Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

# ENERGY-EFFICIENT OPTIMIZATION OF MACHINE LEARNING MODELS: BALANCING PERFORMANCE AND SUSTAINABILITY

**Nisha Yadav* and Pinky Laxmikant Sasmal**

*Department of Information Technology,*

*Pillai College of Arts, Commerce & Science (Empowered Autonomous), Navi Mumbai, Maharashtra, India 410206*

**Corresponding author E-mail: nishayadav@mes.ac.in*

| Received: 31 January 2026 | Revised: 28 February 2026 | Accepted: 20 March 2026 | Published: 31 March 2026 |
|---|---|---|---|

***DOI:*** **https://doi.org/10.5281/zenodo.19562639**

**Abstract:**

*Artificial Intelligence (AI) technologies have expanded rapidly and now play an important role in many fields including healthcare, transportation, and communication. Despite these benefits, the development of large machine learning models requires considerable computational power, which leads to high electricity consumption and increased carbon emissions. Modern deep learning models often contain billions of parameters, making their training process energy-intensive and environmentally costly. As a result, improving the energy efficiency of AI systems has become an important research priority. This research investigates several techniques that can reduce the energy consumption of machine learning models while preserving their predictive performance. The study focuses on optimization approaches such as model pruning, quantization, knowledge distillation, and energy-aware scheduling. These methods aim to reduce computational complexity, minimize memory usage, and lower the carbon footprint of AI training and deployment. The study evaluates widely used AI models including BERT, DistilBERT, ResNet-50, and MobileNet to examine the relationship between model performance and energy usage. In addition, emission monitoring tools such as eco2AI and ML $CO_2$ Impact are used to estimate energy consumption and carbon emissions. A sustainability evaluation metric called the Green Score is introduced to measure the balance between model accuracy and environmental impact. The findings show that applying optimization techniques can significantly reduce energy consumption and emissions while maintaining competitive accuracy levels. This research supports the concept of Green AI, encouraging the development of intelligent systems that are both technologically effective and environmentally responsible.*

**Keywords:** *Artificial Intelligence, Green AI, Energy Efficiency, Machine Learning, eco2AI.*

## 1. Introduction

Artificial Intelligence (AI) and deep learning technologies have experienced remarkable progress in recent years, enabling machines to perform complex tasks such as image recognition, language translation, recommendation systems, and medical analysis. Advanced models including BERT, GPT, ResNet, and BLOOM have achieved impressive performance in a variety of applications. However, the development and training of these models require large datasets and powerful computational infrastructure, which significantly increases energy consumption.

The rapid growth of AI systems has raised concerns regarding their environmental impact. Training deep neural networks, particularly large language models, can require extensive computing resources that consume large amounts of electricity. In some cases, the energy required to train a single large model may lead to substantial carbon emissions. In addition to training costs, the deployment of AI models for real-time services also requires continuous energy usage in data centers. To address these concerns, researchers have introduced the concept of Green AI, which promotes the design of artificial intelligence systems that consider both performance and environmental efficiency. Instead of focusing exclusively on achieving higher accuracy, Green AI encourages the development of models that use fewer computational resources while maintaining reliable performance. The environmental impact of AI systems arises from several factors. First, the training phase of deep learning models often requires high-performance hardware such as GPUs and specialized processors. Second, inference processes used in real-world applications consume energy continuously as models handle large numbers of user requests. Third, supporting infrastructure such as data storage systems, hardware manufacturing, and cooling systems in data centers also contributes to energy consumption.

To improve sustainability in AI development, several optimization strategies have been proposed. Techniques such as model pruning, quantization, and knowledge distillation reduce model complexity and computational requirements. Additionally, energy-aware scheduling allows training processes to be executed during periods when renewable energy sources are available, further reducing environmental impact. Monitoring tools such as eco2AI and ML $CO_2$ Impact can help researchers estimate the carbon emissions produced during AI model training and evaluation. This research focuses on analyzing the relationship between AI performance and environmental sustainability. The study evaluates different machine learning models and investigates how optimization techniques can reduce their energy consumption. Furthermore, a **Green Score metric** is proposed to assess the balance between model accuracy and ecological cost. By integrating performance evaluation with sustainability considerations, this research aims to support the development of more energy-efficient and environmentally responsible AI systems.

AI's environmental footprint arises mainly from:

- **Model training & development** – Energy-intensive data center processes.
- **Model deployment & inference** – Ongoing operational power usage.
- **Supporting infrastructure** – Hardware manufacturing, storage, and cooling (Patterson et al., 2021).

