ENERGY-EFFICIENT DEEP LEARNING ARCHITECTURES FOR SUSTAINABLE AI

[1]VaishaliDeshwal, [2]Anuradha Bharti

[1]Departmentof CSE, AjayKumar Garg Engineering College, Ghaziabad, U.P, India

[2] Department of ECE, Meerut Institute of Engineering and Technology, Meerut, UP., India

deshwalvaishali@akgec.ac.in, Anuradhabharti05@gmail.com

Abstract—**Deep learning has revolutionized countless fields with its powerful capabilities, but this progress comes with a hidden cost—high energy consumption and growing environmental impact. As models become larger and more complex, the need for sustainable and energy-efficient AI has become increasingly urgent. This paper takes a closer look at the carbon footprint of modern deep learning systems and underscores the importance of building AI solutions that are not only intelligent but also environmentally responsible. We explore what it truly means for a deep learning system to be "energy efficient" and present key architectural strategies such as model pruning, quantization, knowledge distillation, and the use of lightweight models like MobileNet and EfficientNet. In addition, we examine how hardware-aware designs—including neuromorphic chips, FPGAs, and Edge AI devices—can play a vital role in achieving better energy-performance balance, especially when aligned with the algorithmic design. The paper also highlights ongoing challenges and trade-offs, particularly the tension between accuracy and efficiency, and the often-overlooked energy costs of training versus inference. Looking ahead, we discuss promising research directions involving Green AI metrics and AutoML tools for optimizing energy use. Ultimately, this study calls for a conscious shift in AI development practices—toward models and systems that are not only high-performing, but also sustainable and aligned with environmental goals.**

Keywords—**Energy-Efficient Deep Learning, Sustainable AI, Green AI Architectures, Model Compression Techniques, Edge AI and Neuromorphic Computing, Low-Power Machine Learning, Hardware-Aware Neural Networks**

I. INTRODUCTION

In recent years, the explosive growth of artificial intelligence (AI) has underscored the pressing need for sustainable design principles. The research paper *Sustainable AI: EnergyEfficient Deep Learning Architectures for Edge Devices* by Meenalochini Pandi et al. (2025) explores how edgeoptimized models—leveraging quantization, pruning, model compression, and neural architecture search—can slash energy consumption by up to 70% while maintaining competitive accuracy. This approach not only extends battery life in resourceconstrained devices but also trims the carbon footprint of deploying AI at scale [1]. Similarly, the survey *Towards energyefficient deep learning for sustainable AI* highlights systemic improvements—across infrastructure, data handling, modeling, training, deployment, and evaluation—to reduce resource use throughout the AI lifecycle. These studies collectively emphasize that sustainable AI is not optional—it is a foundational necessity as intelligent systems permeate everyday devices [2].

Sustainable AI is pivotal for aligning technological progress with ecological responsibility, and Ranpara's paper highlights this by focusing on **circular economy principles**. The proposed multi-layered framework integrates energy-conscious computational models, machine learning algorithms, and formal optimization techniques—like mixed-integer linear programming and lifecycle assessments—to guide decision-making for resource reuse and waste reduction. By applying the framework to real-world scenarios such as lithium-ion battery recycling and urban waste management, the paper demonstrates **a 25% reduction in energy consumption** and an **18% improvement in resource recovery efficiency**, showcasing how AI can actively enable sustainability. This underscore sustainable AI's importance in minimizing environmental impact while preserving performance and scalability, making it indispensable in today's resource-constrained world.

