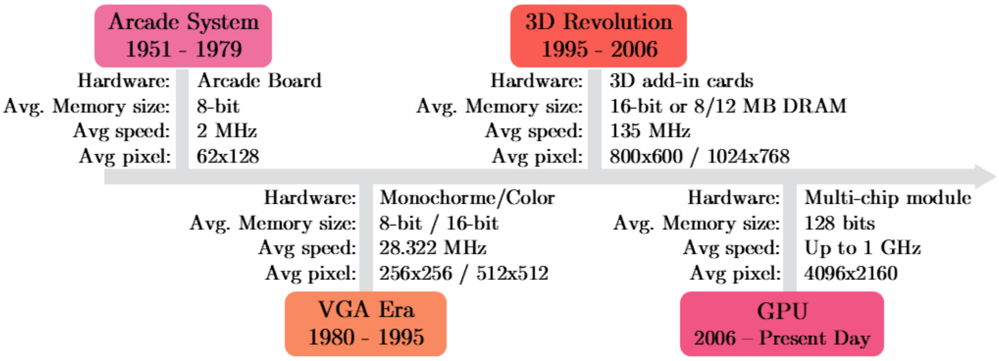


Fig. 2: Evolution of GPU in Chronological Arrangement (1951-2024)

The evolution of GPUs traces back to simple arcade boards and display adapters in the early days of computing. As depicted in Figure 2, the “3D Revolution” from 1995-2006 brought 3D graphics cards capable of higher resolutions [30]. A major turning point was the introduction of NVIDIA’s CUDA in 2006 [31], enabling general-purpose computing on GPUs (GPGPU). This paved the way for GPUs to be leveraged for AI and scientific computing workloads that could harness their parallel processing capabilities. Today’s GPUs like AMD’s Zen 2 [32] and NVIDIA’s latest Tensor Core GPUs [33] are highly advanced, integrating cutting-edge memory technologies and reaching teraflop computing performance. Their unparalleled ability to accelerate deep learning training and inference has made GPUs indispensable for modern AI. However, this incredible computing power comes with significant drawbacks in terms of energy consumption, cost, and environmental impact.

4.1 Drawbacks of Current GPUs for Green AI

While enabling breakthroughs in AI, the widespread use of power-hungry GPUs poses challenges for environmentally sustainable computing:

1. High Energy Consumption: Modern GPUs contain thousands of cores operating in parallel, resulting in significant energy demands and heat generation that requires

complex cooling solutions. The energy-intensive nature of GPU [34] computing has raised concerns about its environmental footprint.

2. Limited Memory Capacity: While equipped with high-bandwidth memory, GPUs have limited overall memory capacity compared to traditional servers [35]. As AI models become increasingly large and data-intensive, this constraint can limit the scalability and efficiency of computations.
3. Steep Cost: GPUs remain an expensive investment for organizations and individuals, both in terms of hardware acquisition and operating costs like power consumption and cooling infrastructure. This financial barrier limits accessibility and could stifle innovation in AI.
4. Programmability Challenges: Programming frameworks like CUDA require specific expertise and can involve significant code modifications to leverage GPU parallelism effectively across different architectures.

As AI systems continue growing in complexity and scale, the heavy reliance on energy-intensive GPU computing poses challenges for environmental sustainability and equitable access. Addressing these drawbacks is crucial for realizing the full potential of AI in an environmentally responsible and inclusive manner.

5 Green Hardware Alternatives

The search for more sustainable AI solutions necessitates the exploration of energy-efficient hardware. Two prominent technologies, Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs), have emerged as leading alternatives to traditional GPUs [36]. This section delves into the benefits and applications of these green hardware alternatives and presents a case study to illustrate their effectiveness. In addition to TPUs and FPGAs, emerging technologies like neuromorphic computing and quantum computing offer promising future solutions for reducing the energy footprint of AI systems. Neuromorphic chips mimic brain-like structures to enhance energy efficiency, while quantum computing offers the potential for exponential increases in processing power with significantly lower energy consumption.

Initially, we examine the trade-offs between the growing parameter count in models and the associated increases in computational and environmental costs, substantiated by Fig 3. Subsequently, we focus on mitigating these challenges by discussing emerging research in model optimization techniques, such as compression, pruning, and quantization.

