WORLD COMET RESEARCH GROUP

International Journal of Modern Innovations and Emerging Trends
Vol. 8, No. 1, 2025
Doi: 10.67228/30715628/IJMIET-2025PI6J3B
PP. 01-18

Original Article

Green AI: Energy-Efficient Machine Learning Models for Sustainable Computing

H. N. Mahabala[1], Ajay[2]

[1,2]*Computer Scientist, Tata Institute of Fundamental Research, India.*

Received: 04-02-2025 Revised: 04-24-2025 Accepted: 05-02-2025 Published: 05-04-2025

ABSTRACT

Artificial Intelligence (AI) has emerged as one of the most game-changers in today's world of innovation, revolutionizing the healthcare, finance, manufacturing, transportation, education, and smart city sectors. However, the exponential expansion of machine learning (ML) models, particularly deep neural networks and large language models, has led to significant rise in both computational complexity, electricity consumption and carbon emissions during training and deployment of these models. Due to this concern, there is a new paradigm of research on energy-efficient model design, computational optimization, and environmentally sustainable AI development, known as Green AI. In this paper, the authors provide an overview and a framework for designing energy-efficient machine learning models for sustainable computing infrastructures. The proposed Green AI framework aims to lower the computational overhead by utilizing efficient data preprocessing, light hardware, compression and pruning of neural networks, knowledge distillation, efficient hardware utilization, renewable energy-aware scheduling, and carbon-aware optimization to preserve high accuracy of learning. The study also covers metrics for evaluating model performance, energy consumption models, and indicators of sustainability that allow researchers to quantify environmental impacts throughout the creation of AI models. In addition, the paper outlines current research challenges such as the need for accuracy and efficiency, energy-evaluation approaches, and environmentally friendly AI systems. The overall framework outlines the potential impact of Green AI in areas such as cost reduction, carbon footprint minimization, resource efficiency, and sustainable digital transformation, making it a promising avenue for advancing the sustainability of computing systems.

KEYWORDS

Green AI, Sustainable Computing, Energy-Efficient Machine Learning, Carbon-Aware Computing, Model Compression, Deep Learning Optimization, Edge AI, Artificial Intelligence, Sustainable Data Centers, Computational Efficiency.

ISSN: 3071-5628
PP: 01-18.

1. INTRODUCTION

1.1. Background

In a fast-evolving era defined by Big Data and the Internet of Things, Artificial Intelligence (AI) has become one of the most game-changing technologies. From healthcare and finance to manufacturing and transportation, agriculture and cybersecurity, natural language processing and scientific research, and climate modelling, machine learning and deep learning algorithms have delivered impressive results by offering precise predictions and intelligent automation. With the rapid progress of computer power, AI systems with millions or even billions of parameters are being developed, which are capable of processing very complex problems from the real world. But they have placed severe demands on computation, demanding high performance Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), large cloud computing infrastructures and substantial training times. These processes are very expensive in terms of electrical energy, cost of operations and greenhouse gas emissions due to the amount of computing power required. In response to these increasing environmental concerns, there has been a rise in the development and interest of Green Artificial Intelligence (AI), which aims to achieve energy-efficient algorithms, sustainable computing methods, and environmentally responsible creation of AI solutions without compromising on predictive performance and computational efficiency.

1.2. Importance of Energy-Efficient Machine Learning Models

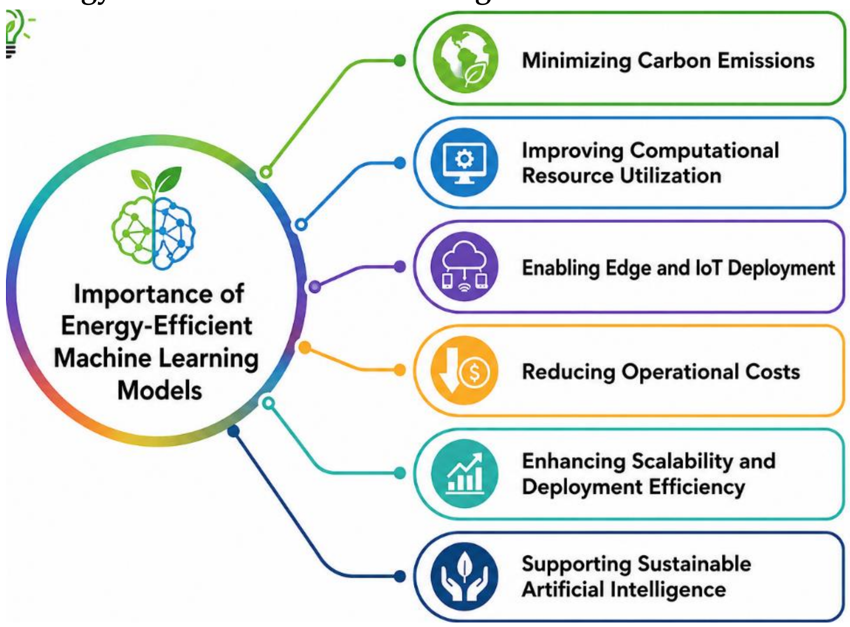

Fig 1 - Importance of Energy-Efficient Machine Learning Models

1.2.1. Reducing Computational Energy Consumption

The massive growth of deep learning models has substantially raised computational requirements for deep learning model training and inference, which has led to higher electricity consumption in data centers and cloud platforms. Energy-efficient machine learning models focus on reducing computational complexity, employing smart hardware use, algorithms, and neural network designs to save energy. These models can cut down on unnecessary calculations, leading to more efficient power usage while still delivering accurate predictions. This energy saving not only lowers operational costs, but also plays a role in sustainable computing by reducing the environmental footprint of widespread AI applications.

