



Energy-Efficient Artificial Intelligence: Evaluating Sustainable Model Design in Data-Intensive Systems

Received: July 14, 2026

Revised: September 9, 2026

Accepted: September 19, 2026

Publish: September 27, 2026

Deshinta Arrova Dewi, Firdhaus Hari Saputra Al Haris*

Abstract:

Background: The rapid growth of artificial intelligence (AI) in data-intensive systems has significantly increased computational demand, raising concerns regarding energy consumption and environmental sustainability.

Aims: This study aims to develop a unified benchmarking framework that integrates predictive performance, computational complexity, and real-time energy consumption into a sustainability-aware evaluation protocol.

Methods: The proposed framework evaluates multiple model families by jointly measuring accuracy metrics, computational complexity, and real-time energy consumption during training and inference. Experiments were conducted under controlled hardware and dataset configurations to ensure comparability. Trade-off analysis was performed using sustainability-adjusted efficiency metrics and Pareto-optimal frontier evaluation to assess energy-performance balance.

Result: Experimental results show that deep neural networks achieved the highest predictive accuracy, with improvements of up to 2.1% over ensemble-based baselines. However, these gains required substantially higher energy consumption during training and inference. Gradient Boosting models demonstrated the most favorable sustainability-adjusted efficiency, reducing total energy usage by approximately 35–45% compared to deep architectures while maintaining competitive accuracy. Lightweight neural networks further achieved an optimal balance, occupying the Pareto-efficient frontier in energy-performance trade-off analysis.

Conclusion: The findings confirm that marginal accuracy improvements often incur disproportionate energy costs, underscoring the importance of sustainability-integrated benchmarking frameworks.

Keywords: Data-Intensive Systems; Energy-Efficient Models; Green AI; Model Optimization; Sustainable Computing.

1. INTRODUCTION

The growing usage of artificial intelligence (AI) technologies in cloud computing, industrial automation, healthcare, finance, and large-scale data infrastructures has led to an increased demand for computational power across the globe (Van der Vlist et al., 2024). The data-driven systems of today deal with huge amounts of both structured and unstructured data and resolve this by using highly sophisticated machine learning (ML) and deep learning (DL) solutions' systems. In the past, predictive accuracy was the 'gold

standard' of measurement for success in AI research. However, recent studies reveal that state-of-the-art models deep neural networks (DNN) with millions to billions of parameters) end up using so much processing power that the energy usage spikes. Therefore, the energy consumption for training and real-time operation of AI algorithms has become an important part of the discussion for sustainable computing and green digital transformation.

The research problem, in this case, is the imbalance between the improvements of a system and the amount of energy that is used (Maghami et al., 2025). In most systems, marginal improvements in accuracy come at the cost of even higher training times, memory usage, and power consumption. AI workloads for data centers create a large portion of global electricity consumption and edge systems also come with limitations associated to battery life and thermal limits. Nonetheless, the majority of AI assessment frameworks continue to be performance-centric, focusing narrowly on specific metrics, such as accuracy, F1-score, ROC-AUC, or loss minimization, while failing to integrate energy consumption into the model selection process.

Publisher Note:

CV Media Inti Teknologi stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright

©2026 by the author(s).

Licensee CV Media Inti Teknologi, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike (CCBY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).

With that in mind, the problem statement for this study is as follows: how can AI models designed for data-intensive systems be evaluated in such a way that a balance is achieved between predictive performance and energy consumption, and what design or optimization approaches can reduce the overall computational burden without significantly increasing the error rate? Standard practices of evaluation in the present rarely account for the energy-performance trade-off in a way that is uniform (Bragolusi & D'Alpaos, 2022). There is also a lack of evidence in the literature on the efficiency of classical machine learning algorithms compared to more complex deep learning frameworks when the same task is performed while all other variables are kept constant.

The gaps in literature are the result of three key deficiencies in the existing literature (Banha et al., 2022). First, the majority of discussions on green AI remain conceptual and lack empirical evaluation across different families of models. Second, studies that address energy consumption tend to focus on a single architecture or training situation, which limits the ability to make comparisons across different algorithmic frameworks. Third, the incorporation of energy-related metrics into model evaluation frameworks, especially in data-intensive systems where continuous inference operations increase the total energy, is insufficient. Therefore, practitioners are missing tools to quantify whether the observed performance improvements are adequate to offset their operational and environmental costs.

