Choosing to Be Green: Advancing Green AI via Dynamic Model Selection

Emilio Cruciani[1], Roberto Verdecchia[2]

[1]European University of Rome, Italy
[2]University of Florence, Italy

Abstract

Artificial Intelligence is increasingly pervasive across domains, with ever more complex models delivering impressive predictive performance. This fast technological advancement however comes at a concerning environmental cost, with state-of-the-art models—particularly deep neural networks and large language models—requiring substantial computational resources and energy. In this work, we present the intuition of *Green AI dynamic model selection*, an approach based on dynamic model selection that aims at reducing the environmental footprint of AI by selecting the most sustainable model while minimizing potential accuracy loss. Specifically, our approach takes into account the inference task, the environmental sustainability of available models, and accuracy requirements to dynamically choose the most suitable model. Our approach presents two different methods, namely *Green AI dynamic model cascading* and *Green AI dynamic model routing*. We demonstrate the effectiveness of our approach *via* a proof of concept empirical example based on a real-world dataset. Our results show that *Green AI dynamic model selection* can achieve substantial energy savings (up to ≈25%) while substantially retaining the accuracy of the most energy greedy solution (up to ≈95%). As conclusion, our preliminary findings highlight the potential that hybrid, adaptive model selection strategies withhold to mitigate the energy demands of modern AI systems without significantly compromising accuracy requirements.

Keywords

Green AI, Green Model Selection, Model Cascading, Model Routing, Energy Efficiency

1. Introduction

The popularization of AI models, ranging from simple classifiers to complex large language models, has taken the world by storm. With the widespread and evergrowing adoption of AI and all the benefits it implied, the environmental resources needed to power such models is also surging, and this trend is no longer negligible [1]. To contrast the invisible impact that AI is having on the limited resources of our planet, the field of *Green AI* [2] rapidly developed, and has seen a considerable growth in the most recent years [3]. By quoting the words of Schwartz et al. [2], Green AI is a field of AI research that yields novel results while considering its computational cost and encouraging the reduction of resources spent. Under the research field of Green AI fall a plethora of heterogeneous solutions, ranging from *ad hoc* hyperparameter tuning to trade-offs between model accuracy and energy consumption, energy-aware model deployment strategies, data-centric techniques [3], and software engineering approaches [4]. Despite the wide array of Green AI solutions that have been conceived to date, Green AI approaches based on selecting different models by factoring in their energy consumption results to date to be an uncharted territory. In this work, we explore the potential that dynamic model selection based on the task at hand, model validation accuracy, and energy efficiency can have on AI environmental sustainability. More specifically, we present the very idea of *Green AI dynamic model selection* by presenting two methods that lend their core intuition from the related literature on dynamic model selection, namely *model cascading* and *model routing* [5, 6]. Intuitively the first method we present, namely *Green AI dynamic model cascading*, subsequently invokes different models from less to more energy greedy till a sufficient level of prediction confidence is achieved. The second method instead,

Green-Aware AI 2025 @ ECAI 2025
emilio.cruciani@unier.it (E. Cruciani); roberto.verdecchia@unifi.it (R. Verdecchia)
https://sites.google.com/view/emiliocruciani (E. Cruciani); https://robertoverdecchia.github.io (R. Verdecchia)
0000-0002-4744-5635 (E. Cruciani); 0000-0001-9206-6637 (R. Verdecchia)

referred to as *Green AI dynamic model routing*, selects exclusively the most efficient model to be used by considering the predicted model accuracy for the input task and the energy efficiency of the model. In addition to a formal presentation of both energy-aware dynamic model cascading and routing methods, we also document the results of an empirical proof of concept evaluation that we execute to showcase the viability of our intuition. The empirical proof of concept, which should be by no means be interpreted as exhaustive or conclusive, is based on an exemplary classification task relying on the widely utilized scikit-learn and keras Python libraries, two AI models of different complexity, and an *ad hoc* implementation of the energy-aware dynamic model cascading and routing methods. The preliminary results we collect point to the potential that dynamic model selection methods have to achieve Green AI. As a complementary portion of our contribution, we also delve into reflecting on the various nuances, potential challenges, and benefits that may arise when building further onto the *Green AI dynamic model selection* approach.