5.1 Parametric Escalation and Associated Cost in Machine Learning

The evolution of GPU technology has played a significant role in enabling the development of larger and more complex models in both the NLP and CV domains. These models, characterized by increased parameters, can learn intricate patterns in data and perform advanced tasks. However, this upward trend in parameter count also ushers in challenges such as escalated storage requirements, increased training times, higher costs, and an expanded carbon footprint.

Popular models from both domains have exhibited an exponential increase in parameter count over time, growing from millions to billions of parameters. This has led to advancements in performance but has also resulted in significantly increased computational costs. The increased financial burden includes the procurement and upkeep of potent hardware needed for training these large models, as well as the associated energy costs.

The Fig 3 showcases the trends in the growth of parameters for popular NLP and CV models over time.

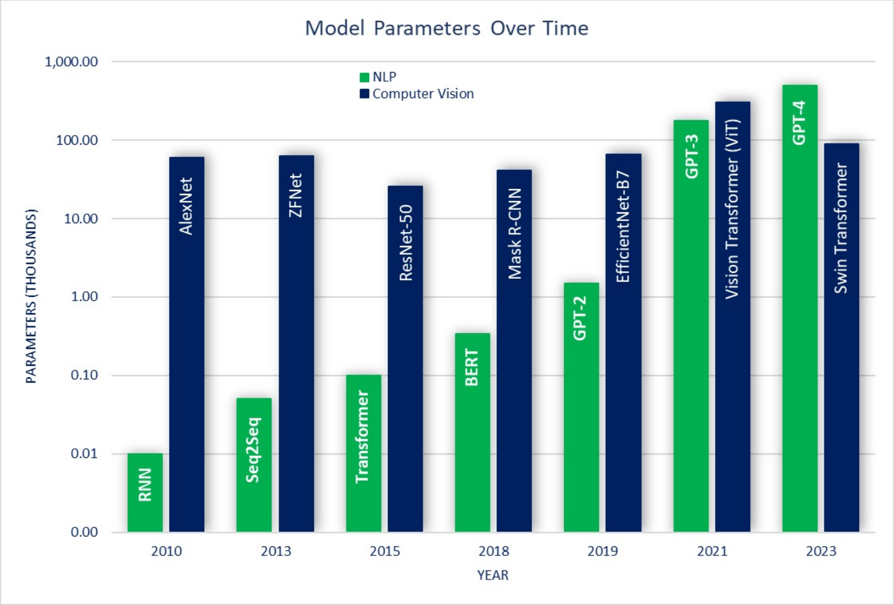


Fig. 3: Increase in Model Parameters in NLP Research Over Time & Increase in Model Parameters in Computer Vision Research Over Time

The implications of this trend are clear: while we now possess highly capable models, their training costs and environmental impacts have grown significantly. This has led to a surge in research focused on model pruning, compression, and quantization techniques, aiming to reduce these models' size, computational requirements, and environmental footprint without significantly compromising performance.

5.2 Mitigating GPU Dependencies

As deep learning models continue to increase in size and complexity, strategies to decrease computational requirements and model sizes have become paramount. This is necessary not only to mitigate the dependency on expensive, high-performance GPUs but also to allow for more efficient training, reduced memory footprint, and the deployment of models on edge devices. Among the most effective techniques being developed are model compression, pruning, and quantization.

Case Study	Technique	Description	Benefits
DistilBERT	Knowledge Distillation	Training a smaller model to mimic a larger model (BERT)	97% of BERT's performance, 60% faster, significantly less power
Pruning and Quantization	Model Optimization	Removing redundant parameters (pruning) and reducing weight precision (quantization)	Reduced model size and energy use by up to 50%
MobileNetV2	Depthwise Separable Convolutions	Uses separable convolutions to reduce computations	High accuracy, efficient on mobile and edge devices, lower energy consumption
ProxylessNAS	Neural Architecture Search (NAS)	Optimizes neural network architectures for specific hardware constraints	Substantial energy savings, maintains competitive performance

Table 1: Summary of Case Studies in Green AI

Model Compression: Model compression techniques aim to reduce the size of the model without significantly impacting its performance. Methods such as knowledge distillation [37] are popular in this domain. In knowledge distillation, a smaller model (student) is trained to mimic the behavior of a larger, more accurate model (teacher). This approach has demonstrated success in several instances. For example, DistilBERT, a distilled version of the BERT model, retains 97% of BERT's performance but is 40% smaller in size.