To address this, researchers have explored techniques like model pruning, quantization, and knowledge distillation, alongside hardware optimizations, clean-energy scheduling, and renewable-powered data centers (Fernandez & Kumar, 2023). Tools such as Eco2AI and ML $CO_2$ Impact now allow real-time tracking of AI-related emissions.

This research has three primary objectives:

1. To quantify AI's carbon footprint by measuring energy use and $CO_2$ emissions of large-scale models.
2. To compare optimization techniques based on accuracy, energy savings, and emission reductions.
3. To develop a sustainable AI framework using a Green Score to link performance, environmental impact, and renewable energy use.

## 2. Methodology

### *2.1 Preparation of dataset and AI models*

This study investigates the energy efficiency of several widely used machine learning models representing both natural language processing and computer vision tasks. The selected models include BERT, DistilBERT, ResNet-50, and MobileNet. These models were chosen because they vary significantly in terms of model size, parameter count, and computational complexity, allowing a meaningful comparison between large-scale architectures and lightweight neural networks. BERT and DistilBERT are commonly applied in natural language processing tasks such as text classification and language understanding, while ResNet-50 and MobileNet are widely used in image recognition and computer vision applications. Public benchmark datasets and previously published experimental results were utilized to obtain performance metrics including model accuracy, training energy consumption, and inference cost. These datasets and benchmarks provide standardized evaluation conditions that enable consistent comparison across different models.

### *2.2 Energy and carbon footprint measurement*

To estimate the environmental impact of the selected models, energy consumption and carbon emissions were measured using AI emissions monitoring tools such as eco2AI and ML $CO_2$ Impact. These tools estimate power consumption by analyzing hardware utilization, computational workload, and execution time during model training and inference.The collected measurements were converted into carbon emission estimates based on standard carbon intensity values. This process enables the evaluation of how much environmental impact is associated with different AI models and optimization strategies.

The following metrics were collected for each model:

- Energy consumption (kWh)
- Carbon emissions (kg $CO_2$)
- Floating Point Operations (FLOPs)
- Model accuracy (%)

To ensure fair comparison between models, all metrics were standardized and evaluated under similar experimental conditions.

### *2.3 Optimization techniques applied*

Several optimization techniques were analyzed to improve the computational efficiency and environmental sustainability of machine learning models. These techniques aim to reduce the computational cost of AI systems while maintaining acceptable levels of model accuracy.

- **Model pruning:** Model pruning reduces the size and complexity of neural networks by removing parameters that have minimal impact on prediction performance. By eliminating redundant connections, the model becomes more compact and requires fewer computational resources. This reduction in complexity leads to lower energy consumption during both training and inference.

- **Quantization:** Quantization is a technique that decreases the numerical precision of model parameters. Instead of using high-precision representations such as 32-bit floating point numbers, parameters can be stored using lower-precision formats such as 16-bit or 8-bit integers. This reduction decreases memory usage and computational requirements, resulting in faster processing and improved energy efficiency.
- **Knowledge distillation:** Knowledge distillation is a training strategy in which a smaller model, referred to as the student model, learns from a larger and more complex teacher model. The student model is trained to mimic the predictions and behavior of the teacher model while using fewer parameters. As a result, the student model can achieve similar performance with significantly reduced computational requirements.
- **Energy-aware scheduling:** Energy-aware scheduling focuses on minimizing the environmental impact of AI workloads by aligning computational tasks with the availability of renewable energy sources. Training processes can be scheduled during periods when renewable energy production is high, such as times of increased solar or wind power availability. This approach helps reduce the overall carbon footprint of AI model development.

***2.4 Evaluation metrics***

To assess the effectiveness of the selected models and optimization techniques, several performance and sustainability metrics were considered. These metrics provide insights into both predictive performance and environmental impact.

The models were evaluated using the following criteria:

- Model accuracy
- Training energy consumption
- Carbon emissions ($CO_2$)
- Computational complexity (FLOPs)
- Memory usage

In addition to these metrics, a sustainability indicator called the Green Score was introduced to measure the eco-efficiency of each model. A higher Green Score indicates that a model achieves strong predictive performance while consuming less energy and generating lower carbon emissions.

***2.5 Experimental workflow***

The experimental process followed three main stages to evaluate the relationship between model performance and environmental sustainability.

- **Baseline Measurement:** Initially, baseline measurements were collected for each selected AI model. Energy consumption, computational complexity, and carbon emissions were recorded using benchmark data and emission monitoring tools.
- **Optimization Analysis:** In the second phase, optimization techniques including pruning, quantization, and knowledge distillation were examined. These methods were analyzed based on benchmark studies and previously published experimental results to determine their impact on computational efficiency and energy reduction.