2. Related Work

The field of Green AI has experienced a swift growth in popularization in the past few years and is increasingly becoming an established discipline [3]. The rise of interest in the topic could have stemmed from diverse effort of the research community to quantify the environmental impact of AI, ranging from generic high level figures of CO_2 emissions [7] to fine-grained measurements of specific models, e.g., deep learning ones [8]. The overall picture studies of this nature draw is consistent, the environmental impact of AI is an issue that needs to be addressed. Answering such call, numerous research endeavors focused on improving the energy efficiency and environmental sustainability of AI. The proposed solutions to achieve Green AI are heterogeneous and span a wide range of approaches.

A family of Green AI techniques focuses on designing AI models by factoring in their energy consumption [9], e.g., by improving model execution times [10], optimizing models for specific hardware components [11], compressing models [12], or seeking more energy efficient model implementations [13, 14]. In contrast, another family of Green AI techniques focus instead on the *a posteriori* optimization of models *via* hyperparameter fine-tuning [15, 16, 17, 18, 19].

As a green AI research area that might somewhat be more related to the topic considered in this research, a set of studies investigated how different model deployment strategies can impact their energy consumption. Contributions of this type consider solutions such as inference on the edge [20, 21], model deployment in virtualized cloud fog networks [22], and distributed machine learning [23].

Taking a different standpoint, other Green AI approaches consider instead exclusively the data the models are trained with, rather than the design of the algorithm themselves, a discipline referred to as *Data Centric Green AI* [24, 25, 26, 27, 28].

All of the above mentioned areas of Green AI research result orthogonal to the topic considered in this study, as they focus exclusively on the optimization of one specific model. In contrast, in our work, we do not aim to improve the energy efficiency of a single model, but rather to select the most fitting one (or set thereof) by keeping AI energy efficiency in mind. To the best of our knowledge, this topic has to date only marginally be explored in the related literature.

The work of Nijkamp et al. [29] is potentially the one that is most closely related to the approaches presented in this contribution. Nijkamp et al. consider an ensemble learning context within the text processing domain, where a subset of pre-trained and trained models are selected for inference and results are merged *a posteriori*. The selection of models can be executed either statically, where an optimal subset of models is chosen for an entire domain considered, or dynamically, i.e., an optimal subset is selected for every queried property within the domain. Differently from such approach, we do not focus on ensemble learning, and trigger the inference of multiple models only in the case of energy-aware model cascading (see also Section 3.1).

In another work that considers ensemble learning, Omar et al. [30] consider the impact that three different design decisions for ensemble learning, namely ensemble size, fusion methods, and partitioning methods, can have on energy consumption. As for the previous study, our contribution differs by not