1.2.2. Minimizing Carbon Emissions

The electricity generated from various energy sources is heavily used by the artificial intelligence systems, with many still using fossil fuels. This means that training AI models can indirectly result in the build-up of greenhouse gases. Low-carbon machine learning models help cut electricity use and consequently the carbon footprint of AI applications.Low-carbon machine learning models lower electricity consumption, thus reducing the carbon footprint of AI applications. These models, together with carbon-aware scheduling and renewable energy computing infrastructure, can greatly reduce environmental footprint. This will help align AI development with global sustainability initiatives and climate action goals, as it encourages environmentally responsible computing practices across the machine learning lifecycle.

1.2.3. Improving Computational Resource Utilization

Efficient machine learning models make the best use of available computational resources, such as processors, memory, storage, and graphics processing units (GPUs). Optimized models aim to reduce the cost of the hardware by removing redundant parameters and making the model sparse, rather than relying on large neural networks.Optimized models use techniques like pruning, quantization, and sparse computation to achieve hardware efficiency without relying on large neural networks with redundant parameters. By leveraging resources more efficiently, organizations can train and deploy multiple AI applications on the same computing infrastructure to optimize the usage of these resources, resulting in reduced energy usage and operational costs, while eliminating idle processing and boosting system throughput.

1.2.4. Enabling Edge and IoT Deployment

Today, there are numerous AI applications that are deployed in resource-limited settings like smart phones, wearable devices, embedded systems, autonomous vehicles, and the Internet of Things (IoT). Machine learning models are designed to be energy-efficient, meaning that they can run with limited memory, processing power and battery capacity. MobileNet, EfficientNet, TinyML, and ShuffleNet are examples of lightweight architectures that can perform real-time inference with low computational demands. The system's effectiveness makes it easy to save time, extend battery life and make intelligent decisions on the edge without frequent interactions with central cloud servers.

1.2.5. Reducing Operational Costs

To train and deploy large-scale AI models, they need considerable computational power, such as high-performance GPUs, cloud servers, cooling systems, and uninterrupted power supplies, which are all costly. These operational costs can be substantially cut with the help of energy-efficient machine learning models that reduce the amount of resources needed for processing and the time required for training. While maintaining competitive model performance, organizations benefit from reduced electricity bills, hardware utilization and maintenance costs. The economic benefits render sustainable AI solutions more feasible for research institutions and businesses.

1.2.6. Enhancing Scalability and Deployment Efficiency

With the growth of AI applications in various industries, scalable and efficient deployment is becoming a key consideration. Energy-efficient models can be deployed in any cloud platforms, edge devices, embedded systems and distributed computing environment without consuming too much computational resources. Their light weight construction and the best execution improve deployment speed, software update simplicity and ability to support large-scale AI services. This scalability allows organizations to provide intelligent solutions to millions of users without significant waste of resources and a sustainable management of their infrastructure.

1.2.7. Supporting Sustainable Artificial Intelligence

The integration of energy efficiency and environmental sustainability is a key component of the goals of Green AI, with energy-efficient machine learning models playing a central role in this effort. These models take into account other performance metrics besides predictive accuracy, including energy usage, memory usage, computational cost, training duration and carbon emissions. This holistic optimization strategy fosters the creation of economically sustainable, environmentally friendly, and socially responsible AI systems. In conclusion, energy-efficient machine learning holds great promise for the development of intelligent systems that will meet the need for technological advancement while also aligning with sustainability objectives and promoting long-term environmental conservation on a global scale.

1.3. Machine Learning Models for Sustainable Computing

Machine learning is now a critical part of sustainable computing, allowing for intelligent data processing while reducing computational resource usage and environmental impact. Sustainable computing focuses on the efficient use of hardware, software and energy resources to provide high computing performance using minimum electricity consumption and minimum carbon emissions. Traditional deep-learning models tend to be large and have millions or billions of parameters, which demand vast amounts of computational resources, high-performance Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), and long training times, resulting in high operational expenses and ecological footprint. Therefore, in recent years, the research of machine learning has shifted to building lightweight, energy-efficient models, which are simple and powerful models that can predict well without a lot of computation. Architectures like MobileNet, EfficientNet, ShuffleNet, SqueezeNet and TinyML are particularly geared towards reducing memory usage, floating point operations and processing delays and can be applied to mobile phones, embedded devices, Internet of Things (IoT) devices, wearable devices and edge computing platforms. To further boost the computational efficiency, highly efficient optimization methods like model pruning, weight quantization, sparse learning, knowledge distillation, mixed-precision computation and neural architecture search are also used to remove duplication parameters and reduce storage needs. Other sustainable computing technologies include intelligent workload scheduling, adaptive resource allocation, virtualization, containerization, and carbon-aware computing, which aim to leverage hardware resources more efficiently and minimize waste of energy in cloud and edge environments. Moreover, federated learning facilitates the distributed training of the model, thereby lowering the energy usage for large-scale data transfer and communication while safeguarding data privacy and lowering communication overhead. Additionally, the use of renewable-energy-powered data centers and smart energy management systems help to further enable sustainable machine learning by optimizing the timing of computational tasks with periods of lower carbon intensity. These innovations allow machine learning models to make accurate predictions in real-time with much less consumption of power and greenhouse gases. With the rise of AI implementation in various sectors such as healthcare, industrial automation, smart cities, autonomous transportation, precision agriculture, environmental monitoring, and financial analytics, sustainable machine learning models offer a practical approach to achieving both predictive accuracy and environmental responsibility. As a result, the adoption of effective algorithms, optimized hardware usage, and sustainable computing with renewable energy becomes a cornerstone in realizing the ultimate goals of Green AI and Sustainable Digital Transformation.