The significance of sustainable AI extends beyond theoretical promise to **practical, measurable impact**, as Ranpara's study further illustrates. By optimizing logistics, the framework reduced transportation-related emissions by 30%, and AI-driven classification improved urban-waste sorting accuracy by 20%. Such advancements demonstrate how energy-efficient architectures can serve multiple facets of a circular economy: enhancing production, improving resource use, and reducing emissions. Moreover, this work aligns with global sustainability initiatives—especially UN Sustainable Development Goals—by offering a scalable, scientifically grounded architecture. In essence, sustainable AI—through frameworks like Ranpara's—bridges high-level environmental objectives and real-world applications, enabling intelligent systems capable of delivering economic, ecological, and societal benefits. [3]

While efficient, compact architectures are critical, it is equally important to address the substantial environmental impact of large deep learning models. Zewe's MIT News article

(January 2025) emphasizes that training generative AI models with billions of parameters, such as GPT4, consumes vast electricity and water resources—exerting stress on electric grids and municipal water systems. Further, the arXiv study *Holistically Evaluating the Environmental Impact of Creating Language Models* examined models ranging from 20 M to 13 B parameters, concluding that a single series of model developments emitted 493 metric tons of CO_2 and consumed nearly 2.8 million L of water—half of which stemmed from development stages often overlooked in reporting [4].

Sustainable AI is essential for navigating the complex balance between technological innovation and ecological stewardship. Ren et al. confront two prevailing narratives: one stresses the substantial environmental footprint of large language models (LLMs)—including high energy consumption, CO_2 emissions, and water usage—while the other suggests LLMs may outperform human labor in eco-efficiency. Their comparative analysis reveals that, in the U.S., conventional LLMs like Llama370B consume substantially less energy—by factors of 40 to 150—compared to humans performing the same tasks, and lightweight LLMs like Gemma2Bit achieve ratios of 1,200 to 4,400 This duality emphasizes that sustainable AI isn't just about reducing computational waste, but about enabling smarter, more efficient workflows that can outpace traditional systems—effectively positioning LLMs as a potential lever for environmental progress [5].

II. MOTIVATION

The past decade has witnessed a meteoric climb in deep learning model sizes and complexity, driven by breakthroughs in architectures like CNNs and transformers. This escalation, however, comes at a steep environmental price. In their 2023 study, Xu et al. analyzed the **energy efficiency of training multiple neural network architectures**, revealing that larger and deeper models consume disproportionately more power during training—and, as a result, emit significantly higher CO_2—compared to smaller, optimized networks. Their empirical findings underscore an urgent need to balance model performance with ecological cost, highlighting that even accuracy gains may not justify the exponential rise in energy consumption [6].

The environmental toll of deep learning research extends beyond model training to include the broader experimental pipeline. In a compelling study titled *From Clicks to Carbon: The Environmental Toll of Recommender Systems*, researchers compared deep learning–based recommender systems with traditional ML models and found a staggering 42× increase in CO_2 emissions over a decade (2013–2023), despite improvements in hardware efficiency. On average, a single deep learning research experiment consumed ~6,854 kWh—eight times more energy than conventional methods, without necessarily offering proportional performance gains. This work highlights that without transparency in energy usage and deliberate architecture optimization, the rise of deep learning could become an environmental burden rather than a technological boon [7].

Energy efficiency in deep learning refers to the amount of computational work achieved (e.g., training or inference) per unit of energy consumed—typically measured in kilowatt-hours (kWh)—often normalized against a performance metric such as accuracy or throughput. Charles Tripp et al.'s 2024 empirical study *"Measuring the Energy Consumption and Efficiency of Deep Neural Networks: An Empirical Analysis and Design Recommendations"* introduces a comprehensive energy model grounded in node-level watt-meter measurements. They analyzed 63,527 training runs across varying depths, architectures, and hardware, revealing non-linear relationships between network design and energy use that challenge simple assumptions like "fewer parameters = more efficient". Based on this, they recommend combining hardware-aware design, memory hierarchy optimization, and algorithmic changes to truly maximize energy efficiency [8].

Complementing this empirical foundation, GarcíaCarbajal et al.'s 2025 journal paper, *Towards an Energy Consumption Index for Deep Learning Models*, introduces a standardized **Energy Consumption Index (ECI)** to quantify energy used per task across training and inference. By evaluating well-known architectures—AlexNet, ResNet, EfficientNet, ConvNeXt, and Swin Transformer—on GPUs like TITAN XP, the study uncovered dramatic efficiency differences tied to architecture, device, and phase of model lifecycle. The ECI facilitates fair comparisons and promotes transparent reporting of energy-use data. Such a metric aligns directly with the motivation behind this work: to develop an *Energy-Efficient Deep Learning Architecture for Sustainable AI*, it becomes imperative to incorporate standardized energy efficiency metrics during design, ensuring proposed architectures not only perform but also demonstrate quantifiable environmental benefits [9].