Model Pruning: Model pruning involves eliminating unnecessary parameters or weights in a model. This technique effectively reduces the size of the model and the computational resources required. Different approaches such as magnitude-based pruning (removing weights with small absolute values) and structural pruning (removing entire neurons or layers) have shown promise. The Lottery Ticket Hypothesis [21], suggesting the existence of smaller, highly performant subnetworks within larger networks, has further underlined the potential of pruning techniques.

Quantization: Quantization is another effective technique to reduce the computational requirements and size of a model [18]. It involves reducing the precision of the numbers used to represent the weights in the model, typically from 32-bit floating-point numbers to lower-precision formats like 16-bit integers or even lower. Quantization has been shown to scale down the memory footprint and computational requisite of models remarkably, with a relatively small impact on model performance [38].

Impact on Reducing GPU Dependencies: Model compression, pruning, and quantization techniques have significantly reduced the dependency on high-end GPUs. They have allowed researchers and practitioners with less computational power to engage with state-of-the-art models. Moreover, these methods have considerably reduced the energy consumption and environmental impact of model training, thereby contributing to the sustainability of the field.

Additionally, the decrease in model sizes has enabled the deployment of advanced models on edge devices, such as mobile phones and IoT devices, thereby widening the applicability of deep learning models in real-world scenarios [39].

6 Case Studies in Green AI

Case studies in Natural Language Processing (NLP) and Computer Vision illustrate successful implementations of Green AI practices.

6.1 NLP Implementations

Energy-Efficient NLP Models: NLP is a computationally intensive field within AI, often requiring substantial energy resources to train and operate large models. However, recent advancements have focused on developing energy-efficient NLP models to mitigate environmental impacts.

One notable example is the development of the DistilBERT model by Hugging Face [40]. DistilBERT is a smaller, faster, and more efficient version of BERT (Bidirectional Encoder Representations from Transformers) [41]. By employing knowledge distillation, where a smaller model is trained to mimic the behavior of a larger model, DistilBERT achieves 97% of BERT's performance while being 60% faster and requiring significantly less computational power [40].

Another approach involves pruning and quantization techniques. Pruning removes redundant parameters from a model, while quantization reduces the precision of the model's weights. These techniques can substantially decrease the model size and energy consumption without greatly affecting performance. Research by [42] demonstrated that applying these methods to transformer models like GPT-3 can reduce energy usage by up to 50% .

6.2 Computer Vision Applications

Sustainable Computer Vision: Computer vision tasks, such as image recognition and object detection, are typically energy-intensive due to the large datasets and complex computations involved. However, innovative practices and models have been developed to enhance energy efficiency in this domain.

A case study on the MobileNetV2 [43] architecture highlights its design for efficiency. MobileNetV2 uses depthwise separable convolutions, significantly reducing the number of computations compared to traditional convolutional networks. This architecture maintains high accuracy while operating efficiently on mobile and edge devices, leading to lower energy consumption during inference.

Another example is the use of efficient NAS methods. Efficient NAS algorithms, such as ProxylessNAS [44], optimize neural network architectures for specific hardware constraints, balancing performance and energy efficiency. In practical applications, these optimized models have shown substantial energy savings while maintaining competitive performance metrics.

6.3 Other Domains

Robotics: In robotics, energy efficiency is crucial for extending the operational time of robots and reducing the environmental impact of their deployment. One example is the development of energy-efficient algorithms for autonomous navigation.

Researchers have created algorithms that optimize the robot's path and motion to minimize energy consumption while ensuring reliable operation [45]. These advancements are particularly important for field robots operating in remote or resource-limited environments.

Healthcare: Healthcare applications of AI also benefit from green practices. For instance, in medical imaging, AI models designed for efficient image processing can reduce the energy required for tasks like disease detection and diagnosis. An example is the use of lightweight convolutional neural networks (CNNs) in detecting abnormalities in X-ray and MRI scans [46]. These models are optimized to run efficiently on standard medical equipment, thereby reducing the overall energy footprint of medical diagnostics.