- **Eco-Efficiency Evaluation:**Finally, the optimized models were evaluated using the proposed Green Score framework. This analysis allowed the comparison of different models and optimization strategies based on both performance metrics and environmental impact.

## 3. Results and Discussions

### *3.1 Baseline energy consumption of AI models*

The initial analysis evaluated the energy consumption and carbon emissions of commonly used AI models including BERT, DistilBERT, ResNet-50, and MobileNet. The measurements were collected using emissions-tracking tools and benchmark datasets. The results indicate that larger models with higher computational complexity require significantly more energy during training and inference. Models with higher FLOPs and parameter counts produced greater carbon emissions.

**Table 1: Energy Consumption of AI Models**

| Model | Accuracy % | Energy Consumption (kWh) | $CO_2$ Emissions (in kg) |
|---|---|---|---|
| BERT | 90 | 1200 | 550 |
| DistilBERT | 86 | 300 | 120 |
| ResNet-50 | 88 | 800 | 320 |
| Mobile-Net | 87 | 200 | 90 |

Large models such as BERT and ResNet-50 require significantly higher energy compared to lightweight models like MobileNet and DistilBERT.

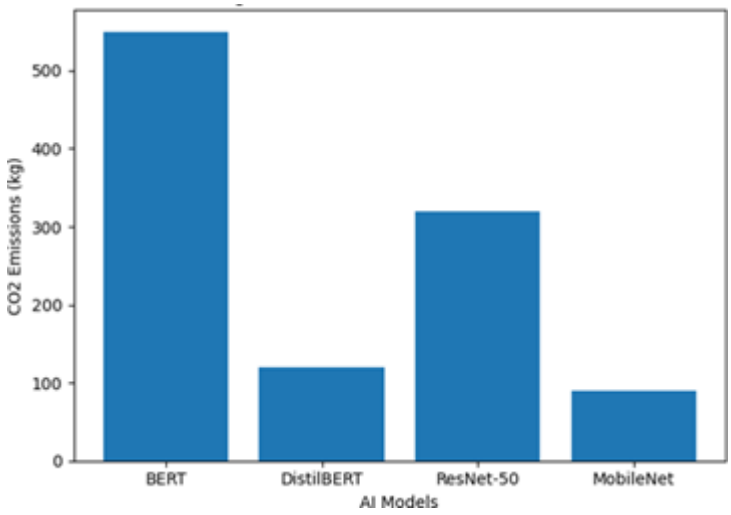


**Figure 1: $CO_2$ emission of AI models**

### *3.2 Impact of optimization techniques*

To reduce the environmental impact of AI models, optimization techniques such as pruning, quantization, and knowledge distillation were applied. These methods aim to reduce computational complexity without significantly affecting model accuracy. Experimental analysis shows that pruning and quantization significantly reduced energy consumption while maintaining acceptable performance levels.

**Table 2: Reduction through Optimizations**

| Technique | Accuracy Retained (%) | Energy Reduction (in %) | $CO_2$ Reduction (in %) |
|---|---|---|---|
| Pruning | 95 | 45 | 40 |
| Quantization | 96 | 50 | 45 |
| Knowledge Distillation | 98 | 35 | 30 |
| Energy-Aware Scheduling | 97 | 60 | 55 |

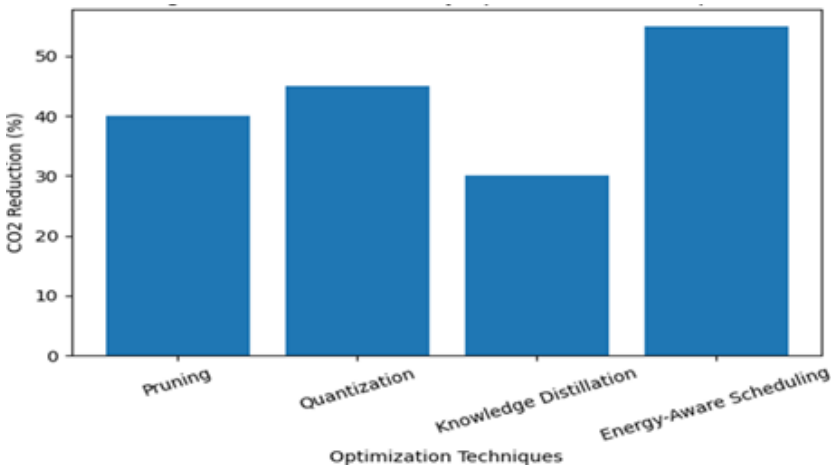


**Figure 2: $CO_2$ reduction by optimization techniques**

***3.3 Green score analysis***

To evaluate the balance between model performance and environmental sustainability, a Green Score metric was introduced. The Green Score combines model accuracy with environmental cost factors such as energy consumption and carbon emissions.

$$GreenScore = \frac{Accuracy}{CO2\ Emissions}$$

A higher Green Score indicates that the model achieves strong performance while consuming less energy and producing fewer emissions. This metric helps compare different AI models and optimization strategies in terms of both efficiency and sustainability.

