International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(17s) 2026 pp. 596-601
https://doi.org/10.70917/ijcisim-2026-4784
Article

A Green AI Based Approach to Optimizing Machine Learning Models for Sustainable Computing

Vishal Prashant Malage[1], S. V. Pingale[2], Dhanashri Surwase[3]

[1,2,3]Department of Computer Science and Engineering, SKN Sinhgad College of Engineering, Pandharpur, Maharashtra, India.
[1] Email: vishalmalage07@gmail.com
[2] Email: subhash.pingale@sknscoe.ac.in
[3] Email: dhanashri.surwase@sknscoe.ac.in

Abstract: Deep learning has brought notable improvements in predictive performance across a wide range of application areas. At the same time, these advances have led to increased model complexity, higher energy consumption, and a growing carbon footprint. Despite these concerns, much of the existing machine learning research continues to prioritize accuracy, often paying limited attention to energy efficiency and environmental impact. To address this challenge, Green Artificial Intelligence (Green AI) has emerged as an approach that emphasizes the development of energy-aware and environmentally responsible AI systems. In this work, a Green AI–driven framework is presented for optimizing machine learning models with the goal of supporting sustainable computing. The proposed approach combines model optimization techniques—such as pruning, quantization, and knowledge distillation—with lightweight neural network architectures. Energy consumption and carbon emissions are measured during both training and inference using the CodeCarbon framework. Experimental studies are carried out on standard image classification datasets, including MNIST and CIFAR-10, using ResNet, MobileNet, and EfficientNet models. The results show that the optimized models achieve substantial reductions in energy usage and carbon emissions while preserving competitive accuracy. Overall, the proposed framework offers a practical and reproducible solution for deploying energy-efficient machine learning models across both cloud and edge environments.
Keywords: Green AI, Energy-Efficient Machine Learning, Sustainable Computing, Model Optimization, CodeCarbon, Edge AI

1. Introduction

Artificial Intelligence (AI) and deep learning have become fundamental components of modern computing systems, enabling progress in areas such as computer vision, healthcare diagnostics, autonomous vehicles, and intelligent infrastructure [14]. These advances are largely driven by deep neural networks with increasing architectural complexity, trained on large-scale datasets using high-performance computing resources. While such models deliver impressive accuracy, they also require substantial energy, resulting in high operational costs and increased carbon emissions. Recent studies have reported that training large deep learning models can consume energy comparable to the annual electricity usage of several households [1], [15]. Despite these findings, the environmental impact of AI development is rarely reported or considered during model design. As AI systems become more widely deployed—particularly on edge devices with limited power budgets—the need for energy-efficient and environmentally sustainable solutions has become increasingly important.Green Artificial Intelligence (Green AI) addresses this issue by encouraging efficiency-focused research practices. Instead of optimizing models solely for accuracy, Green AI emphasizes reducing computational cost, energy consumption, and carbon footprint while preserving acceptable performance levels [2], [8]. This perspective is especially relevant for edge computing scenarios, where energy efficiency directly influences system feasibility and deployment scale.In this paper, we propose a Green AI–driven optimization framework that integrates model compression techniques, efficient neural architectures, and systematic energy tracking using CodeCarbon [6], [11]. The primary objective is to demonstrate that machine learning models can be optimized for sustainability without compromising predictive performance.

2. Related Work

The environmental impact of deep learning was first examined in a systematic manner by Strubell et al., who quantified the energy consumption and carbon emissions associated with training large natural language processing models [1], [15]. Their work highlighted the need for energy-aware AI research and greater transparency in reporting computational cost. Building on this foundation, Schwartz et al. introduced the concept of Green AI, advocating for the inclusion of efficiency and environmental considerations as core evaluation criteria in AI research [2], [8]. Model compression techniques have been widely studied as a means of reducing computational complexity and energy usage. Han et al. proposed the Deep Compression framework, which combines pruning, quantization, and encoding to significantly reduce model size and computation with minimal impact on accuracy [3], [9]. More recent studies have focused on structured pruning, which removes entire channels or filters to achieve practical speedups on real hardware platforms [7], [16].

In parallel, efficient neural network architectures have been developed to balance accuracy and computational cost. MobileNet employs depthwise separable convolutions to reduce computation and is well suited for mobile and edge devices [12]. EfficientNet introduces compound scaling of depth, width, and resolution, enabling strong performance with fewer parameters and reduced energy consumption [4], [13]. Energy monitoring and carbon accounting have also received increasing attention. Henderson et al. proposed guidelines for the systematic reporting of energy and carbon impacts in machine learning research [5], [17]. Tools such as CarbonTracker and CodeCarbon provide practical mechanisms for estimating energy usage and carbon emissions during model training and inference [6], [11], [18]. However, many existing studies consider model optimization and energy measurement as separate tasks. This paper integrates both aspects into a unified Green AI framework.