7 Efficient Datacenters

Efficient datacenters are critical in reducing the environmental impact of AI operations. This section outlines the design principles, renewable energy sources, and best practices for creating energy-efficient datacenters. Accordingly, we present strategies for designing and operating datacenters that prioritize energy efficiency and renewable energy sources. Figure 4 represents key features of sustainable and efficient datacenter design.

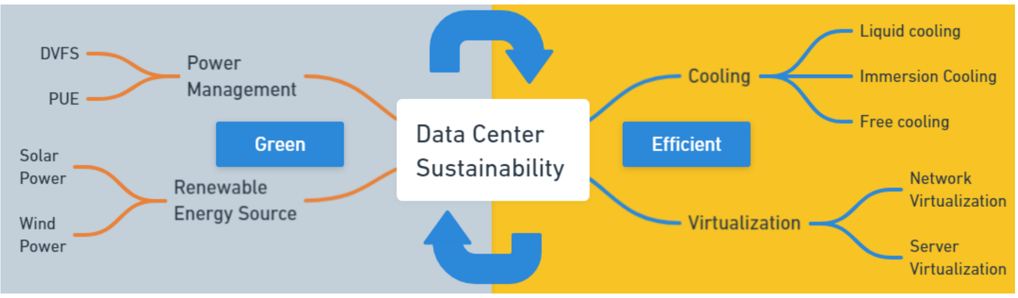


Fig. 4: Key features of sustainable and efficient datacenter design

7.1 Design Principles

Cooling Systems: Innovative cooling technologies are essential to reduce the energy consumption of datacenters. Traditional cooling methods, which rely heavily on air conditioning, are highly energy-intensive [47]. Modern alternatives include:

1. Liquid Cooling: Direct liquid cooling systems, where a coolant is brought into direct contact with heat-generating components, have been shown to be more efficient than traditional air cooling [48].
2. Immersion Cooling: In this method, servers are submerged in a thermally conductive but electrically insulating liquid. This technique significantly reduces the energy required for cooling by efficiently dissipating heat [49, 50].

3. Free Cooling: Utilizing natural air or water sources from the surrounding environment to cool datacenters. Free cooling can substantially cut down energy use by reducing the reliance on mechanical cooling systems [51].

Research indicates that innovative cooling technologies can reduce the energy consumption of datacenters by up to 50%, contributing significantly to overall energy efficiency [52].Cooling techniques like immersion cooling are more practical in certain geographic regions with high ambient temperatures, while liquid cooling may be more feasible in regions with limited water resources. The scalability of these techniques is also dependent on the size of the data centers.

Power Management: Effective power management strategies are vital for optimizing energy use in datacenters. Key strategies include:

1. Dynamic Voltage and Frequency Scaling (DVFS): Adjusting the voltage and frequency according to the current workload, which reduces power consumption during periods of low activity [53].
2. Power Usage Effectiveness (PUE) Optimization: PUE is a metric used to determine the energy efficiency of a datacenter. Optimizing PUE involves minimizing the energy used by the facility infrastructure (e.g., cooling and power distribution) relative to the energy used by the IT equipment [54].

Implementing these power management strategies can lead to substantial reductions in energy use.

7.2 Renewable Energy Sources

Solar and Wind: The adoption of renewable energy sources such as solar and wind power is crucial for powering datacenters sustainably. These sources offer several benefits:

1. Solar Power: Installing solar panels on or near datacenter facilities can provide a significant portion of the required energy. Solar farms can also be established in regions with high solar irradiance to support datacenters.
2. Wind Power: Wind turbines can be deployed in areas with consistent wind patterns. Offshore wind farms, in particular, offer high energy output potential with minimal land use.

The integration of solar and wind power into datacenter operations has been shown to reduce carbon emissions and operating costs, making them attractive options for green energy solutions.