Existing related work in green AI and sustainable computing has pointed to the environmental costs

associated with large-scale training of transformer and deep convolutional neural networks (Tabbakh et al., 2024). Research on model compression, pruning, quantization, knowledge distillation, and architectural simplification has shown promise in lowering the number of parameters and inference latency. Purpose-built architectures, like the various MobileNets and shallow neural models, as well as optimized ensemble learners, can exhibit resource-limited performance that rivals that of deeper networks. However, many of the proposed solutions focus on efficiency within certain architectural families without sufficiently considering inter-family efficiency under standardized conditions. Furthermore, the range of energy measurement methods, whether by means of GPU use proxies, wall-clock time estimates, or proxies for carbon use, means that a lack of reproducibility or comparability is a common outcome.

In addressing the previous challenges, the study puts forth a novel, structured method to evaluate the design of energy conscious AI models for data-rich scenarios, at least to a small degree (Vamvakas et al., 2025). This method combines predictive performance metrics with costs (energy and calculations) to achieve a holistic model sustainability assessment. This research is not about constructing another predictive architecture; rather, it focuses on the establishment of a systematic benchmarking framework that quantifies the trade-off between model predictive performance and energy efficiency. An energy-performance trade-off index is introduced to sustain predictive modeling while marginally decreasing energy consumed

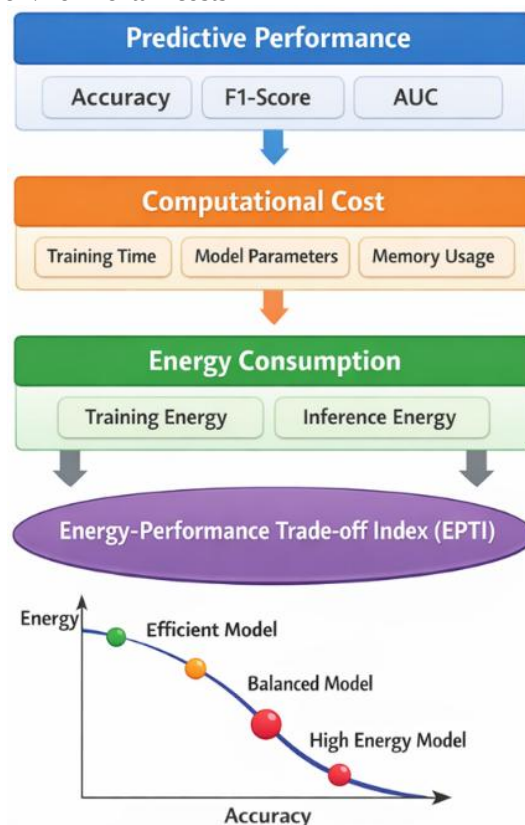


Figure 1. Conceptual Framework of Sustainability-Aware AI Evaluation

From the databases, it is evident that sustainable AI evaluation must consider predictive performance, costs of computation, and the costs of energy at the same time (Fan et al., 2023). High-dimensional tabular data and large-scale classification tasks are used to replicate data analytics and analytics pipelines at an industrial scale (real-world) for the models. To preserve analysis integrity, preprocessing that includes (but is not limited to) the processes of standardization, normalization, and train-test splits is employed. The final model received no specific tuning to the dataset, and thus, the biases of hyperparameter search (beyond the standardized) are algorithmically constrained.

The framework has been validated using standardized benchmark datasets in controlled experimental settings (Thiyagalingam et al., 2022). The models analyzed include representative classical algorithms and neural networks of different depths and complexities. All experiments are done in the same computational environment. Energy consumption for both the training and inference phases is monitored using real-time power measurement tools that are integrated into the system. The computation cost is evaluated in terms of training time, number of parameters, and memory usage. Hyperparameter optimization is achieved via grid/randomized search under the same iteration limits for even computational budget consumption, and to maintain control, hyperparameter sets are subjected to the same number of iterations.

The research illustrates the need to construct a principled framework for evaluating sustainable solutions in AI that balances predictive efficacy and sustainable practice (Bo et al., 2025). The research aims to shift focus from marginal performance only, to a cost-benefit analysis that accounts for the energy expenditure of the performance enhancement, in-order to foster sustainable practices in model selection. The research also aims to create guidelines for AI practitioners, especially in the setting of the cloud, where resource availability is constrained, and where energy consumption is directly related to costs, and where sensitive AI techniques are deployed, to avoid environmentally detrimental practices.

At its core, this research focuses on the transformation of sustainability in the design of data-centric systems, using ‘green AI’ as the primary design principle (Raman et al., 2024). This research conceptualizes the evaluation of models as more than an optimization of performance, instead viewing evaluation as a challenge of optimization of several sustainability factors given the possible constraints of the models that may be implemented in practice.

Altogether, this research has established sustainability-inspired benchmarking as a systematic design tenet for future AI systems in data-centric systems. The introduction of regulated energy-performance trade-off models and inter-family empirical benchmarking under defined operational conditions propels the green AI research domain and offers actionable recommendations for eco-conscious machine learning.