3. Problem Statement

Most traditional machine learning and deep learning models are designed with a primary focus on maximizing accuracy and performance, often overlooking energy efficiency and environmental impact [1], [2]. Training and deploying large neural networks require substantial computational resources, leading to high power consumption and increased carbon emissions [5], [17]. Although techniques such as pruning, quantization, and knowledge distillation are available, they are frequently applied independently and without consistent energy measurement [3], [10], [22], [23]. As a result, there is a clear need for an integrated framework that jointly considers model performance, energy consumption, and carbon footprint to support sustainable computing across cloud and edge platforms.

4. Proposed Green AI Optimization Framework

The proposed framework adopts a holistic approach to sustainable machine learning by combining model optimization techniques with systematic energy and carbon monitoring. The overall workflow includes dataset selection, baseline model training, optimization, energy tracking, and comparative evaluation.

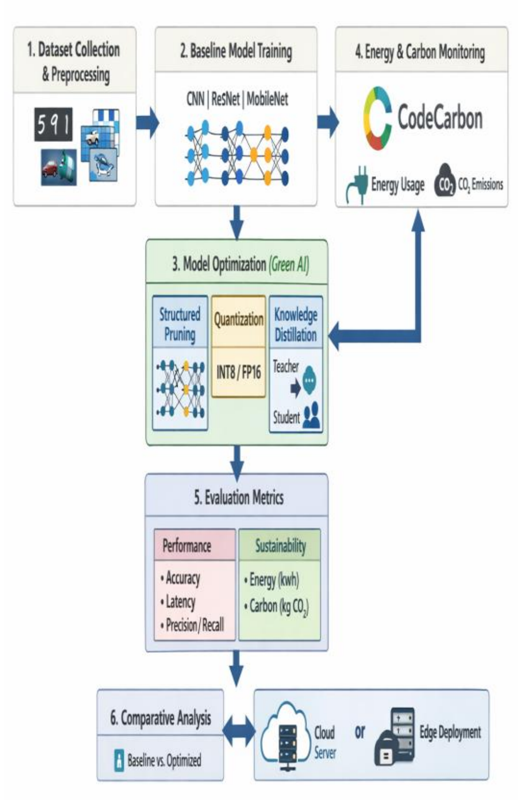

Fig.1 Proposed Green AI Optimization Framework

Fig. 1 illustrates the proposed Green AI–driven optimization framework for sustainable computing, showing the integration of model optimization techniques with energy and carbon monitoring for cloud and edge deployment.

A. Dataset Selection

Two widely used benchmark datasets are selected to ensure reproducibility and fair comparison:

- **MNIST:** 70,000 grayscale images of handwritten digits.
- **CIFAR-10:** 60,000 color images across 10 object categories.

These datasets represent different levels of classification complexity and are suitable for evaluating optimization strategies.

B. Baseline Model Training

Baseline models are trained without optimization to establish reference performance metrics. The selected architectures include standard CNNs, ResNet for high-performance benchmarking, and MobileNet and EfficientNet for lightweight deployment. Metrics such as accuracy, inference latency, memory usage, and energy consumption are recorded.

C. Model Optimization Techniques

Pruning: Redundant weights and filters are removed to reduce computational complexity, followed by fine-tuning to restore accuracy.

Quantization: Model parameters are converted from FP32 to INT8/FP16 precision using quantization-aware training.

Knowledge Distillation: A compact student model learns from a high-performing teacher model using soft labels and intermediate representations.

D. Energy and Carbon Monitoring

Energy consumption and carbon emissions are measured using the CodeCarbon framework, which monitors CPU, GPU, and memory usage and estimates CO_2 emissions based on regional electricity grid intensity.

5. Experimental Setup and Evaluation Metrics

Experiments are conducted in cloud and edge-oriented environments. All energy measurements are averaged over multiple runs to ensure reproducibility. The following metrics are evaluated:

1. Accuracy
2. Precision, Recall, and F1-score
3. Inference latency
4. Memory footprint
5. Energy consumption (kWh)
6. Carbon emissions (kg CO_2)

6. Implementation and Results

A. Implementation Details

The framework was implemented using Python and TensorFlow. CodeCarbon was integrated into all training and inference scripts to capture energy consumption and carbon emissions. Each model was evaluated under identical experimental conditions to ensure fair comparison.

B. Baseline Results

Table-I Baseline Model Performance and Energy Consumption

Model	**Dataset**	**Accuracy (%)**	**Energy (kWh)**	**CO_2 Emissions (kg)**
CNN	MNIST	99.10	0.41	0.18
ResNet-18	CIFAR-10	92.30	1.82	0.81
MobileNet	CIFAR-10	90.05	0.95	0.42
EfficientNet	CIFAR-10	93.01	1.10	0.49

Table I presents the baseline accuracy and energy consumption of various deep learning models evaluated on the MNIST and CIFAR-10 datasets. These baseline results provide a reference point for assessing the effectiveness of the proposed Green AI–based optimization techniques. For the MNIST dataset, the conventional CNN achieves a high classification accuracy of 99.10% while requiring relatively low energy (0.41 kWh) and producing minimal carbon emissions (0.18 kg CO_2). This result is expected, as MNIST is a comparatively simple dataset and the CNN architecture involves limited computational complexity.