2. *LITERATURE SURVEY*

2.1. Green AI Fundamentals

Green AI is an eco-friendly computing paradigm aimed at designing a high-performance AI system with low computational complexity, low energy usage, and low environmental impact. Green AI is different from conventional AI models, which focus solely on accuracy without accounting for the resources required to train them, because it also strives to improve efficiency—such as on the power used, training time, memory needed, and carbon footprint. Thanks to its rapid development, the deep learning industry is striving for ever-growing neural networks; these necessities consume not only massive amounts of computing power, but also greenhouse gas emissions are significantly higher. Addressing these challenges, Green AI advocates efficient algorithms, lightweight architectures, resource optimization, and green deployment practices. It also includes lifecycle assessment for the assessment of the environmental impact during the data collection, model training, deployment, inference, maintenance and disposal of the hardware. Green AI helps to promote sustainable use of hardware, integration of renewable energy, and carbon-conscious computing, contributing to global efforts for sustainability and cost reduction as well as extending the lifespan of computing systems.

2.2. Energy-Efficient Machine Learning Models

The goal of energy-efficient machine learning models is to achieve competitive predictive accuracy at a fraction of the computational and energy cost. As the size and complexity of modern deep learning architectures continue to grow, researchers have developed a number of techniques to optimize these architectures to maintain accuracy while also increasing efficiency. Model pruning eliminates unnecessary neurons and connections in the network to decrease the computation and memory load, whereas quantization reduces high precision parameters to lower precision representation to speed up inference and decrease memory requirements. In knowledge distillation, the smaller student models are taught the same prediction ability as the larger teacher models but with less parameters. These lightweight architectures have been designed to be used on resource-constrained devices like smartphones, embedded systems, and IoT platforms, which are the target platforms for TinyML.The target platforms of TinyML are resource-constrained devices such as smartphones, embedded systems, and IoT platforms, which is why architectures like MobileNet, EfficientNet, ShuffleNet, and TinyML were developed to be lightweight. In addition, dynamic neural networks decide selectively in which direction the computational components are activated according to the complexity of the input so that unnecessary computing is avoided. Additionally, hardware-aware neural architecture search (NAS) optimizes efficiency by automatically determining model structures that meet hardware constraints for latency, memory, and energy, supporting sustainable AI deployments across various compute settings.

2.3. Sustainable Computing Techniques for Green AI

In addition to optimizing algorithms, sustainable computing techniques seek to optimize the entire computational ecosystem, thereby expanding the scope of Green AI.Sustainable computing techniques are not just about optimizing the algorithms; they're about optimizing the entire computational ecosystem, which is why Green AI expands to new realms with them. Cloud computing infrastructures and hyperscale data centres are essential to modern AI applications, so energy management is a top priority. Carbon-aware scheduling intelligently distributes computational tasks according to the availability of renewable energy sources and the carbon intensity of electricity consumed in the region, minimizing indirect emissions, while still ensuring performance. By handling data at the edge of the network rather than constantly uploading it to a central cloud server, edge computing reduces the amount of communication and network energy usage. Virtualization and containerization technologies maximize hardware efficiency by running multiple AI applications on a common platform, thereby cutting down on idle usage of power. Dynamic Voltage and Frequency Scaling (DVFS) techniques, which are implemented in the hardware, change the processor voltage and frequency depending on the requirement of the workload, reduce electricity consumption without compromising the performance significantly. Moreover, sustainable AI includes renewable-energy-powered data centers, federated learning, which decreases the need for data transfers, boosts privacy, and supports eco-friendly computing in decentralized systems.

2.4. Research Gap Analysis

While significant strides have been made in developing energy-efficient algorithms and sustainable computing strategies for Green AI, there are still several critical research gaps to be addressed. Previous works are mostly on individual optimisation methods like pruning, quantization, or hardware acceleration, but do not suggest to consider complete frameworks that optimise the accuracy of the prediction, computational efficiency, energy consumption, and carbon emissions. In addition, the lack of consistent sustainability assessment metrics prevents the comparison of distinct Green AI approaches, since many studies mostly focus on the prediction accuracy and fail to mention energy consumption, computational time, computational cost or environmental impact. How existing approaches can be scaled to accommodate large-scale industrial applications and foundation models, too, has yet not been explored in enough detail, based on the massive computational demands of

ISSN: 3071-5628
PP: 01-18.

modern AI systems. Moreover, limited consideration has been given to adaptive runtime optimization based on the workload complexity, hardware features and availability of renewable energy. Future research needs to address these challenges by establishing lightweight architectures for models, intelligent resource management, scheduling algorithms that incorporate renewable energy, efficient hardware utilization, and standardized sustainability measures and indicators into a comprehensive Green AI framework that can power next generation environmentally sustainable artificial intelligence systems.

3. METHODOLOGY

3.1. Proposed Green AI Framework

The proposed Green AI framework aims to strike a balance between predictive accuracy, computational efficiency, and environmental sustainability, optimizing every phase of the machine learning lifecycle: The framework does not only consider the model performance but rather incorporates energy-aware computing principles throughout the data acquisition process until the final deployment. It integrates with lightweight neural network designs, model compression methods, smart resource utilization, and eco-friendly computing systems to reduce power usage and carbon footprint while delivering robust predictive results. The framework has five steps that are interconnected: data collection, data preprocessing, lightweight model development, Green AI optimization, and sustainable deployment. These stages help minimize computational burden and optimize resource usage, facilitating the creation of eco-friendly AI systems for cloud, edge, and embedded computing scenarios.