III. ARCHITECTURAL TECHINIQUES FOR EFFICIENCY

Model pruning and quantization have become indispensable techniques for reducing the computational and energy overhead of deep neural networks. A cutting-edge study titled *Automatic Joint Structured Pruning and Quantization for Efficient Neural Network Training and Compression* (Qu et al., 2025) introduces the GETA framework, which seamlessly combines structured pruning with quantization-aware training. GETA builds a quantization-aware dependency graph (QADG) to guide pruning, and employs a partially projected stochastic gradient method ensuring bit-width constraints across layers are satisfied. Results show GETA produces smaller, high-performance models—often outperforming traditional two-stage methods—demonstrating that co-optimization of pruning and quantization enhances energy efficiency without compromising accuracy [10].

Further illustrating the synergy between pruning and quantization, Balaskas *et al.* (2023) in their work *Hardware-Aware DNN Compression via Diverse Pruning and Mixed-Precision Quantization*, propose an automated, hardware-sensitive framework. Leveraging reinforcement learning, their method simultaneously applies coarse- and fine-grained structured pruning and per-layer mixed-precision quantization to tailor models for embedded accelerators. Evaluated on CIFAR-10/100 and ImageNet, this approach achieves an average **39% reduction in energy consumption** with only a minimal 1.7% drop in accuracy, outperforming baseline compression strategies. By adapting pruning intensity and quantization precision to hardware capabilities, this work exemplifies how architectural efficiency can align with sustainability goals in AI[11].

Knowledge distillation enables efficient AI by transferring learning from a large, high-capacity *teacher* model to a compact *student* model, reducing compute and energy costs while preserving performance. Wu et al.'s 2025 paper, *Knowledge Distillation with Adaptive Influence Weight (KDAIF)*, advances this concept by integrating influence functions—borrowed from robust statistics—to assign dynamic, data-driven weights to training samples under the principles of Sustainability, Accuracy, Fairness, and Explainability (SAFE). KDAIF not only allows the student to learn more efficiently (reducing training time and energy) but also enhances model transparency by highlighting which examples matter most. Experiments on CIFAR10/100 and GLUE benchmarks show that KDAIF outperforms existing distillation methods in both performance and learning efficiency—thereby offering a powerful template for future energy-conscious architectures [12].

Moving beyond traditional neural networks, Konstantaropoulos et al. (2025) introduce a compelling energy-saving framework in *Dynamic Activation with Knowledge Distillation for EnergyEfficient Spiking Neural Network Ensembles*. Here, a high-performance ANN teacher guides an ensemble of lightweight spiking neural networks (SNNs), each learning distinct aspects of the task. By dynamically activating only relevant student SNNs per input, this ensemble reduces computation and energy—achieving up to **20× fewer FLOPs** and a **65% drop in energy use**, with only a ~2 % accuracy loss on CIFAR10. This architecture demonstrates how knowledge distillation can be combined with neuromorphic designs to build highly efficient, context-aware systems ideal for energy-constrained or edge deployments [13].

Lightweight architectures are streamlined neural network designs that balance compactness with performance efficiency—typically achieved through innovative building blocks like depthwise separable convolutions, inverted residuals, and compound scaling. MobileNet and EfficientNet exemplify this paradigm. MobileNet architectures (V1–V4) have become foundational in designing energy-efficient AI for edge and mobile environments. These models utilize depthwise separable convolutions: a depthwise convolution followed by a pointwise convolution, significantly reducing computation without substantial accuracy loss. A recent 2025 study, *Energy-Efficient AI for Medical Diagnostics: Performance and Sustainability Analysis of ResNet and MobileNet*, empirically demonstrates that MobileNet consumes considerably less power and trains faster than ResNet when classifying thoracic diseases—translating to lower energy cost and carbon emissions [14].