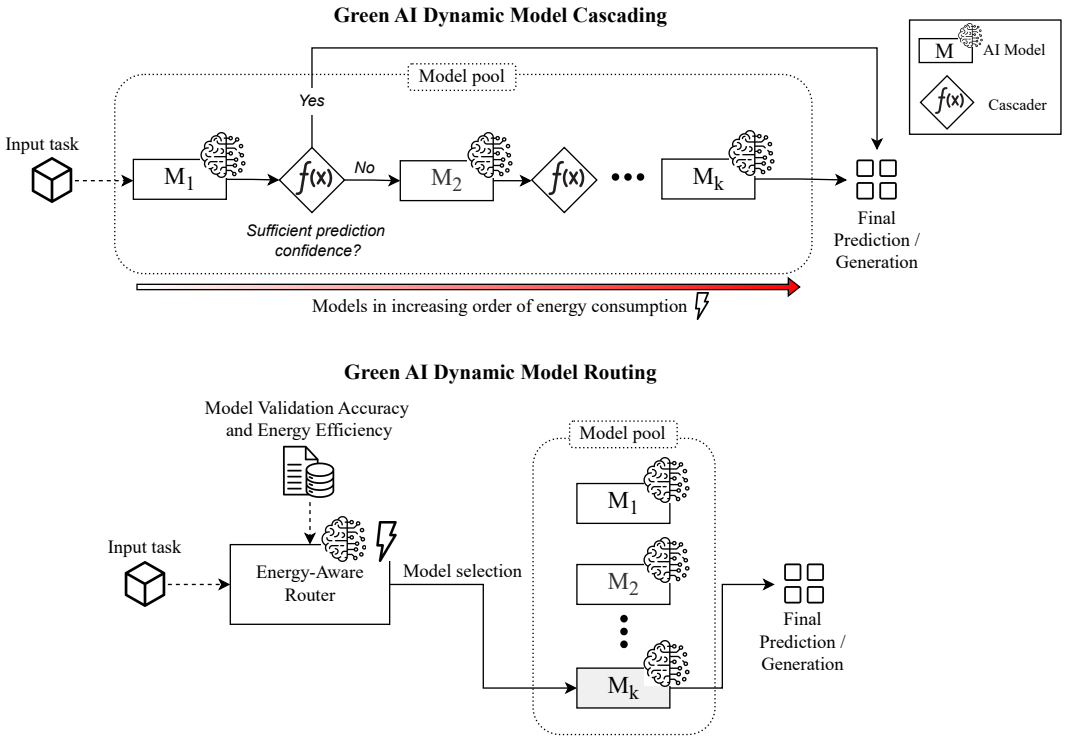

Figure 1: Overview of the dynamic model selection methods for energy efficient inference.

considering the context of ensemble learning, but rather dynamic model selection for energy efficiency. A related work by Matathammal et al. [31] presents EdgeMLBalancer, an approach that balances resource utilization *via* dynamic model switching in the context of edge-devices. In contrast to such work, our dynamic model selection approach is not concerned with the allocation between different resource-constrained edge devices, is not specific to real-time object detection, and is not based on the MAPE-K Feedback Loop to conduct the selection of models (see also Section 3).

As mention of another branch of related work, at the core of this study lies a plethora of foundational research endeavors conducted in the realm of model selection [32, 33, 34], with particular emphasis on approaches based on model cascading and model routing [5, 6]. Our contribution builds upon such literature, by borrowing the intuition of such approaches to embed environmental sustainability as part of the dynamic model selection process.

3. Green AI Dynamic Model Selection Methods

In this section we present two methods to build energy aware classification models *via* cascading and routing. An overview of the proposed dynamic model selection methods for energy efficiency are depicted in Figure 1 and are further described below.

Intuitively the first strategy, named *Green AI dynamic model cascading*, is based on a cascading methods where models at an increasing level of energy consumptions are invoked subsequently when required. The second method instead, named *Green AI dynamic model routing*, is based on an upfront energy-aware router component that selects the best suited model based on the task at hand, the validation accuracy of models, and their energy efficiency.

As a note on terminology, in the following documentation both methods we present consider a labeled dataset $D = (\mathbf{x}_i, y_i)_{i=1\ldots,n}$ of n data points with $\mathbf{x}_i \in X$ being the feature vector of data point i and $y_i \in Y$ being its label. Also, we let M be a classification model, namely a function $M : X \to Y$

that given an input $\mathbf{x} \in X$ predicts its class as $\hat{y} = M(\mathbf{x}) \in Y$.

3.1. Green AI Dynamic Model Cascading

For the cascading model $C : X \to Y$, we need: a sequence of $k \geq 2$ models $M_1, M_2, \ldots, M_k$, ordered by increasing energy consumption and typically increasing complexity and accuracy; a family of prediction confidence functions $\alpha_i(\mathbf{x}) : X \times Y \to [0, 1]$, depending on each model M_i; a parameter $\epsilon \in [0, 1]$ to control the confidence tolerance.

At inference time, for an input instance $\mathbf{x}$, the cascading mechanism proceeds as described in Algorithm 1. In particular, the cascading model evaluates the first model M_1 and obtains both the predicted label $M_1(\mathbf{x})$ and its confidence score $\alpha_1(\mathbf{x})$. If the confidence satisfies $\alpha_1(M_1(\mathbf{x})) \geq 1 - \epsilon$ for some $\epsilon \in (0, 1)$, we accept the prediction and terminate. Otherwise, we move to the next model M_2 and repeat the procedure. This continues until either a model M_i produces a sufficiently confident prediction or the first $k - 1$ models are exhausted, in which case we use the last model M_k as a fallback.