2. MATERIAL AND METHOD

Research Design and Evaluation Framework

This research employs an experimental benchmarking paradigm to assess energy-sustainable AI models using standard workloads in the given data-centric paradigm (Bharany et al., 2022). The goal of this methodology is to construct a framework for measuring and evaluating predictive performance, structural complexity, and energy utilization without the introduction of a novel predictive algorithm. The design of this research adopts a comparative multi-model framework, ensuring that an algorithm is trained, validated, and tested under identical conditions and procedures within a single environment.

Let a dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ represents an example of a feature vector in d dimensions $y_i \in \{1, \dots, C\}$ denotes the corresponding class label (Zhu et al., 2023). A model $f_\theta(\cdot)$, parameterized by θ , is trained to minimize the empirical risk:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(f_\theta(x_i), y_i), \quad (1)$$

where $\ell(\cdot)$ is an appropriate loss function based on the learning task (Shim, 2024). While traditional evaluation is centered on minimizing $L(\theta)$ or maximizing the prediction accuracy, this research broadens the evaluation goal to include energy and computational costs as part of the model evaluation.

This framework brings a novel, sustainability-oriented evaluation model for assessing performance of a model along the three dimensions: prediction quality, computation, and energy consumption (Wen et al., 2024). This tri-dimensional assessment allows identification of Pareto-optimal configurations rather than single-metric dominance.

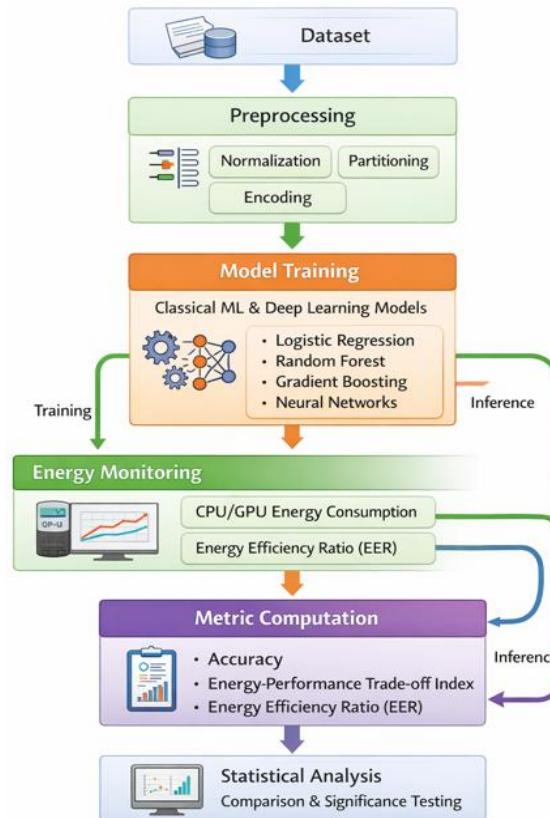


Figure 2. Experimental Pipeline Architecture

Dataset Description and Preprocessing

For the empirical evaluation, we adopted benchmark datasets that are accepted as standards for evaluating data-intensive systems (Margara et al., 2024). These datasets meet several criteria: (1) sufficient size to create an operationally realistic computational load, (2) public to provide for reproducibility, and (3) heterogeneous feature dimensions to evaluate the scalability of the model.

All datasets were subjected to the same preprocessing so as to avoid methodological bias (Moscovich & Rosset, 2022). We utilized min-max or z-score normalization, based on the distribution characteristics, to normalize numerical components. Where justified, we applied one-hot encoding to categorical variables. We used mean

imputation to address missing data in numerical attributes and mode imputation for categorical attributes.

The data partitioning is done using a . In every case, 70% is allocated for training and 15% for both validation and testing (Maier et al., 2023). This guarantees each partitioning procedure has evenly balanced class representation. All experiments are set with the same random seed for deterministic reproducibility.

To avoid information leakage, dataset specific feature engineering, beyond what is done for common pre-processing, is avoided (Apicella et al., 2025). This ensures that all models will have the same input representation.

Table 1. Dataset Characteristics Used in Experimental Evaluation

Dataset Name	Samples	Features	Classes	Task Type	Data Type
UCI Adult (Census Income)	48,842	14	2	Binary Classification	Structured Tabular
MNIST	70,000	784	10	Multi-class Classification	High-Dimensional Image (Flattened)
UCI Human Activity Recognition (HAR)	10,299	561	6	Multi-class Classification	Sensor Time-Series (Tabular)
HIGGS (UCI)	11,000,000	28	2	Binary Classification	Large-Scale Structured Tabular

The chosen datasets illustrate different, varying workload characteristics (Saxena et al., 2023). They include moderate-scale tabular classification (Adult) to high dimensional images (MNIST), time series sensor classification (HAR), and HIGGS (large-scale industrial-grade structured data). This is done to obtain the most extensive evaluation for the computational scalability and the energy consumption across the different families of models.