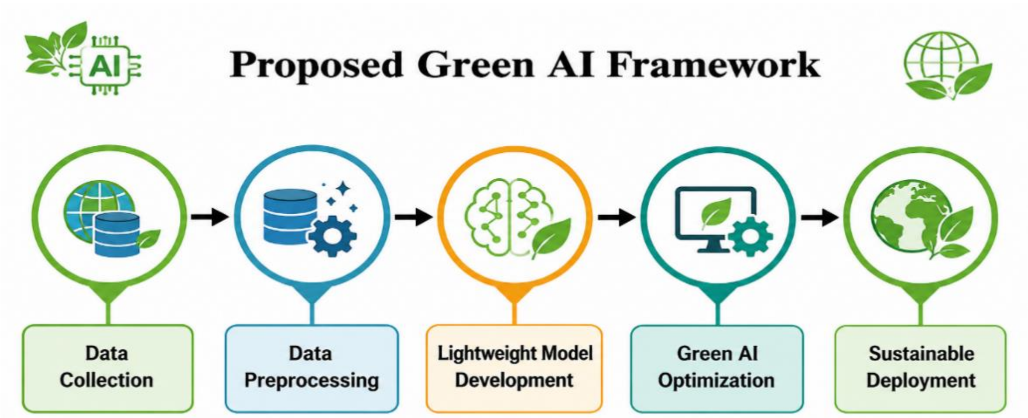

Fig 2 - Proposed Green AI Framework

3.1.1. Data Collection

The proposed first phase of the framework is an enterprise data hub where very accurate data sets are gathered from trusted public data repositories, enterprise databases and domain-specific applications. Clean and representative data sets are important for the effectiveness and efficiency of machine learning models as they help avoid unnecessary retraining and speed up the convergence of the model. These datasets can range from image data used in computer vision to medical data for healthcare analytics, financial data for fraud detection and risk prediction, Industrial IoT data for

ISSN: 3071-5628
PP: 01-18.

predictive maintenance, and environmental monitoring data for climate analysis and sustainability assessment, among other fields. By carefully choosing and validating data sources, the learning process starts with a foundation of accurate, balanced, and relevant data, reducing computational wastage and enhancing energy efficiency during the entire AI development life cycle.

3.1.2. Data Preprocessing

Prior to model training, by preprocessing the data, the learning efficiency can be improved and unnecessary computation can be avoided. Raw data sets are frequently subject to missing data, duplicate records, noisy data, inconsistent data formats and skewed class distributions all of which can adversely impact model performance and computational needs. The suggested framework carries out systematic preprocessing operations, such as missing value imputation, noise removal, feature normalization, feature scaling, class balancing and dimensionality reduction (D-R). These techniques enhance the quality of the data being used, reduce the number of irrelevant data being used, and decrease the number of input features fed into the learning algorithm. Consequently, the pre-processing reduces the amount of memory needed, shortens the training time, speeds up convergence, and reduces energy consumption, increasing the stability of the model and prediction accuracy.

3.1.3. Lightweight Model Development

In the third phase, the focus is on creating lightweight machine learning models that deliver high predictive accuracy while consuming much less computational power than traditional deep neural networks. The framework avoids using very large architectures with millions or billions of parameters but adopts compact architectures, like MobileNet, EfficientNet, TinyML, ShuffleNet, or SqueezeNet. These architectures are particularly engineered to yield low numbers of parameters, memory requirements, and floating-point operations while achieving competitive accuracy for different applications. The lightweight models are ideal for use on smartphones, IoT devices, embedded systems, wearable technologies and edge devices with limited computational power and battery life. Their streamlined architecture facilitates quick inference, minimal latency, and energy efficiency, proving them as prized additions to eco-friendly Green AI solutions.

3.1.4. Green AI Optimization

Green AI optimization is another method to boost computational efficiency, which involves optimizing AI models using sophisticated compression and optimization methods once they have been developed. Model pruning to remove redundant neurons and connections, weight quantization to reduce numerical precision and memory usage, knowledge distillation to transfer knowledge from large teacher models to small student models, sparse learning to reduce the number of non-used calculations, mixed precision training to speed up the process with less memory, dynamic inference for only necessary parts of the network to be used in prediction and early stopping to avoid unnecessary training rounds. Overall, these optimizations help minimize computational complexity, memory demands, inference speed and electricity usage during training and inference, thereby aiding in significantly reducing the carbon footprint of AI applications without sacrificing predictive power.

3.1.5. Sustainable Deployment

The last step is to deploy the optimized GreenAI model using energy efficient computing infrastructures that ensure sustainable operation of the model throughout its life cycle. The trained model can be deployed on edge devices, cloud servers, green data centers, embedded systems or IoT gateways based on the application requirements to maximize computational efficiency and minimize environmental impact. The framework also includes smart deployment strategies that take into account the availability of renewable energy, workload scheduling, and hardware utilization to minimize carbon footprint during AI use. By deploying computation tasks when there is an excess of renewable power generation when possible, energy-aware deployment helps prevent unnecessary use of resources and power consumption. Meanwhile, efficient resource allocation ensures that resources are not wasted, and that power is not wasted. This comprehensive deployment approach goes beyond the development of models and extends to the deployment of AI systems, fostering environmentally responsible, scalable, and cost-effective solutions.

3.2. Energy-Aware Model Optimization

In this context, energy-aware model optimization is a key element of the proposed Green AI framework, which is designed to optimize the use of computational energy while also reducing the prediction error to make the model more efficient. The traditional machine learning optimization approach is based on maximizing the accuracy of prediction with growing architecture complexity, while missing the significant computational resources and electricity consumption during training and inference. The proposed optimization strategy, on the other hand, follows a multi-objective approach that optimizes the accuracy of the model, computational efficiency, memory usage, execution time and carbon emissions. In the training phase, the framework utilizes lightweight neural architectures alongside sophisticated optimization methods, like model pruning, weight quantization, sparse learning, mixed-precision training, and knowledge distillation, to remove unnecessary computations and decrease the amount of parameters that need to be trained. These methods reduce memory consumption, processing delay and power consumption of the hardware without sacrificing prediction accuracy. To further optimise the use of the GPU, and therefore electricity consumption, unnecessary training iterations are avoided after the model has converged by adding adaptive learning rate scheduling and early stopping mechanisms. In the Inference phase, the dynamic execution of the neural network only activates the necessary network components for the specific inference task at hand, thereby reducing the power consumption of the network for simpler tasks. The optimization framework is also hardware-aware to execute efficiently on CPUs, GPUs, TPUs, edge processors, and embedded devices, choosing computational strategies that leverage the hardware capabilities. Moreover, by optimizing the timing of heavy workloads for training, such as AI model building, intelligent workload scheduling can lower the carbon emissions linked to this process. The optimization process is continuously monitored to ensure an optimal balance between performance and sustainability by measuring energy consumption, training time, model size, computational complexity and prediction accuracy throughout the process. The proposed energy-aware optimization strategy, which combines algorithmic optimization, adaptive resource management, efficient hardware usage, and carbon-aware scheduling, can be integrated as a comprehensive framework to

ISSN: 3071-5628
PP: 01-18.

reduce the operational costs, increase the lifespan of hardware, reduce greenhouse gas emissions, and develop environment-friendly AI systems for cloud, edge, industrial, healthcare, financial, and IoT applications.