Algorithm 1 Energy-Aware Cascading Inference

Require: Instance $\mathbf{x}$

for $i = 1$ to $k - 1$ **do**
  if $\alpha_i(\mathbf{x}) \geq 1 - \epsilon$ **then**
    return $M_i(\mathbf{x})$
return $M_k(\mathbf{x})$

3.2. Green AI Dynamic Model Routing

An alternative method to reduce inference cost is to directly learn a routing function that selects, for each input, the most appropriate model in terms of energy-accuracy tradeoff. In this case, we define a routing model $R : X \to \{1, \ldots, k\}$, which maps each input instance to one of the k available models $M_1, M_2, \ldots, M_k$, again ordered by increasing energy consumption.

At inference time, for a given input $\mathbf{x} \in X$, the routing model selects an index $i = O(\mathbf{x})$ and returns the prediction $M_i(\mathbf{x})$, as described in Algorithm 2.

Algorithm 2 Energy-Aware Routing Inference

Require: Instance $\mathbf{x}$

$i \leftarrow O(\mathbf{x})$
return $M_i(\mathbf{x})$

The goal is to train a model O so that it selects the least energy-consuming model capable of producing a correct prediction. To train the routing model, we assume access to a validation dataset $D_{\text{val}} \subseteq D$, and we construct training labels for the routing task as follows: for each input $\mathbf{x} \in D_{\text{val}}$, we identify the lowest-index model M_i that correctly classifies $\mathbf{x}$ (i.e., $M_i(\mathbf{x}) = y$); if no such model exists, we select the lowest-index model regardless of accuracy to minimize energy cost. This label construction assumes that we can evaluate the correctness of each model on the validation set and that the energy consumption associated with each model is known.

Formally, we define an oracle routing function $O^* : X \to \{1, \ldots, k\}$ that given $\mathbf{x}$ returns:

$$O^*(\mathbf{x}) = \begin{cases} \min \{i \in \{1, \ldots, k\} \ : \ M_i(\mathbf{x}) = y\} & \text{if } \exists i \text{ s.t. } M_i(\mathbf{x}) = y, \\ 1 & \text{otherwise.} \end{cases}$$

We then train the model O to approximate the oracle O^*, using standard classification techniques.

4. Empirical Proof of Concept

In this section, we demonstrate the advantages of our approach through a concrete example, comparing the performance and energy consumption of our dynamic model selection classifiers with those of their basic components. As described in Section 3, we consider a cascading model C and a routing model R.

We consider the standard multi-class classification task on the scikit-learn `digits` dataset[1]. The dataset consists of 1797 8x8 grayscale images of hand-written digits (0-9). We use 60% of the dataset for training, a 20% for validation, and 20% for testing. The validation set is only used by the routing model R to training the oracle, while it is not used by the cascading model C for a fair comparison.

For simplicity, in this proof of concept we use only two components for each of the two models: a shallow (depth 5) decision tree G (the simpler and greener model), and a deep (5 hidden layers of decreasing sizes) feedforward neural network A (the more accurate and energy-costly model). The confidence score α_G used for the cascading model C corresponds to the fraction of training samples in the reached leaf node that belong to the predicted class. We use $\varepsilon = 0.2$ as a parameter for C, i.e., for an instance $\mathbf{x}$ we use the prediction of G when the confidence score $\alpha_G(\mathbf{x})$ is at least $1 - \varepsilon = 0.8$, or the prediction of A otherwise. As an oracle for the routing model R we use a logistic regressor with balanced class-weights (to account for potentially imbalanced classes coming from the predictions of G and A in the validation set).

For each of the competitors, we measure its *accuracy* (i.e., fraction of correctly classified instances), *total prediction time* (measured on the whole test set, in ms), and *energy consumption* (measured on the whole test set, in $\mu\mathrm{W\,h}$). Moreover, for our methods C and R we also report the fraction of predictions performed by the simpler component G and the time overhead required for the model selection procedure itself[2]. The running time is measured using the python function `time.perf_counter()`. The energy consumption is estimated using the python package *CodeCarbon*[3]. In order to smooth the time and energy measurements, they are averaged over 1 000 predictions of the entire test set. The example is run on a MacBook Air (M1, 16GB Memory) and the measurements are reported in Table 1.