IV. HARDWARE – AWARE NEURAL DESIGN

Neuromorphic processors, inspired by the brain's structure and function, offer a transformative path toward ultra-efficient AI. Vogginger et al. (2024) in *Neuromorphic hardware for sustainable AI data centers* highlight how neuromorphic systems like Intel Loihi and IBM TrueNorth drastically reduce energy consumption and alleviate data-transfer bottlenecks by merging memory and processing, processing spikes only when triggered—achieving 10–100× lower power usage compared to traditional accelerators. Complementing this, B2Bdaily's recent overview underscores real-world gains: Innatera's T1 and Loihi 2 NPUs achieve up to 100× power savings while supporting real-time inference in edge devices. These neuromorphic architectures deliver low-latency, event-driven processing, making them ideal for sustainable AI systems that require both performance and minimized energy footprint [15].

FPGAs enable custom, energy-optimized AI pipelines tailored to specific models. The 2025 study *Optimizing Deep Learning Acceleration on FPGA for RealTime and ResourceEfficient Image Classification* by Mouri Zadeh Khaki & Choi shows that an FPGA-accelerated CNN implementation can lower energy consumption by 40% while maintaining accuracy, thanks to integer arithmetic support and hardware-level optimizations like pipelining and clock gating. Earlier work in *FPGA/DNN CoDesign* emphasizes the co-design methodology, where models and FPGA architecture are optimized together to yield up to 2.5× improvement in energy efficiency over GPUs, all without sacrificing performance. These findings demonstrate that FPGAs are a flexible and sustainable solution, enabling just-in-time customization of AI workloads for maximum power efficiency [16].

Deploying AI directly on edge devices further promotes sustainable AI by minimizing data transmission and cloud reliance. Peccia et al. (2024) in *Efficient Edge AI: Deploying CNNs on FPGA with the Gemmini Accelerator* achieved 36.5 GOP/s/W efficiency running real-time YOLOv7 on a Xilinx ZCU102 FPGA—demonstrating that edge-AI solutions can rival server-level performance at a fraction of the energy cost. Additionally, Susskind et al.'s 2023 study *ULEEN* presented a novel weightless neural network architecture, running on FPGA and ASIC, delivering up to **13 million inferences per Joule**—over 7× more efficient than traditional quantized models, while maintaining high accuracy. These advancements confirm that edge AI, driven by hardware-aware design, can sustain sophisticated intelligence on constrained devices without compromising sustainability [17].

Algorithm–hardware codesign is a pivotal techniquein developing deep learning systems that are not only performant but also energyconscious, advancing the goalof sustainable AI. Algorithm–hardware codesign integrates neural architecture design directly with hardware constraints, enabling simultaneous optimization of both dimensions for superior energy efficiency. In **Fan et al. (2021),** *Algorithm and Hardware Codesign for Reconfigurable CNN Accelerator*, the authors present a framework that jointly searches for neural subnetworks and hardware configurations to achieve Paretooptimal trade-offs. Their results show up to **8.5× higher energy efficiency, 3× lower latency**, and **2–6% improved accuracy** over manually designed models like ResNet101 and MobileNetV2 This codesign methodology ensures that model choices—from layer types to structure—are directly informed by hardware performance metrics, making it particularly effective for sustainable AI where minimizing energy per inference is crucial [18].

V. CHALLENGES

Accuracy vs. efficiency: navigating accuracy vs. efficiency within deep learning architectures is a critical challenge. Balancing marginal performance gains with energy consumption—and ensuring fault-tolerant operation—must be integral to any sustainable AI design paradigm. Yang, Adamek & Armour's 2024 study *DoubleExponential Increases in Inference Energy: The Cost of the Race for Accuracy* conducted an extensive evaluation of over 1,200 ImageNet classification models. Their analysis revealed a steep **diminishing return** on accuracy gain: every incremental boost in top-1 accuracy requires **exponentially larger inference energy**, highlighting that pursuing marginal gains can be energetically prohibitive. To guide sustainable model selection, the authors introduced an **energy-efficiency scoring system**, promoting models that offer balanced accuracy and power consumption. This work illustrates a core challenge in sustainable AI: beyond a certain point, minor improvements in accuracy are not justified by the disproportionate energy they require—demanding a more nuanced, eco-aware approach to model design [19].