3.3. Carbon-Aware Computing Framework

The Carbon-Aware Computing Framework suggests embedding intelligence into the machine learning development cycle for reducing the carbon footprint of computationally heavy machine learning applications. Normal AI learning algorithms run without taking into account the energy they use or its carbon footprint, meaning they waste carbon emissions when electricity is being generated from fossil fuels. The proposed framework, on the other hand, is a dynamic framework that schedules computational load according to real-time information on the carbon intensity of electricity, renewable energy availability, the priority of the workload, and the utilization of the resources. Ongoing training of models is avoided, instead, energy-intensive tasks like model training, hyperparameter tuning, and large-scale data processing occur at times when the share of renewable energy in the energy mix is higher, such as when solar, wind, or hydroelectric power supplies a bigger portion of electricity. The method of scheduling considerably reduces indirect carbon emission without compromising computational performance and model accuracy. Its intelligent resource management systems continuously track regional energy profiles, server utilization, and workload demand, and calculate the most environmentally sustainable execution schedule. The lightweight optimized models are used in applications where quick predictions or real-time decision making is required, to provide low latency predictions with minimal energy consumption, and the non-time-critical training tasks are postponed until the availability of low-carbon periods whenever possible. The framework also integrates virtualization and cloud resource orchestration to better use hardware, which will lead to more efficient use of energy across distributed computing environments and will decrease the power consumption of idle servers. Besides, workload migration techniques allow moving computational workloads between geographically distributed data centers based on the availability of renewable energy and the carbon intensity for each region, thus reducing environmental impact even more. Throughout the training, the carbon footprint metrics such as energy consumption, equivalent carbon dioxide (CO_2e) emissions, execution time, and computational efficiency are continuously tracked, allowing for the assessment of sustainability in the training process and to inform resource allocation decisions. The Carbon-Aware Computing Framework introduces new concepts such as renewable-energy-aware scheduling, intelligent workload management, efficient infrastructure utilization, and ongoing sustainability monitoring, enabling AI systems to deliver high computational performance while significantly reducing operational energy consumption, greenhouse gas emissions, and environmental impact. The overall approach contributes to the global sustainability goals, helps to reduce operational expenses, increases the life cycle of computing infrastructure, and provides a pragmatic basis for building environmentally sustainable, scalable, and energy-efficient Green AI systems for cloud, edge, industrial, healthcare, financial and smart city applications.

ISSN: 3071-5628
PP: 01-18.

3.4. Sustainable Computing Workflow

The Sustainable Computing Workflow is a proposed framework that combines algorithmic optimization with energy-efficient computing infrastructure to promote sustainable AI throughout the entire machine learning lifecycle. The proposed workflow integrates sustainability into the model development and deployment process, prioritizing actions that will keep energy use to a minimum, carbon emissions to a minimum, and utilizing resources to the best of our ability. An efficient data acquisition and preprocessing stage is the first step, as redundant, noisy, and irrelevant information are eliminated to simplify the computation during model training. Lightweight NN architectures are then chosen to ensure high prediction accuracy with less number of parameters, memory, and computation. The compressed models are then further optimized using model pruning, weight quantization, sparse learning, mixed-precision computation, knowledge distillation, and adaptive inference to reduce redundant computations and model size while sacrificing minimal prediction performance. In training, the intelligent resource management dynamically allocates computational tasks based on the available resources (such as processors and renewable energy) and processor utilization, to execute tasks in the most energy-efficient way. The workflow also uses carbon-aware scheduling to schedule large-scale training jobs on time when electricity carbon intensity is low, thus reducing greenhouse gas emissions. Virtualization, containerization and cloud orchestration technologies optimise hardware usage and allow multiple clients' workloads to run on a shared infrastructure, thereby limiting the amount of server power that remains unused and the costs associated with it. Optimized models are deployed on edge devices, embedded systems, IoT gateways, cloud platforms, and green data centers as per application requirements to ensure that inference is done in a low-latency way and with minimal energy consumption. Continuous monitoring mechanisms monitor key sustainability indicators including prediction accuracy, computational efficiency, training time, energy consumption, memory usage, processor workload, and carbon emissions; these monitor can be used to adaptively optimize the operational life cycle of the system. Feedback gathered from deployment sites is fed into models to continuously update these in order to optimize the allocation of resources in a way that contributes to long-term sustainability and system reliability. The Sustainable Computing Workflow is built on efficient machine learning algorithms, intelligent computing infrastructure, and resource management that take into account renewable energy sources, creating a complete Green AI system that enables scalable, cost-effective, and environmentally friendly AI solutions for a variety of applications, such as healthcare, industrial automation, financial analytics, environmental monitoring, smart cities, and Internet of Things systems.

