BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

GREEN AI: BUILDING ENERGY-EFFICIENT ARTIFICIAL INTELLIGENCE : *A Review of Techniques, Trends and Research Gaps*

[1]Shreya Deshpande, [2]Prof. Geeta Kodabagi

[1,]Student, Roll No. A49, TE (A) Sem-I, AY 2026-27
[1,2]Department of Artificial Intelligence & Data Science Engineering
[1,2]Ajeenkya DY Patil School of Engineering, Pune, India

Abstract

Artificial Intelligence (AI) has become a major asset now in various industries. However, developing and running high-performance AI models usually require massive computing power, which leads to enormous power and energy usage. Green AI aims at significantly reducing these demands and environmental damage by finding ways to make AI models both highly accurate and environmentally friendly. This work gives an overview of Green AI key concepts and the most recent researches done in this direction. It starts by introducing Green AI along with its role in sustainable computing and proceeds to explore the various ways that Green AI is being implemented such as pruning, quantization, model compression, efficient network design, TinyML, edge deployment, and hardware-level optimization. Also, this paper estimates the ecological cost of AI and compares ten articles that have been published in the period 2015 - 2025. The findings highlight several problems yet to be solved for instance the trade-off between model accuracy and energy consumption, no agreement on a standard method of disclosing energy and carbon emissions, the computational burden of massive neural networks, the limited implementation by the industry, and the necessity of research on environmentally friendly Large Language Models (LLMs) and generative AI.

Keywords: Green AI, energy-efficient AI, pruning, quantization, model compression, TinyML, carbon footprint, sustainable computing.

Introduction

Artificial Intelligence, or AI, is a tool that industries have been using extensively to solve image, language, or decision-making type problems. However, these AI capabilities involve a certain resource cost. The energy and resource consumption of running and training large models for lots of users is high. As a result of its widespread deployment AI's demand for power has a financial and environmental impact.

Green AI aims to produce the most eco-conscious yet most functional models possible by reducing the amount of resources spent. Along the lines of Schwartz et al.'s [5] paper on Green AI, the scientists should report their models on computing efficiency together with accuracy rather than only on the level

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

of accuracy. As a result, there has been a rise in studies on estimating, reducing and disclosing the energy use, and carbon emissions from machine learning.

We have done a review of the work on Green AI for the technical seminar and have explained the problem statement, defined the objectives of the research, have shown how this topic aligns with present day tech trends, have compared ten studies chosen by us, and, finally, have described the aspects where there is a need for further investigations.

Problem Statement

In accordance with the notion of Green AI, an increasing number of scientists have set out to investigate how energy consumption and carbon emissions related to machine learning can be measured, minimized and communicated.

Objectives

The study has the following objectives:

- To study recent trends and future prospects of GreenAI.
- To elaborate why Green AI is a significant factor in sustainable computing.
- To summarize the techniques which help reduce the energy expenditure of AI systems.
- To look at AI's environmental aspects and the way it will be Green AI which allows for the carbon emissions.
- To identify areas where Green AI can be used and where it may offer solutions at a future level.
- Under supervision from the mentor, to compose a technical seminar-based review report.

Relevance to Current Technology Trends

It is obvious why Green AI is the focus of so many as businesses increasingly adopt measures to make computations less demanding on the environment, such as lowering power consumption and carbon emissions. Industry activities currently underway are, for example, Google's exploration in data-center cooling, Microsoft Azure’s green AI initiatives, OpenAI's work to boost efficiency in AI training, and NVIDIA’s efforts to engineer hardware that can support energy-efficient AI operations.

Energy-Efficient Techniques

The research in the current report presents various methods in line with the common goal of making AI systems less environmentally unfriendly.

Pruning

The pruning is basically the act of trimming the neural network in a way that we keep the weights or connections which are most responsible for the output and remove or reduce those weights or connections that are not contributing much to the output. Pruning allows the model to get even better when it has been trained on fewer samples or when we get a new dataset in the future [10]. But it is not

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

a one step process in most of the time; we have to give the model a second training or maybe some extra tuning after pruning.

Quantization

Quantization replaces exact calculations with less accurate ones to save memory. It can reduce memory usage and energy consumption [8]. However, the accuracy of some jobs is affected slightly. Zhou et al. [1] showed in their study comparing Small Language Models that with energy efficiency in consideration quantization was superior to pruning but both did not affect performance very much.

Model Compression and Efficient Architectures

The model is made considerably easier or even smaller if several methods like pruning and quantization are combined with other approaches to achieve this goal of reducing the number of operations [10]. MobileNet and EfficientNet based architectures, for instance, can provide very high performance by utilizing very limited parameters and low power consumption, respectively [7]. These kinds of efficient designs that can save power are especially suitable for environments with constraints due to limited resources, on the other hand, they may not work as well when applying to large-scale data-driven models.

TinyML and Edge AI

TinyML along with edge AI allow AI apps to run on gadgets with minimum memory, processors and energy. Among the object-detection systems optimized specifically for use on the edge we see some examples [2]. In the domain of TinyML it is found that a combination of pruning with quantization can save energy and yet result in a relatively tiny accuracy loss [4].