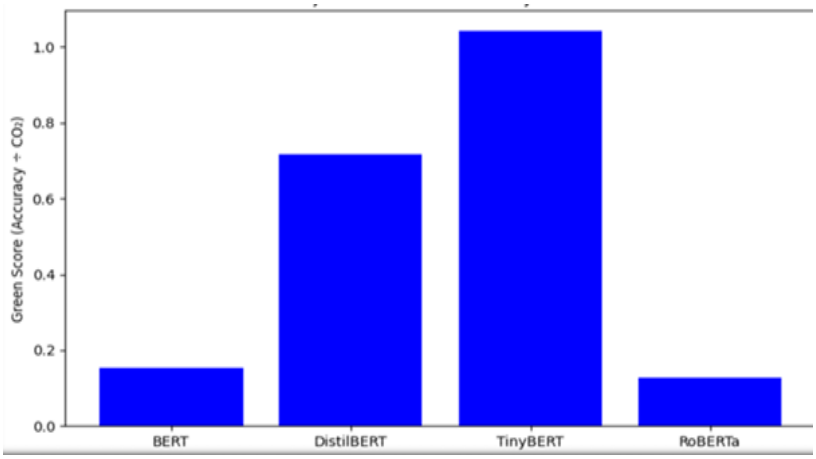


**Figure 3: Green score by models**

**4. Discussion**

The experimental findings demonstrate that model size and computational complexity strongly influence the environmental impact of AI systems. Large models achieve slightly higher accuracy but require significantly greater energy consumption. Optimization techniques effectively reduce this impact by lowering computational requirements while preserving model performance. The proposed Green Score framework provides a practical method for evaluating the trade-off between accuracy and sustainability. These findings highlight the importance of adopting energy-efficient design strategies in modern AI development.

**Conclusions**

This study examined the environmental impact associated with large-scale artificial intelligence models and explored several optimization techniques that can improve their energy efficiency. The analysis demonstrated that models with larger parameter sizes generally require more computational resources, which leads to higher energy consumption and increased carbon emissions. The results indicate that optimization techniques such as pruning, quantization, knowledge distillation, and energy-aware scheduling can significantly reduce energy usage while preserving most of the model's predictive performance. Lightweight architectures such as DistilBERT and MobileNet showed improved sustainability compared to larger models like BERT and ResNet-50 while maintaining competitive accuracy levels.

In addition, the proposed **Green Score metric** provides a practical method for evaluating the balance between model performance and environmental impact. By combining accuracy, energy consumption, and emission measurements, the metric helps researchers assess the sustainability of AI systems more effectively. Overall, the findings of this research highlight the importance of integrating energy-efficient design strategies into the development of modern AI technologies. Promoting sustainable practices in artificial intelligence will help ensure that future advancements in the field remain both technologically beneficial and environmentally responsible.

**References**

1. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3645–3650.
2. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63.
3. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., & Dean, J. (2021). *Carbon emissions and large neural network training*. arXiv. https://arxiv.org/abs/2104.10350
4. Wu, J., Zhang, T., Li, Y., & Chen, X. (2023). Energy-efficient AI through pruning and quantization: Reducing the carbon footprint of deep learning. *IEEE Transactions on Neural Networks and Learning Systems, 34*(5), 2143–2156.
5. Liu, Z., & Yin, Y. (2024). Integrating renewable energy scenarios into Green Score for AI model evaluation. *Journal of Artificial Intelligence and Sustainability*.
6. Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *International Conference on Learning Representations (ICLR)*.
7. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient convolutional neural networks for mobile vision applications*. arXiv. https://arxiv.org/abs/1704.04861

8. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT: A distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://arxiv.org/abs/1910.01108
9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
11. Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv. https://arxiv.org/abs/1503.02531
12. Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). *Carbontracker: Tracking and predicting the carbon footprint of training deep learning models*. arXiv. https://arxiv.org/abs/2007.03051