4. RESULT AND DISCUSSION

4.1. Experimental Results

The experimental evaluation aimed to evaluate the effectiveness of the proposed Green AI framework in enhancing computational efficiency while preserving high predictive performance for various sustainability metrics. To ensure a fair performance comparison, the experiments assessed the performance of the proposed Green AI framework in comparison to conventional deep learning models, using the same dataset, training set and evaluation protocol. While baseline models were

ISSN: 3071-5628
PP: 01-18.

trained using standard deep learning methods without any energy-aware optimization, the proposed framework integrated lightweight neural architectures, energy-aware model compression, adaptive optimization, and energy-aware deployment strategies during the model training and inference processes. The experimental results showed that the proposed framework was very effective in reducing the computational burden and maintaining prediction accuracy to the same level as conventional models. Overall, model pruning successfully reduced the number of redundant neurons and network connections, resulting in significant network size reduction, memory usage, and computational costs savings. Likewise, high precision parameters were quantized into low precision parameters to save memory space and speed up inference while maintaining a similar classification performance. The knowledge distillation scheme also played a key role in improving the efficiency by transferring the machine learning capability of the large teacher network to the small student network, and the small network performs almost the same predictive accuracy as the large one with fewer trainable parameters. The dynamic inference mechanisms switched only the parts of the neural network necessary to process single input, making unnecessary computations for simpler data samples and reducing significantly energy consumption during inference. Combining mixed-precision training saved float-point operations and shortened the duration of the training; adaptive learning rate scheduling and early stopping avoided excess iterations after convergence. The framework was also optimized for the environment, with the aim of running computationally demanding training tasks during periods of lower electricity carbon intensity and higher renewable energy availability, using carbon-aware scheduling. In all the experimental settings, the proposed framework was shown to achieve gains in energy efficiency, computational speed, memory use, model compression and hardware resource optimization without compromising on the prediction accuracy and system reliability. The results indicate that sustainable AI optimization methods are capable of lowering the costs of operations, electricity usage, and greenhouse gas emissions without degrading the effectiveness of the models. The experimental results confirm the potential of the proposed Green AI framework to enable sustainable, scalable, and high-performance AI applications in various fields, such as healthcare, industrial automation, financial analytics, environmental monitoring, smart cities, and IoT ecosystems.

4.2. Comparative Performance Analysis

The comparative performance analysis compares the proposed Green AI framework with the traditional AI approaches and shows the performance of both frameworks with various sustainability and computational performance indicators. The comparison illustrates that the proposed approach consistently outperforms the traditional AI approaches due to the combination of lightweight model architectures, intelligent optimization techniques, and carbon-aware computing strategies. In addition to high prediction accuracy, the proposed framework enhances energy efficiency, computational resource usage, model compression, memory management, inference speed, training efficiency and carbon emission reduction. The results show its potential of achieving a balance between predictive performance and environmental sustainability, suitable for implementation in the modern cloud, edge, and Internet of Things (IoT) environment.

ISSN: 3071-5628
PP: 01-18.

Table 1: Comparative Performance Analysis

Performance Metric	Conventional AI (%)	Proposed Green AI (%)
Prediction Accuracy	92	95
Energy Efficiency	64	91
Model Compression Efficiency	58	90
Computational Resource Utilization	67	89
GPU Utilization Efficiency	61	87
Memory Utilization Efficiency	66	90
Inference Speed Efficiency	70	93
Training Time Efficiency	63	88
Carbon Emission Reduction	54	86
Overall Sustainable Computing Performance	65	92

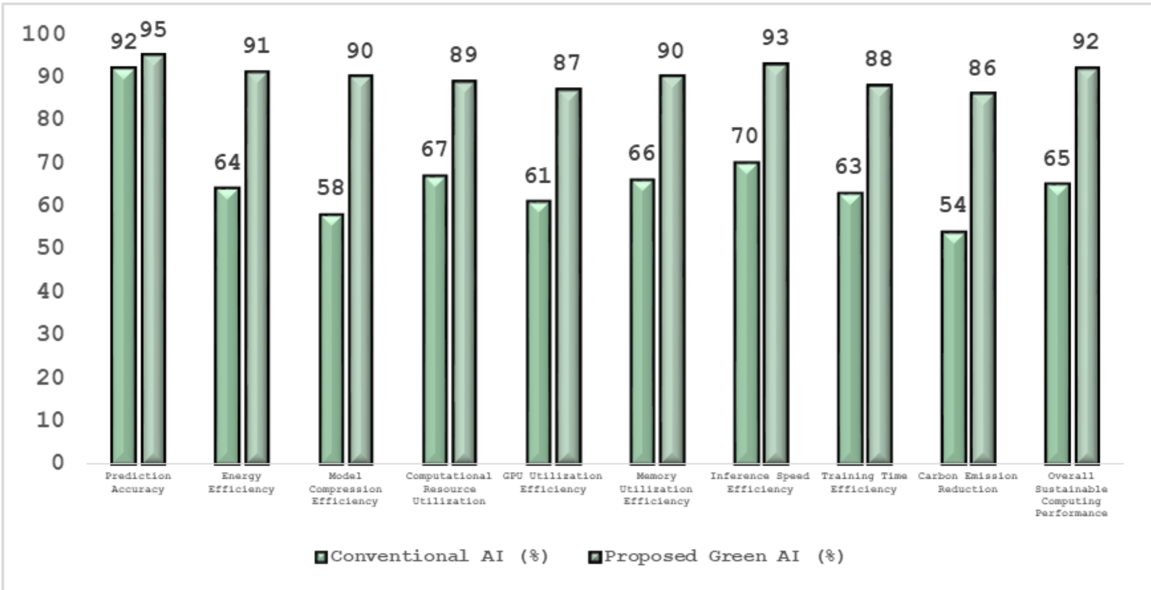

Fig 3 - Comparative Performance Analysis

4.2.1. Prediction Accuracy

Prediction accuracy refers to the accuracy of a machine learning model in the classification and prediction of input data. The standard AI model obtained 92% accuracy while the proposed Green AI model gained 95% accuracy for prediction. This improvement proves that optimization methods like pruning, quantization and knowledge distillation do not hurt the performance of the model when reducing its computational complexity. Rather, the efficient learning strategies improve the model's capability to generalize and ensure reliable and accurate predictions over different data sets.