7.3 Best Practice

Selecting the optimal location and designing the architecture of datacenters are critical for enhancing energy efficiency. Placing datacenters in cooler climates can reduce the need for artificial cooling. Proximity to renewable energy sources and reliable grid infrastructure also supports sustainable operations. Designing datacenters with energy efficiency in mind includes optimizing server rack layout for better airflow, using energy-efficient building materials, and incorporating natural cooling elements into the structure. Studies have shown that datacenters located in cooler climates and designed with energy efficiency principles can achieve significant savings in energy cost [55]. On the other hand, Virtualization technologies [56] play a significant role in improving the energy efficiency of datacenters. Consolidating multiple virtual servers onto a single physical server reduces the number of physical machines required, lowering both energy consumption and cooling needs. Optimizing network resources through virtualization reduces the need for physical networking hardware, leading to energy savings. By enhancing resource utilization and reducing the physical hardware footprint, virtualization can lead to significant improvements in energy efficiency.

Managerial Implications: The findings of this study provide actionable insights for AI managers and policymakers. By adopting Green AI practices, organizations can reduce operational costs through energy savings while contributing to sustainability goals. The proposed framework can guide AI project managers in balancing model performance with environmental impact, particularly in resource-constrained environments. Additionally, the recommendations can assist policymakers in formulating regulations that promote sustainable AI development across industries.Policymakers can foster the adoption of Green AI practices by introducing tax incentives for companies adopting energy-efficient hardware and mandating carbon footprint reporting for AI models. Additionally, setting up governmental funding for research in sustainable AI technologies would accelerate progress.

8 Limitations of the Study

While this paper provides a comprehensive framework for advancing Green AI, several limitations need to be acknowledged:

Scope of Coverage: The paper primarily focuses on widely discussed techniques and hardware solutions for Green AI, such as model optimization, TPUs, and FPGAs. However, it does not delve into newer, emerging technologies or less explored areas that might also contribute significantly to sustainable AI.

Empirical Validation: Much of the discussion is based on theoretical insights and case studies from existing literature. There is a lack of original empirical data or experimental results that could substantiate the claims and recommendations made in the paper. Future work should aim to include empirical evaluations of the proposed sustainable AI practices.

Interdisciplinary Integration: While the paper identifies the need for interdisciplinary collaborations, it does not extensively cover how such integrations can be effectively implemented. Detailed strategies for fostering collaborations between AI researchers, environmental scientists, and policy makers are essential for a holistic approach to Green AI.Additionally, regular workshops and joint publications can help bridge the gap between these disciplines.

Regulatory and Policy Frameworks: The paper highlights the importance of regulatory frameworks and policy recommendations but provides only a broad overview. Specific policy proposals and detailed regulatory guidelines are necessary to drive actionable change in AI sustainability practices.

Scalability and Accessibility: The paper suggests techniques and hardware alternatives that might not be easily scalable or accessible to all organizations, especially smaller enterprises or researchers with limited resources. Addressing these disparities is crucial for widespread adoption of Green AI practices. Many Green AI practices, such as building energy-efficient datacenters or adopting advanced cooling systems, may be challenging for smaller organizations or those in developing countries due to resource limitations. Cloud-based AI services and open-source tools offer scalable, accessible solutions that can bridge this gap.

By recognizing these limitations, future research can build upon this work, incorporating more diverse technologies, empirical data, detailed interdisciplinary strategies, specific policy guidelines, and solutions that are scalable and accessible to a broader audience.

9 Data Availability Statement

This study did not generate or analyze any datasets. As such, data sharing is not applicable to this article. The research presented here is based on theoretical analysis, literature review, or other methodologies that did not involve the creation or examinati‘1 on of datasets. Consequently, there are no data associated with this study to be made available.

10 Conclusion

This paper has proposed a framework for Green AI, incorporating the latest research on energy-efficient computing. We have also addressed the scalability of Green AI practices for smaller organizations and developing countries. While GPUs have significantly propelled the advancements in AI, their high energy consumption and environmental impact necessitate a shift towards more sustainable practices. This paper has proposed a comprehensive framework for Green AI, covering model optimization techniques, energy-efficient hardware alternatives, and strategies for designing efficient datacenters. By exploring case studies in NLP and Computer Vision, we have demonstrated the feasibility and benefits of adopting sustainable AI practices. However, to fully realize the potential of Green AI, future research must address the current limitations by incorporating empirical validation, fostering interdisciplinary collaborations, and developing specific policy guidelines. Future research should focus on emerging technologies like quantum computing and their potential to further reduce AI's environmental footprint. Additionally, making sustainable AI practices scalable and accessible to a wider audience is crucial for their widespread adoption. By aligning AI development with global sustainability goals, we can ensure that the advancements in this field are not only technologically innovative but also environmentally responsible, paving the way for a more sustainable future in AI.