Table 1
Performance and energy consumption of classifiers

Classifier	Fraction of G	Overhead (ms)	Accuracy	Time (ms)	Energy ($\mu\mathrm{W\,h}$)
(G) Decision Tree	1.00	0.00	0.73	0.13	0.13
(A) Neural Network	0.00	0.00	0.98	37.44	40.80
(C) **Cascading**	0.65	0.22	0.92	29.48	32.12
(R) **Routing**	0.62	0.19	0.88	33.13	30.83

We observe that our hybrid methods C and R, respectively based on cascading and routing, are effective in balancing accuracy and energy consumption relative to their components, the lightweight model G and the accurate, energy-intensive model A. Importantly, the computational overhead introduced by the model selection process is negligible, adding only $\approx$0.2 ms to an average inference time of $\approx$30 ms. Quantitatively, the cascading method C retains 94.7% of the accuracy of A (a reduction of only 0.05), while improving inference speed by 21.25% (7.95 ms) and reducing energy consumption by 21.27% (8.68 $\mu\mathrm{W\,h}$). The routing method R trades slightly more accuracy, retaining 89.7% of A's performance (a reduction of 0.1), for even greater energy efficiency, cutting consumption by 24.44% (9.97 $\mu\mathrm{W\,h}$) and improving speed by 11.51% (4.31 ms). These results highlight the potential of hybrid inference strategies to deliver substantial energy savings with only modest accuracy degradation.

[1] https://scikit-learn.org/1.5/auto_examples/datasets/plot_digits_last_image.html
[2] The energy overhead is not reported because it is minimal and difficult to measure accurately.
[3] https://codecarbon.io

5. Discussion

In this section, we discuss the key aspects we deem paramount to be considered while further developing the idea of Green AI model selection. More specifically, we cover aspects regarding the generalizability of our intuition, potential impediments connected to porting the presented techniques to practice, and other nuances that may arise while further developing Green AI dynamic model selection.

On the generalizability to other tasks. While our study focuses on classification, the proposed approach can in principle be extended to other inference paradigms, including generative tasks, e.g., text generation and image synthesis, and large language models [35, 36]. However, cascading in such settings poses new challenges: Unlike classification, generation tasks often produce variable-length outputs and lack a clear, standardized notion of "confidence", making it harder to decide when to accept early outputs. Furthermore, in autoregressive models, energy cost and quality are tightly coupled over long sequences, which complicates dynamic routing or early termination. Designing confidence surrogates or lightweight proxies for generation quality is an open and non-trivial research direction [37].

On the consumption of the oracle. Routing strategies rely on a pre-trained oracle that predicts the most suitable model to use for a given input. While our results show that routing can outperform static selection in terms of energy efficiency, this advantage must be weighed against the energy and latency overhead introduced by the oracle itself. Although this cost can often be amortized, especially when the oracle is lightweight compared to the target models, it is nonetheless a relevant factor in real-world deployments and should be included in life-cycle assessments [38].

On the precision-energy consumption tradeoff. A central trade-off in energy-aware inference lies in balancing precision and energy consumption [39, 40]. Tuning the confidence threshold ϵ in cascading or adjusting the routing oracle's decision boundary directly affects both the fraction of queries routed to low-cost models and the overall accuracy. This trade-off is highly application-dependent: For some critical tasks even minor drops in accuracy may be unacceptable, while for others tolerating occasional misclassifications may be worthwhile for substantial energy savings.

On the specificity of the task at hand. The benefits of dynamic model selection depend heavily on the characteristics of the task. Tasks with highly skewed input difficulty, where many inputs are easily handled by simple models, stand to benefit most from cascading or routing [41]. In contrast, tasks that are uniformly hard may offer little opportunity for savings, as most queries will require the most complex model regardless. This suggests that per-task calibration or meta-learning strategies could further enhance the adaptability of energy-aware approaches. Specific task might also dictate the dynamic model selection strategy. For example, by considering image generation quality and energy consumption [42], model routing might be to date the only solution applicable.