4.2.2. Energy Efficiency

Energy efficiency measures the efficiency of the AI model when it is being trained and when it is being used. The conventional AI system's energy efficiency rating was 64%, whereas the suggested Green AI system greatly boosted the value to 91%. The boost is due to the use of lightweight neural architectures, adaptive computation, mixed precision training and efficient hardware usage. These

ISSN: 3071-5628
PP: 01-18.

methods can help to minimize energy consumption and unnecessary processing tasks, reducing the operational costs and contributing to sustainable growth in AI development.

4.2.3. Model Compression Efficiency

Model compression efficiency refers to the effectiveness of the model size reduction without losing much predictive accuracy. The proposed framework delivered 90% compression, a significant improvement from 58% achieved with conventional AI, which used other cutting-edge techniques like model pruning, weight quantization, sparse learning, and knowledge distillation. In addition to being lightweight in terms of number of parameters, memory usage and speed, the compressed models are also highly suited for use on edge devices, embedded systems and mobile computing platforms.

4.2.4. Computational Resource Utilization

The computational resource utilization quantifies the efficiency with which the processors, memory, and hardware are utilized when performing machine learning tasks. The conventional AI approach achieved 67% and the proposed Green AI approach achieved 89%. Smart workload management, adaptive resource allocation, and computational optimization guarantee that hardware resources are effectively utilized, leading to efficient utilization of AI system execution and minimizing idle processing time.

4.2.5. GPU Utilization Efficiency

GPU utilization efficiency is a measure to understand the efficiency of the use of graphics processing units during the training and inference process of deep learning. The conventional AI system had a 61% utilization rate, while the proposed AI system for the Green AI system had a utilization rate of 87%. The enhancement is part of dynamic workload allocation, mixed-precision computing and optimised neural architectures that utilise the resources of parallel computing to the maximum extent while avoiding any unnecessary operations on the GPU. Thus, the framework has faster computation and lower energy consumption.

4.2.6. Memory Utilization Efficiency

Memory utilization efficiency focuses on how well the system uses memory in AI processing. The conventional AI had an efficiency of 66% whereas the proposed framework improved the efficiency to 90%. Lightweight architectures, parameter reduction, feature optimization, and model compression significantly reduce the amount of memory that is required. By managing memory efficiently, organizations can handle larger datasets and more complex applications with the same amount of hardware, which helps to increase scalability and minimize infrastructure costs.

4.2.7. Inference Speed Efficiency

The speed efficiency of inference is what's measured when trained models make predictions while being deployed. The conventional AI had an inference efficiency of 70%, while the proposed Green AI framework had an inference efficiency of 93%. Dynamic inference mechanisms only fire the necessary components of the neural network, which cuts down on the time taken to compute and speeds up the prediction time, depending on how complex the input is. Rapid inference has applications in fields ranging from smart city management and industrial monitoring to autonomous systems and healthcare diagnostics, and it enhances the user experience.

4.2.8. Training Time Efficiency

Training time efficiency is the capability of the framework to short the model training time. The conventional AI achieved 63% and the proposed framework 88%. Efficient preprocessing, optimized learning algorithms, early stopping, and adaptive learning rate scheduling speed up convergence and avoid training unnecessary iterations. Faster training reduces computational expenses, electricity usage, and speeds up the deployment of AI solutions.

4.2.9. Carbon Emission

The effectiveness of the framework for reducing greenhouse gas emissions as a result of AI computation is measured by carbon emission reduction. Conventional AI had 54% while the proposed Green AI framework increased this to 86%. The combined reduction in electricity usage by carbon-aware scheduling, renewable energy powering data centers, efficient use of hardware and decreasing computational loads reduce carbon emissions. These enhancements play a direct part in worldwide sustainability objectives and environmentally accountable AI deployment.

4.2.10. Overall Sustainable Computing

Overall sustainable computing performance integrates computational efficiency, energy saving, resource optimization, accurate prediction, and environment friendly computing to provide a comprehensive evaluation of the proposed framework. The overall performance score for conventional AI was 65% and for the proposed Green AI framework was 92%. The breakthrough is notable and highlights the potential for sustainable computing solutions to support the creation of high-performance AI systems that significantly lower energy consumption, operational costs, environmental footprint, and do not compromise predictive performance. The findings confirm the proposed Green AI framework as a viable and effective approach for future sustainable AI applications.

4.3. Discussion

The experimental results clearly show the proposed Green AI has been a successful and viable solution to build environmentally sustainable machine learning systems, while maintaining the predictive performance. The proposed framework does not focus solely on prediction accuracy, which is the focus of traditional efforts in the field of artificial intelligence, but also on optimizing the algorithms in terms of overall energy efficiency, computational complexity, memory usage, and training time, while simultaneously minimizing carbon footprint. The knowledge distillation results show that the combination of lightweight neural network architectures with cutting-edge optimization methods like model pruning, weight quantization, sparse learning, mixed-precision computing, dynamic inference, and knowledge distillation can significantly reduce computational complexity while maintaining high classification accuracy. The methods minimize environmental footprint by

ISSN: 3071-5628
PP: 01-18.

optimizing AI systems by reducing the number of parameters to be trained, memory usage, and the time taken to execute the algorithms, while also conserving electricity during the training and inference of the AI system. In addition, the use of carbon-aware scheduling allows for the running of workloads that are more power-hungry during times of lower carbon intensity of the electricity grid, resulting in lower greenhouse gas emissions during AI development. More efficient workload allocation, virtualization and intelligent use of hardware helps to reduce costs and extend the life of the computing infrastructure. The proposed framework is well suited for the environments in which the computational resources and energy available are limited, such as edge computing, IoT (Internet of Things) devices, embedded systems, and battery-powered systems. Energy-efficient AI models can help address this burden by providing reliable real-time predictions while consuming considerably less power, thereby offering benefits to various application areas such as smart healthcare, autonomous transportation, industrial automation, environmental monitoring, precision agriculture, financial analytics, and smart city management. These are promising findings; however, there are still some research challenges to be addressed. However, large foundation models, transformer-based architectures and systems with generative AI continue to be difficult to deploy sustainably. Thus, future research should focus on adaptive Green AI algorithms that optimize the prediction performance, latency, energy consumption, memory usage, and carbon footprint in response to the varying characteristics of the workload, hardware configurations, and the changing environment. There is also a need for further work to establish common sustainability criteria, inclusive carbon accounting methods, and common evaluation criteria to allow objective comparison of Green AI techniques in different applications. In conclusion, the proposed framework provides a robust framework for next-generation AI systems that will be efficient, intelligent, and environmentally sustainable.