Acknowledgements. The Authors declare that there is no conflict of interest.

Declarations

- Data availability: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

[1] Makridakis, S. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures* **90**, 46–60 (2017).

[2] Wu, X. *et al.* A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* **135**, 364–381 (2022).

[3] Liu, S. *et al.* Edge computing for autonomous driving: Opportunities and challenges. *Proceedings of the IEEE* **107**, 1697–1716 (2019).

[4] Wu, C.-J. *et al.* Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* **4**, 795–813 (2022).

[5] Miao, Q. *et al.* Dao to hanoi via desci: Ai paradigm shifts from alphago to chatgpt. *IEEE/CAA Journal of Automatica Sinica* **10**, 877–897 (2023).

[6] Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022).

[7] Gholami, A. *et al.* Ai and memory wall. *IEEE Micro* 1–5 (2024).

[8] Libertson, F., Velkova, J. & Palm, J. Data-center infrastructure and energy gentrification: perspectives from sweden. *Sustainability: science, practice and policy* **17**, 152–161 (2021).

[9] Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green ai. *Communications of the ACM* **63**, 54–63 (2020).

[10] Fund, S. Sustainable development goals. *Available at this link: https://www. un. org/sustainabledevelopment/inequality* (2015).

[11] Yigitcanlar, T., Mehmood, R. & Corchado, J. M. Green artificial intelligence: Towards an efficient, sustainable and equitable technology for smart cities and futures. *Sustainability* **13**, 8952 (2021).

[12] Abedin, M. *et al.* Material to system-level benchmarking of cmos-integrated rram with ultra-fast switching for low power on-chip learning. *Scientific Reports* **13**, 14963 (2023).

[13] You, J., Chung, J.-W. & Chowdhury, M. Zeus: Understanding and optimizing gpu energy consumption of dnn training **1**, 119–139 (2023).

[14] Hao, K. Training a single ai model can emit as much carbon as five cars in their lifetimes. *MIT Technology Review* **75**, 103 (2019).

[15] Jones, N. *et al.* How to stop data centres from gobbling up the world's electricity. *Nature* **561**, 163–166 (2018).

[16] Kulkarni, U. *et al.* Ai model compression for edge devices using optimization techniques 227–240 (2021).

[17] Liu, Z., Sun, M., Zhou, T., Huang, G. & Darrell, T. Rethinking the value of network pruning (2018).

[18] Jacob, B. *et al.* Quantization and training of neural networks for efficient integer-arithmetic-only inference 2704–2713 (2018).

[19] Bukhari, A. H. *et al.* Predictive analysis of stochastic stock pattern utilizing fractional order dynamics and heteroscedastic with a radial neural network framework. *Engineering Applications of Artificial Intelligence* **135**, 108687 (2024).

[20] Reddy, M. I., Rao, P. V., Kumar, T. S. & K, S. R. Encryption with access policy and cloud data selection for secure and energy-efficient cloud computing. *Multimedia Tools and Applications* **83**, 15649–15675 (2024).

[21] Frankle, J. & Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks (2018).

[22] Lin, M. *et al.* Channel pruning via automatic structure search. *arXiv preprint arXiv:2001.08565* (2020).

[23] Surianarayanan, C., Lawrence, J. J., Chelliah, P. R., Prakash, E. & Hewage, C. A survey on optimization techniques for edge artificial intelligence (ai). *Sensors* **23**, 1279 (2023).

[24] Gray, R. M. & Neuhoff, D. L. Quantization. *IEEE transactions on information theory* **44**, 2325–2383 (1998).

[25] Wright, J. *et al.* Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE* **98**, 1031–1044 (2010).

[26] Zhang, Z., Xu, Y., Yang, J., Li, X. & Zhang, D. A survey of sparse representation: algorithms and applications. *IEEE access* **3**, 490–530 (2015).

[27] Liu, C. *et al.* Progressive neural architecture search 19–34 (2018).

[28] Han, J. & Orshansky, M. Approximate computing: An emerging paradigm for energy-efficient design 1–6 (2013).