On the energy cost of loading models. Energy measurement methodologies must carefully account for the cost of loading models into memory, especially when switching between models incurs overhead due to I/O or hardware constraints [43]. In scenarios where models are not kept in memory persistently, the benefit of selecting a low-cost model may be offset by the loading cost. This points to the importance of deployment-aware design: in serverless or constrained edge environments, keeping a subset of models "warm" may be necessary for real energy savings.

On the development cost of maintaining models updated. The practical deployment of multi-model systems introduces non-trivial maintenance costs. Each model in the pool must be monitored, updated, and re-validated to cope with data drift, distribution shifts, or evolving application requirements. This adds complexity to the life-cycle management of the AI system and raises questions about the

long-term cost–benefit balance. Approaches such as continual learning may help reduce redundancy and maintain performance with a smaller, more efficient model pool [44].

On the optimization of model carbon footprint. In this work we primarily focused on model energy consumption, i.e., the raw energy consumed by the models. In a broader perspective however, the carbon footprint of the models, i.e., the total amount of greenhouse gases required to produce the energy consumed by the models, might be instead the primary metric we want to optimize for [45]. Distinguishing between energy consumption and carbon footprint is necessary in contexts where models are not powered by the same energy grid, e.g., in a distributed deployment scenario. By considering the core intuition behind the presented methods, we argue that these can be effortlessly adapted by considering the measured carbon footprint of the models instead of the energy consumption considered in this contribution.

6. Conclusions and Future Work

With the great technological advancements AI brought, its evergrowing popularization, and its non-negligible environmental impact, we are responsible to conceive novel solutions that preserve technological advancements while optimizing environmental sustainability. In this work we present the concept of *Green AI dynamic model selection*, which lives at the intersection of dynamic model selection and model environmental sustainability. The core contribution presented in this study is twofold, by proposing two distinct techniques through which Green AI can be achieved by dynamically selecting models according to their environmental sustainability. The two methods are referred to in this work as *Green AI Dynamic Model Cascading* and *Green AI Dynamic Model Routing*. To support our documented intuition, we report an empirical proof of concept, which showcases in practical terms the potential of the proposed idea. While the results of our experimentation are by no means to be considered as generalizable or conclusive, they support us in arguing that *Green AI dynamic model selection* is a Green AI strategy that is worth to be further investigated. To further support our intuition, we also further delve into speculating on the core concepts, impediments, and benefits of *Green AI dynamic model selection*, in the hope that our contribution can support other researchers in making AI greener.

References

[1] C.-J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai, et al., Sustainable ai: Environmental implications, challenges and opportunities, Proceedings of Machine Learning and Systems 4 (2022) 795–813.

[2] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green ai, Communications of the ACM 63 (2020) 54–63.

[3] R. Verdecchia, J. Sallou, L. Cruz, A systematic review of green ai, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13 (2023) e1507.

[4] L. Cruz, J. P. Fernandes, M. H. Kirkeby, S. Martínez-Fernández, J. Sallou, H. Anwar, E. Barba Roque, J. Bogner, J. Castaño, F. Castor, et al., Greening ai-enabled systems with software engineering: A research agenda for environmentally sustainable ai practices, ACM SIGSOFT Software Engineering Notes 50 (2025) 14–23.

[5] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, 2001, pp. I–I. doi:10.1109/CVPR.2001.990517.

[6] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87. doi:10.1162/neco.1991.3.1.79.

[7] A. Lacoste, A. Luccioni, V. Schmidt, T. Dandres, Quantifying the carbon emissions of machine learning, arXiv preprint arXiv:1910.09700 (2019).