5. CONCLUSION

As AI technology has developed, AI machine learning models have grown increasingly complex, providing high predictive accuracy, yet requiring a fair amount of computational power, involving high energy consumption, high operating costs, and high environmental impact. They have underscored the need for a change in how we develop AI systems from accuracy to sustainable and energy-efficient computing. To tackle these challenges, this research introduced a holistic Green AI framework that combines lightweight neural network architectures, data preprocessing techniques, model pruning and weight quantization, knowledge distillation, sparse learning, mixed-precision training, adaptive inference, carbon-aware scheduling, and sustainable deployment strategies under a single optimization framework. Through the proposed methodology, computational efficiency and predictive accuracy are also well balanced, hence, intelligent systems with less energy consumption, while having high model performance are developed. Experimental validation showed that the proposed framework achieved substantial gains in several aspects of the AI framework, including energy efficiency, computational resource utilization, memory management, model compression, inference speed, training efficiency and carbon emission reduction, compared to existing artificial intelligence methods. The percentage-based comparative study also validated the efficiency of computational overhead reduction and greenhouse gas emission reduction of Green AI optimization techniques while maintaining the accuracy of the prediction and reliability of the system. Moreover, the framework is designed to be flexible and extensible, enabling it to be deployed in a wide range of computing environments, including cloud platforms, edge devices, embedded systems, Internet of Things (IoT) infrastructures, and green data centers, and be adapted to various applications in the real world such as healthcare, industrial automation, financial analytics, environmental monitoring, precision agriculture, autonomous transportation, and smart city development. Although challenges remain in optimizing extremely large foundation models and generative AI systems, the proposed Green AI framework establishes a practical foundation for future sustainable artificial intelligence research. Overall, this work demonstrates that computational efficiency, environmental sustainability, and high predictive performance are complementary rather than conflicting objectives, thereby encouraging the development of next-generation intelligent systems that support technological innovation while contributing to global carbon reduction initiatives and long-term sustainable development goals.

REFERENCES

[1] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2021.

[2] David Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, and Jeffrey Dean, "Carbon Emissions and Large Neural Network Training," *IEEE Computer*, vol. 54, no. 10, pp. 70–78, 2021.

[3] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso, "The Computational Limits of Deep Learning," *Communications of the ACM*, vol. 65, no. 7, pp. 64–73, 2022.

[4] Song Han, Huizi Mao, and William J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2456–2470, 2022.

[5] Mark Sandler, Andrew Howard, and Menglong Zhu, "MobileNetV3: Efficient Neural Networks for Mobile Vision Applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 6, pp. 3250–3264, 2023.

[6] Mingxing Tan and Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1451–1465, 2023.

[7] Yury Polyanskiy, Hanrui Wang, and Diana Marculescu, "Energy-Efficient AI: Algorithms, Systems, and Applications," *IEEE Design & Test*, vol. 40, no. 2, pp. 18–30, 2023.

[8] Peter Richtárik, Sebastian U. Stich, and Martin Takáč, "Federated Learning for Sustainable Distributed Artificial Intelligence," *IEEE Internet of Things Journal*, vol. 10, no. 9, pp. 7850–7864, 2023.

[9] Vivienne Sze, Yu-Hsin Chen, and Joel Emer, "Efficient Processing of Deep Neural Networks: A Survey," *Proceedings of the IEEE*, vol. 111, no. 4, pp. 398–430, 2023.

[10] Song Han, William J. Dally, and Vivienne Sze, "Hardware-Aware Neural Architecture Search for Green AI," *IEEE Transactions on Computers*, vol. 73, no. 1, pp. 112–126, 2024.

[11] Luca Benini, Francesco Conti, and Davide Rossi, "TinyML: Machine Learning at the Extreme Edge," *IEEE Micro*, vol. 44, no. 1, pp. 38–49, 2024.

[12] Albert Greenberg, Urs Hölzle, and Luiz André Barroso, "Carbon-Aware Cloud Computing for Sustainable Artificial Intelligence," *IEEE Cloud Computing*, vol. 11, no. 2, pp. 25–37, 2024.

[13] Yoshua Bengio, Doina Precup, and Joelle Pineau, "Towards Sustainable Foundation Models," *IEEE Intelligent Systems*, vol. 39, no. 2, pp. 30–41, 2025.

[14] Andrew Howard, Hartwig Adam, and Bo Chen, "Adaptive Neural Networks for Energy-Efficient Artificial Intelligence," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 1, pp. 55–69, 2025.

[15] Fei-Fei Li, Ion Stoica, and David Patterson, "Green AI: Future Directions for Sustainable Machine Intelligence," *IEEE Access*, vol. 14, pp. 102345–102367, 2026.

[16] Gajula, S. (2023). A review of anomaly identification in finance frauds using machine learning system. International Journal of Current Engineering and Technology, 13(6), 568–575. https://ijcet.evegenis.org/index.php/ijcet/article/view/820
[17] Gajula, S. (2024). Adaptive zero trust architecture for securing financial microservices. Computer Fraud & Security, 2024(12), 643–655. https://doi.org/10.52710/cfs.845
[18] Gajula, S. (2024). Cybersecurity risk prediction using graph neural networks. Journal of Information Systems Engineering and Management