Hardware Optimization

The main objective of hardware optimization is to enhance the computing system's performance by efficiently processing deep learning tasks [9]. It is also indicated in the papers that merging hardware optimization gains with compression techniques may bring additional savings in the energy consumption [3].

Measuring Energy and Carbon Footprint

With their publication, Henderson et al. [6], outlined the ways to assess and report the energy use and greenhouse gas emissions of machine learning activity, with emphasis on training. Having the data at hand is the first and most important thing: it enables a fair comparison between methods and allows the scientists to assess, if the introduced idea is really helping the environment.

Literature Survey

Methods, findings, and limitations of the ten studies that were analyzed in the review are collated and described in Table 1. The table is ordered in a way that newest entries come first while oldest end the list.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Table 1: Literature Survey Matrix

Author/Year	Methodology	Key Findings	Limitations
Zhou et al. [1] (2025)	Comparison of the effectiveness of pruning and quantization methods in the case of small language models	Quantizing the layers helped to cut down the energy needs without a significant deterioration of the model's capacity	Evaluated only on Small Language Models
Bhushan et al. [2] (2025)	TinyML-based edge AI	Running a machine-learning based object recognition algorithm from power saving perspective	Relevant mainly to TinyML and edge scenarios
Isenkul [3] (2025)	Weight pruning, activation quantization and hardware upgrades	Cutting the power demand without affecting the model output	Mainly tested on specific real-time applications
Greg [4] (2025)	Quantization and pruning for TinyML	A joint pruning-quantization approach that cuts down energy use with very little sacrifice in accuracy	Review-based work mostly; few instances of real-life testing
Schwartz et al. [5] (2020)	Green AI introduction and computational-cost evaluation	Emphasizing measuring efficiency alongside with accuracy	Mainly a theoretical framework
Henderson et al. [6] (2020)	CO2-emission measurement in the AI training process	Different approaches have been proposed to calculate or estimate AI carbon footprint.	Training has got the lion's share
Howard et al. [7] (2019)	MobileNet / EfficientNet-style architectures	These methods work very accurately as fewer parameters and with lower power are required.	But it is not suitable for extremely largescale models.
Jacob et al. [8] (2018)	Quantization	Reduced energy consumption and memory utilization via computation with low precision.	In some tasks there is a minimal loss of accuracy.
Sze et al. [9] (2017)	Energy-efficient DNN hardware design	Energy-saving methods based on hardware optimization	These methods focus on hardware while algorithms are largely ignored.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Author/Year	Methodology	Key Findings	Limitations
Han et al. [10] (2016)	They suggested pruning of the neural network (deep pruning)	They were able to cut the model size and the amount of computation down without the accuracy being affected.	Requires retraining after pruning

Research Gap

The studies reviewed point to several areas where further work is needed:

- Although many ways are there to increase energy efficiency and accuracy simultaneously, yet achieving the both simultaneously is not easy.
- In the area of energy use and carbon emissions of AI, measuring and reporting by standards is currently a major obstacle that the research community is still trying to address through the development of a unified framework.
- It is still the case that developing and operating big AI models take a considerable amount of computing resources.
- Green AI is still not widely used by industries.
- Energy use of Large Language Models (LLMs) and generative AI is the area that needs to be more researched due to the continued increasing use.
- Future solutions must not only work for all different scales, and be cheap to deploy, but also environmentally-friendly in the long-term.

Conclusion

Green AI can significantly lower energy bills and electricity use by means of the changes in computing hardware and software. Pruning, quantization, and lightweight network design besides other methods are used to increase the efficiency of AI systems. Nevertheless, the studies that have been done in the field reveal that measuring carbon consumption as well as energy use reliably is still one of the challenges which the research community face. At the same time it is not clear whether methods for optimization can be scaled effectively to such large models and in what measure the changes are being adopted by the industry at large. Only when these issues are resolved will Green AI become a realistic approach for bringing highly efficient AI systems to life without sacrificing their environmental friendliness.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

References

[1] Y. Zhou et al., "Pruning vs. quantization for small language models," in Findings of the Association for Computational Linguistics: EMNLP, 2025.

[2] A. Bhushan et al., "TinyML-based energy-efficient edge AI for sustainable computing," Sci. Rep., vol. 15, 2025.

[3] M. Isenkul, "Energy-efficient deep learning using pruning, quantization, and hardware optimization," J. Real-Time Image Process., vol. 22, 2025.

[4] J. Greg, "Quantization and pruning strategies for energy-efficient TinyML models," ResearchGate preprint, 2025.

[5] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Commun. ACM, vol. 63, no. 12, pp. 54–63, 2020.

[6] P. Henderson et al., "Towards the systematic reporting of the energy and carbon footprints of machine learning," J. Mach. Learn. Res., vol. 21, no. 248, pp. 1–43, 2020.

[7] A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019.

[8] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.

[9] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," Proc. IEEE, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[10] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. Int. Conf. Learn. Represent. (ICLR), 2016.