Complementing this, Siddique, Basu & Hoque (2021) in *Exploring FaultEnergy Tradeoffs in Approximate DNN Hardware Accelerators* explore accuracy loss when deploying approximate computing to reduce energy. Their experiments on MNIST and FashionMNIST show that **small, quantization-inspired energy savings can trigger massive accuracy drops under hardware faults**, with up to 66% misclassification versus only 9% in precise accelerators. The study highlights that energy-efficient approximations may introduce vulnerabilities—not only reducing accuracy but also affecting reliability. Consequently, sustainable AI must address **accuracy vs. efficiency** as a multi-faceted trade-off, incorporating robustness and resilience at the hardware layer and ensuring that energy savings do not incur unacceptable performance or safety costs [20].

Training vs. inference energy use: Balancing training and inference energy use is a critical challenge in building **energy-efficient deep learning architectures for sustainable AI.** While training deep learning models is undeniably resource-intensive, inference often accounts for a larger share of lifetime energy usage. Yarally et al. (2023), in *Uncovering Energy-Efficient Practices in Deep Learning Training*, revealed that careful adjustments—like reducing model complexity and using Bayesian hyperparameter tuning—can halve training energy. However, even with optimized training, inference remains the dominant phase, especially in real-world applications involving continuous deployment and updates [21].

Efforts to minimize energy use across both phases require different approaches. Yarally et al. showed that reducing unnecessary experiments and tuning hyperparameters with cost-aware strategies effectively lowers training energy without sacrificing accuracy. However, the inference phase demands tailored architecture design—models must be compact, hardware-aware, and quantized for efficient deployment. The *From Computation to Consumption* study further supported this, indicating that training and inference cannot be optimized in isolation; profiling GPU and memory usage during both phases is essential to identify dominant energy sinks [22].

VI. FUTURE SCOPE

Establishing robust Green AI metrics and refining AutoML to optimize energy use—are essential for guiding the next generation of **energy-efficient deep learning architectures for sustainable AI.**

As the environmental footprint of AI grows, standardized metrics are urgently needed to guide sustainable development. Clemm et al. (2024), in *Towards Green AI: Current Status and Future Research*, advocate for life-cycle–based assessment frameworks that integrate model, data, and infrastructure stages—enabling holistic measurement of energy use and carbon emissions [23].

Building on this, Budennyy et al. (2022) developed **Eco2AI**, an open-source toolkit that quantifies both energy consumption and CO_2 emissions during training and inference, promoting transparency and reproducibility. These efforts mark a pivotal shift: moving from arbitrary reporting to evidence-based sustainability standards. Future research must expand these metrics to benchmarking competitions, journals, and conferences—ensuring that energy efficiency becomes a mandatory performance dimension alongside accuracy [24]. AutoML has enormous potential to automate sustainable model design, but its own computational cost can be prohibitive. Hennig et al. (2024) presented a multi-objective AutoML framework—*Towards Leveraging AutoML for Sustainable Deep Learning*—which jointly optimizes accuracy and energy consumption when tuning Deep Shift Neural Networks, achieving high accuracy with significantly reduced compute [25].

VII. CONCLUSION

As AI continues to evolve and expand its role in everyday life, the importance of building it sustainably has never been more critical. This work explored a wide range of strategies focused on improving the energy efficiency of deep learning systems. We looked at how architectural techniques—like pruning, quantization, knowledge distillation, and lightweight models such as MobileNet and EfficientNet—help cut down on energy use without sacrificing performance. We also examined how hardware-aware solutions, including neuromorphic chips, FPGAs, and edge AI, can make AI smarter and more power-conscious. Alongside these, algorithm–hardware co-design emerged as a promising direction to simultaneously boost speed and reduce environmental impact[26-27].