[8] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for modern deep learning research, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13693–13696.
[9] B. D. Rouhani, A. Mirhoseini, F. Koushanfar, Delight: Adding energy dimension to deep neural networks, in: Proceedings of the 2016 International Symposium on Low Power Electronics and Design, 2016, pp. 112–117.
[10] E. García-Martín, N. Lavesson, H. Grahn, E. Casalicchio, V. Boeva, Energy-aware very fast decision tree, International Journal of Data Science and Analytics 11 (2021) 105–126.
[11] K. Rungsuptaweekoon, V. Visoottiviseth, R. Takano, Evaluating the power efficiency of deep learning inference on embedded gpu systems, in: 2017 2nd international conference on information technology (INCIT), IEEE, 2017, pp. 1–5.
[12] H. Yang, Y. Zhu, J. Liu, Energy-constrained compression for deep neural networks via weighted sparse projection and layer input masking, arXiv preprint arXiv:1806.04321 (2018).
[13] N. Marini, L. Pampaloni, F. Di Martino, R. Verdecchia, E. Vicario, Green AI: Which Programming Language Consumes the Most? , in: 2025 IEEE/ACM 9th International Workshop on Green and Sustainable Software (GREENS), IEEE Computer Society, Los Alamitos, CA, USA, 2025, pp. 12–19. URL: https://doi.ieeecomputersociety.org/10.1109/GREENS66463.2025.00007. doi:10.1109/GREENS66463.2025.00007.
[14] S. Georgiou, M. Kechagia, T. Sharma, F. Sarro, Y. Zou, Green ai: Do deep learning frameworks have different costs?, in: Proceedings of the 44th International Conference on Software Engineering, 2022, pp. 1082–1094.
[15] L. H. P. de Chavannes, M. G. K. Kongsbak, T. Rantzau, L. Derczynski, Hyperparameter power impact in transformer language model training, in: Proceedings of the second workshop on simple and efficient natural language processing, 2021, pp. 96–118.
[16] M. Magno, M. Pritz, P. Mayer, L. Benini, Deepemote: Towards multi-layer neural networks in a low power wearable multi-sensors bracelet, in: 2017 7th IEEE international workshop on advances in sensors and interfaces (IWASI), IEEE, 2017, pp. 32–37.
[17] M. Barlaud, F. Guyard, Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai, in: 2020 25th international conference on pattern recognition (ICPR), IEEE, 2021, pp. 1566–1573.
[18] D. Stamoulis, T.-W. R. Chin, A. K. Prakash, H. Fang, S. Sajja, M. Bognar, D. Marculescu, Designing adaptive neural networks for energy-constrained image classification, in: 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), IEEE, 2018, pp. 1–8.
[19] M. Sponner, B. Waschneck, A. Kumar, Adapting neural networks at runtime: Current trends in at-runtime optimizations for deep learning, ACM Computing Surveys 56 (2024) 1–40.
[20] S. Gondi, V. Pratap, Performance and efficiency evaluation of asr inference on the edge, Sustainability 13 (2021) 12392.
[21] X. Yang, S. Hua, Y. Shi, H. Wang, J. Zhang, K. B. Letaief, Sparse optimization for green edge ai inference, Journal of communications and information networks 5 (2020) 1–15.
[22] B. A. Yosuf, S. H. Mohamed, M. M. Alenazi, T. E. El-Gorashi, J. M. Elmirghani, Energy-efficient ai over a virtualized cloud fog network, in: Proceedings of the twelfth ACM international conference on future energy systems, 2021, pp. 328–334.
[23] B. Güler, A. Yener, Energy-harvesting distributed machine learning, in: 2021 IEEE international symposium on information theory (ISIT), IEEE, 2021, pp. 320–325.
[24] R. Verdecchia, L. Cruz, J. Sallou, M. Lin, J. Wickenden, E. Hotellier, Data-centric green ai an exploratory empirical study, in: 2022 international conference on ICT for sustainability (ICT4S), IEEE, 2022, pp. 35–45.
[25] S. Salehi, A. Schmeink, Data-centric green artificial intelligence: A survey, IEEE Transactions on Artificial Intelligence 5 (2023) 1973–1989.
[26] M. Alswaitti, R. Verdecchia, G. Danoy, P. Bouvry, J. Pecero, Training green ai models using elite samples, IEEE Transactions on Sustainable Computing (2025).
[27] S. Kumar, S. Datta, V. Singh, S. K. Singh, R. Sharma, Opportunities and challenges in data-centric