[29] Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big data* **3**, 1–40 (2016).

[30] Dehal, R. S., Munjal, C., Ansari, A. A. & Kushwaha, A. S. Gpu computing revolution: Cuda 197–201 (2018).

[31] Shams, R. & Kennedy, R. A. Efficient histogram algorithms for nvidia cuda compatible devices 418–422 (2007).

[32] Bhargava, R. & Troester, K. Amd next generation" zen 4" core and 4 th gen amd epyc™ server cpus. *IEEE Micro* (2024).

[33] Hanindhito, B. & John, L. K. Accelerating ml workloads using gpu tensor cores: The good, the bad, and the ugly 178–189 (2024).

[34] James, A. Energy efficiency and design challenges in analogue memristive chips. *Nature Reviews Electrical Engineering* **1**, 6–7 (2024).

[35] Zhao, H. *et al.* Towards fast setup and high throughput of gpu serverless computing. *arXiv preprint arXiv:2404.14691* (2024).

[36] Vandendriessche, J. *et al.* Environmental sound recognition on embedded systems: From fpgas to tpus. *Electronics* **10**, 2622 (2021).

[37] Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network (2015).

[38] Hascoet, T., Zhuang, W., Febvre, Q., Ariki, Y. & Takiguchi, T. Reducing the memory cost of training convolutional neural networks by cpu offloading. *Journal of Software Engineering and Applications* **12**, 307–320 (2019).

[39] Zawish, M., Davy, S. & Abraham, L. Complexity-driven model compression for resource-constrained deep learning on edge. *IEEE Transactions on Artificial Intelligence* (2024).

[40] 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).

[41] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[42] Dettmers, T., Lewis, M., Belkada, Y. & Zettlemoyer, L. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems* **35**, 30318–30332 (2022).

[43] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks 4510–4520 (2018).

[44] Cai, H., Zhu, L. & Han, S. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332* (2018).

[45] Mohammed, A., Schmidt, B., Wang, L. & Gao, L. Minimizing energy consumption for robot arm movement. *Procedia Cirp* **25**, 400–405 (2014).

[46] Shuvo, M. B., Ahommed, R., Reza, S. & Hashem, M. Cnl-unet: A novel lightweight deep learning architecture for multimodal biomedical image segmentation with false output suppression. *Biomedical Signal Processing and Control* **70**, 102959 (2021).

[47] Liu, D., Zhao, F.-Y. & Tang, G.-F. Active low-grade energy recovery potential for building energy conservation. *Renewable and Sustainable Energy Reviews* **14**, 2736–2747 (2010).

[48] Gullbrand, J., Luckeroth, M. J., Sprenger, M. E. & Winkel, C. Liquid cooling of compute system. *Journal of Electronic Packaging* **141**, 010802 (2019).

[49] Pambudi, N. A. *et al.* Preliminary experimental of gpu immersion-cooling **93**, 03003 (2019).

[50] Pambudi, N. A. *et al.* The immersion cooling technology: Current and future development in energy saving. *Alexandria Engineering Journal* **61**, 9509–9527 (2022).

[51] Zhang, H., Shao, S., Xu, H., Zou, H. & Tian, C. Free cooling of data centers: A review. *Renewable and sustainable energy reviews* **35**, 171–182 (2014).

[52] Zhang, Y., Wei, Z. & Zhang, M. Free cooling technologies for data centers: energy saving mechanism and applications. *Energy Procedia* **143**, 410–415 (2017).

[53] Le Sueur, E. & Heiser, G. Dynamic voltage and frequency scaling: The laws of diminishing returns 1–8 (2010).

[54] Kumar, R., Khatri, S. K. & Diván, M. J. Power usage efficiency (pue) optimization with counterpointing machine learning techniques for data center temperatures. *International Journal of Mathematical, Engineering and Management Sciences* **6**, 1594 (2021).

[55] Mukherjee, D., Chakraborty, S., Sarkar, I., Ghosh, A. & Roy, S. A detailed study on data centre energy efficiency and efficient cooling techniques. *International Journal* **9** (2020).

[56] Helali, L. & Omri, M. N. A survey of data center consolidation in cloud computing systems. *Computer Science Review* **39**, 100366 (2021).