Real-world examples, such as TinyML and TinyissimoYOLO, show that it's possible to build powerful AI systems that run on extremely low energy, making them ideal for real-time and portable use. Still, challenges remain. Balancing model accuracy with energy use, and managing the large energy gap between training and inference, are ongoing issues. The future of green AI will also depend on better metrics for measuring energy impact and smarter AutoML tools that can build energy-efficient models automatically.

In the end, making AI sustainable isn't just about improving technology—it's about rethinking how we build, train, and deploy intelligent systems. As AI continues to shape the future, it's essential for researchers, developers, and policymakers to work together to put energy efficiency at the heart of AI design. By taking a thoughtful and collaborative approach, we can ensure that the next generation of AI is not only powerful but also responsible—advancing innovation while respecting the limits of our planet.

REFERENCES

[1] Hidalgo, I., Fenández-de_Vega, F., Ceberio, J., Garnica, O., Velasco, J. M., Cortés, J. C., ... & Díaz, J. (2023). Sustainable Artificial Intelligence Systems: An Energy Efficiency Approach. *Authorea Preprints*.

[2] Mehlin, V., Schacht, S., & Lanquillon, C. (2023). Towards energy-efficient deep learning: An overview of energy-efficient approaches along the deep learning lifecycle. *arXiv preprint arXiv:2303.01980*.

[3] Ranpara, R. (2025). Energy-Efficient Green AI Architectures for Circular Economies Through Multi-Layered Sustainable Resource Optimization Framework. *arXiv preprint arXiv:2506.12262*.

[4] Morrison, J., Na, C., Fernandez, J., Dettmers, T., Strubell, E., & Dodge, J. (2025). Holistically evaluating the environmental impact of creating language models. *arXiv preprint arXiv:2503.05804*.

[5] Ren, S., Tomlinson, B., Black, R. W., & Torrance, A. W. (2024). Reconciling the contrasting narratives on the environmental impact of large language models. *Scientific Reports*, *14*(1), 26310.

[6] Xu, Y., Martínez-Fernández, S., Martinez, M., & Franch, X. (2023). Energy efficiency of training neural network architectures: an empirical study. *arXiv preprint arXiv:2302.00967*.

[7] Vente, T., Wegmeth, L., Said, A., & Beel, J. (2024, October). From clicks to carbon: The environmental toll of recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 580-590).

[8] Tripp, C. E., Perr-Sauer, J., Gafur, J., Nag, A., Purkayastha, A., Zisman, S., & Bensen, E. A. (2024). Measuring the energy consumption and efficiency of deep neural networks: An empirical analysis and design recommendations. *arXiv preprint arXiv:2403.08151*.

[9] Aquino-Brítez, S., García-Sánchez, P., Ortiz, A., & Aquino-Brítez, D. (2025). Towards an Energy Consumption Index for Deep Learning Models: A Comparative Analysis of Architectures, GPUs, and Measurement Tools. *Sensors*, *25*(3), 846.

[10] Qu, X., Aponte, D., Banbury, C., Robinson, D. P., Ding, T., Koishida, K., ... & Chen, T. (2025). Automatic joint structured pruning and quantization for efficient neural network training and compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 15234-15244).

[11] Balaskas, K., Karatzas, A., Sad, C., Siozios, K., Anagnostopoulos, I., Zervakis, G., & Henkel, J. (2024). Hardware-aware dnn compression via diverse pruning and mixed-precision quantization. *IEEE Transactions on Emerging Topics in Computing*, *12*(4), 1079-1092.

[12] Wu, S., Luo, X., Liu, J., & Deng, Y. (2025). Knowledge distillation with adapted weight. *Statistics*, 1-28.

[13] Konstantaropoulos, O., Mallios, T., & Papadopouli, M. (2025). Dynamic Activation with Knowledge Distillation for Energy-Efficient Spiking NN Ensembles. *arXiv preprint arXiv:2502.14023*.