ai, IEEE Access (2024).
[28] H. Järvenpää, P. Lago, J. Bogner, G. Lewis, H. Muccini, I. Ozkaya, A synthesis of green architectural tactics for ml-enabled systems, in: Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Society, 2024, pp. 130–141.
[29] N. Nijkamp, J. Sallou, N. van der Heijden, L. Cruz, Green ai in action: Strategic model selection for ensembles in production, in: Proceedings of the 1st ACM International Conference on AI-Powered Software, 2024, pp. 50–58.
[30] R. Omar, J. Bogner, H. Muccini, P. Lago, S. Martínez-Fernández, X. Franch, The more the merrier? navigating accuracy vs. energy efficiency design trade-offs in ensemble learning systems, arXiv preprint arXiv:2407.02914 (2024).
[31] A. Matathammal, K. Gupta, L. Lavanya, A. V. Halgatti, P. Gupta, K. Vaidhyanathan, Edgemlbalancer: A self-adaptive approach for dynamic model switching on resource-constrained edge devices, arXiv preprint arXiv:2502.06493 (2025).
[32] D. Anderson, K. Burnham, Model selection and multi-model inference, Second. NY: Springer-Verlag 63 (2004) 10.
[33] C. J. Merz, Dynamical selection of learning algorithms, in: Learning from Data: Artificial Intelligence and Statistics V, Springer, 1996, pp. 281–290.
[34] W. Armstrong, P. Christen, E. McCreath, A. P. Rendell, Dynamic algorithm selection using reinforcement learning, in: 2006 international workshop on integrating ai and data mining, IEEE, 2006, pp. 18–25.
[35] S. Bae, J. Ko, H. Song, S.-Y. Yun, Fast and robust early-exiting framework for autoregressive language models with synchronized parallel decoding, arXiv preprint arXiv:2310.05424 (2023).
[36] H. Rahmath P, V. Srivastava, K. Chaurasia, R. G. Pacheco, R. S. Couto, Early-exit deep neural network-a comprehensive survey, ACM Computing Surveys 57 (2024) 1–37.
[37] S. Xu, Y. Lu, G. Schoenebeck, Y. Kong, Benchmarking llms' judgments with no gold standard, arXiv preprint arXiv:2411.07127 (2024).
[38] B. Zhang, A. Davoodi, Y.-H. Hu, A mixture of expert approach for low-cost customization of deep neural networks, arXiv preprint arXiv:1811.00056 (2018).
[39] A. E. Brownlee, J. Adair, S. O. Haraldsson, J. Jabbo, Exploring the accuracy–energy trade-off in machine learning, in: 2021 IEEE/ACM International Workshop on Genetic Improvement (GI), IEEE, 2021, pp. 11–18.
[40] M. Kim, W. Saad, M. Mozaffari, M. Debbah, On the tradeoff between energy, precision, and accuracy in federated quantized neural networks, in: ICC 2022-IEEE International Conference on Communications, IEEE, 2022, pp. 2194–2199.
[41] M. Damani, I. Shenfeld, A. Peng, A. Bobu, J. Andreas, Learning how hard to think: Input-adaptive allocation of lm computation, arXiv preprint arXiv:2410.04707 (2024).
[42] G. Bertazzini, C. Albisani, D. Baracchi, D. Shullani, R. Verdecchia, The Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation, arXiv preprint arXiv:2506.17016 (2025).
[43] C. Ji, F. Wu, Z. Zhu, L.-P. Chang, H. Liu, W. Zhai, Memory-efficient deep learning inference with incremental weight loading and data layout reorganization on edge systems, Journal of Systems Architecture 118 (2021) 102183.
[44] F. Majidi, F. Khomh, H. Li, A. Nikanjam, An efficient model maintenance approach for mlops, arXiv preprint arXiv:2412.04657 (2024).
[45] I. Wang, N. Ardalani, M. Elhoushi, D. Jiang, S. Hsia, E. Sumbul, D. Mahajan, C.-J. Wu, B. Acun, Carbon aware transformers through joint model-hardware optimization, arXiv preprint arXiv:2505.01386 (2025).