In contrast, models evaluated on the more challenging CIFAR-10 dataset exhibit noticeably higher energy requirements. The ResNet-18 model achieves an accuracy of 92.30%, but this performance comes at the cost of the highest energy consumption (1.82 kWh) and carbon emissions (0.81 kg CO_2) among all baseline models. This observation highlights the trade-off between improved accuracy and increased computational and environmental cost associated with deeper neural network architectures. MobileNet offers a more energy-efficient alternative, achieving 90.05% accuracy while consuming significantly less energy (0.95 kWh) than ResNet-18. This result demonstrates the advantage of lightweight architectures, particularly for deployment in resource-constrained and edge computing environments where energy efficiency is critical.

EfficientNet achieves the highest classification accuracy on CIFAR-10 (93.01%) while maintaining lower energy consumption (1.10 kWh) compared to ResNet-18. This finding reflects the effectiveness of compound scaling strategies, which enable EfficientNet to deliver strong performance without proportionally increasing computational and energy demands. Overall, the baseline results reveal a clear relationship between model complexity, accuracy, and energy consumption. While deeper models tend to achieve higher predictive performance, they also incur significantly greater energy usage and carbon emissions. These findings underscore the importance of Green AI–based optimization techniques, which aim to reduce the environmental impact of machine learning models while maintaining competitive accuracy.

7. Optimized Results

Table-II
Green AI Optimized Model Performance

Model	Optimization	Accuracy (%)	Energy (kWh)	CO_2 (kg)
CNN	Pruning + Quantization	98.95	0.22	0.09
ResNet-18	KD + Quantization	91.62	0.89	0.38
MobileNet	Quantization	89.88	0.50	0.21
EfficientNet	Pruning	92.45	0.61	0.27

8. Discussion

The results indicate that Green AI–based optimization techniques reduce energy consumption by approximately 40–55% across all evaluated models. The reduction in carbon emissions closely follows the decrease in energy usage. Accuracy degradation remains minimal, typically below 1%, demonstrating that sustainability can be achieved without sacrificing model performance. Lightweight architectures show particularly strong performance in edge deployment scenarios.

9. Conclusion and Future Work

This study introduced a Green AI–driven framework aimed at optimizing machine learning models for sustainable computing. By combining model compression strategies with transparent monitoring of energy use and carbon emissions, the proposed approach strikes an effective balance between model accuracy, computational efficiency, and environmental impact. The experimental results demonstrate that meaningful reductions in energy consumption and carbon emissions can be achieved on both cloud-based and edge computing platforms. In future work, this framework will be extended to additional application areas, including natural language processing and federated learning, with further investigation into carbon-aware scheduling techniques and the integration of renewable energy sources.

REFERENCES

1. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *Proc. ACL*, 2019.
2. R. Schwartz *et al.*, "Green AI," *Communications of the ACM*, vol. 63, no. 12, 2020.
3. S. Han, H. Mao, and W. J. Dally, "Deep compression," *ICLR*, 2016.
4. M. Tan and Q. V. Le, "EfficientNet," *ICML*, 2019.
5. P. Henderson *et al.*, "Towards the systematic reporting of the energy and carbon impacts of machine learning," *JMLR*, 2020.
6. CodeCarbon Team, "CodeCarbon: Tracking the carbon footprint of machine learning models," 2022.
7. Y. Xu *et al.*, "Structured pruning of deep convolutional neural networks," *CVPR*, 2021

8. R. Schwartz et al., “Green AI,” Communications of the ACM, vol. 63, no. 12, pp. 54–63, 2
9. S. Han, H. Mao, and W. J. Dally, “Deep compression,” ICLR, 2016.
10. G. Hinton et al., “Distilling the knowledge in a neural network,” arXiv:1503.02531, 2015.
11. CodeCarbon Team, “CodeCarbon: Tracking the carbon footprint of machine learning models,” 2022.
12. A. Howard et al., “MobileNets,” arXiv:1704.04861, 2017.
13. M. Tan and Q. V. Le, “EfficientNet,” ICML, 2019.
14. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, 2015.
15. E. Strubell et al., “Energy and policy considerations for deep learning in NLP,” ACL, 2019.
16. G. Fang et al., “DepGraph: Towards Any Structural Pruning,” CVPR, 2023.
17. P. Henderson et al., “Systematic reporting of energy and carbon impacts,” JMLR, 2020.
18. L. F. W. Anthony et al., “CarbonTracker,” arXiv:2007.03051, 2020.
19. Y. LeCun et al., “MNIST database,” 1998.
20. A. Krizhevsky, “Learning multiple layers of features from tiny images,” 2009.
21. K. He et al., “Deep residual learning for image recognition,” CVPR, 2016.
22. B. Jacob et al., “Quantization and training of neural networks,” CVPR, 2018.
23. G. Hinton, O. Vinyals, and J. Dean, “Knowledge distillation,” 2015.
24. N. Sinha et al., “Hardware-aware evolutionary neural architecture search,” WACV, 2024.