[14] Rehman, Z. U., Hassan, U., Islam, S. U., Gallos, P., & Boudjadar, J. (2025). Energy-Efficient AI for Medical Diagnostics: Performance and Sustainability Analysis of ResNet and MobileNet. *Studies in health technology and informatics*, *327*, 1225-1229.

[15] Vogginger, B., Rostami, A., Jain, V., Arfa, S., Hantsch, A., Kappel, D., ... & Maaß, W. (2024). Neuromorphic hardware for sustainable AI data centers. *arXiv preprint arXiv:2402.02521*.

[16] Mouri Zadeh Khaki, A., & Choi, A. (2025). Optimizing Deep Learning Acceleration on FPGA for Real-Time and Resource-

Efficient Image Classification. *Applied Sciences*, *15*(1), 422.

[17] Peccia, F. N., Pavlitska, S., Fleck, T., & Bringmann, O. (2024, August). Efficient Edge AI: Deploying Convolutional Neural Networks on FPGA with the Gemmini Accelerator. In *2024 27th Euromicro Conference on Digital System Design (DSD)* (pp. 418-426). IEEE.

[18] Fan, H., Ferianc, M., Que, Z., Li, H., Liu, S., Niu, X., & Luk, W. (2022, January). Algorithm and hardware co-design for reconfigurable cnn accelerator. In *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)* (pp. 250-255). IEEE.

[19] Yang, Z., Adamek, K., & Armour, W. (2024). Double-Exponential Increases in Inference Energy: The Cost of the Race for Accuracy. *arXiv preprint arXiv:2412.09731*.

[20] Siddique, A., Basu, K., & Hoque, K. A. (2021, April). Exploring fault-energy trade-offs in approximate DNN hardware accelerators. In *2021 22nd International Symposium on Quality Electronic Design (ISQED)* (pp. 343-348). IEEE.

[21] Yarally, T., Cruz, L., Feitosa, D., Sallou, J., & Van Deursen, A. (2023, May). Uncovering energy-efficient practices in deep learning training: Preliminary steps towards green ai. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)* (pp. 25-36). IEEE.

[22] Douwes, C., & Serizel, R. (2024). From Computation to Consumption: Exploring the Compute-Energy Link for Training and Testing Neural Networks for SED Systems. *arXiv preprint arXiv:2409.05080*.

[23] Clemm, C., Stobbe, L., Wimalawarne, K., & Druschke, J. (2024, June). Towards Green AI: Current status and future research. In *2024 Electronics Goes Green 2024+(EGG)* (pp. 1-11). IEEE.

[24] Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. V. E., ... & Zhukov, L. E. E. (2022, December). Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. In *Doklady mathematics* (Vol. 106, No. Suppl 1, pp. S118-S128). Moscow: Pleiades Publishing.

[25] Hennig, L., Tornede, T., & Lindauer, M. (2024). Towards leveraging automl for sustainable deep learning: A multi-objective hpo approach on deep shift neural networks. *arXiv preprint arXiv:2404.01965*.

[26] Deshwal, V. (2024). *Ethics in artificial intelligence*. GLIMPSE - Journal of Computer Science, 3(2), 14–18

[27] Taluja, A., Kumar, H., Agarwal, V., Tomar, K., & Garg, S. (2024). Transformative applications of blockchain and machine learning in healthcare: A review. *GLIMPSE – Journal of Computer Science, 3*(2), 34–38.

ABOUT THE AUTHORS

Vaishali Deshwal received the M.Tech. degree in Computer Science and engineering from Dr. A.P.J. Abdul Kalam Technical University (AKTU), Uttar Pradesh, India. Her research interests are Artificial Intelligence, Machine Learning and deep learning. She is currently working as an Assistant Professor at AKGEC Ghaziabad.

Anuradha Bharti received the M.Tech degree in VLSI Design (Electronics and Communication Engineering) from IGDTUW, New Delhi, India. Her research interests are Artificial Intelligence, Machine Learning and deep learning. She is currently working as an Assistant Professor at MIT Meerut.