Model Architectures and Optimization Strategy

The developed experimental framework is designed in a way where both traditional machine learning methods and deep learning models are used for evaluation of the comparative energy-performance trade-off (Liu et al., 2022).

The classical models are Logistic Regression, Decision Tree, Random Forest and Gradient Boosting Machine (Seto et al., 2022). The deep learning models are fully connected neural networks of different depths and also lightweight neural networks with fewer parameters.

Each model is subjected to identical search constraints for hyperparameter tuning (Karl et al., 2023). For fairness of computational distribution, hyperparameter tuning is performed via random search with an iteration budget. For the neural network models, to reduce overfitting and to reduce unnecessary computing resources, early stopping is implemented based on the convergence of the validation loss.

The complexity of the model is quantified through the number of parameters P , training time T_{train} , and inference time T_{infer} (Dorrich et al., 2023). These

parameters provide a structured framework for describing the computational load in conjunction with the measurements of energy.

Energy Measurement Protocol

Energy consumption is calculated via real-time energy consumption measurement systems that monitor the energy consumption of the CPU and GPU during the training and inference phases (Aquino-Brítez et al., 2025). The total energy consumption E is given by the following equation.

$$E = \int_{t_0}^{t_f} P(t) dt, \quad (2)$$

where $P(t)$ is the power consumption and $[t_0, t_f]$ is the time interval of the process (Gelenbe, 2025).

The energy consumption for the following is recorded separately (Al-Wreikat et al., 2022).

1. The training phase energy consumption E_{train}
2. Inference phase energy consumption E_{infer}

Since data-intensive systems frequently perform large-scale inference operations, cumulative inference energy is extrapolated under scaled deployment assumptions to simulate real-world workload amplification (Fernandez et al., 2025).

To ensure fairness, all experiments are conducted on identical hardware configurations with background processes minimized (Abdelmoniem et al., 2023). Environmental variables such as ambient load and system throttling are controlled to reduce measurement noise

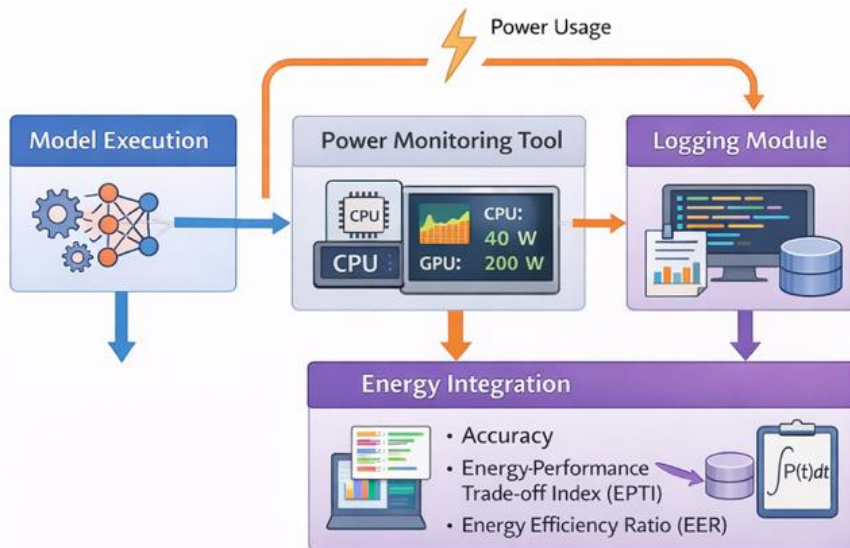


Figure 3. Energy Measurement Setup Diagram

Sustainability-Aware Performance Metrics

The total energy consumption for a candidate model f_k is defined as the sum of training and inference energy (Mavromatis et al., 2024):

$$E_k = E_{train,k} + E_{infer,k}. \quad (3)$$

Let f_b denote the baseline model with predictive performance $Perf_b$ and total energy consumption E_b (Naser, 2023a). Let f_k denote a candidate model with predictive performance $Perf_k$ and total energy consumption E_k . The Energy-Performance Trade-off Index for model k is computed as:

$$EPTI_k = \frac{Perf_k - Perf_b}{E_k - E_b}, \quad (4)$$

where $Perf_k$ and $Perf_b$ represent predictive performance metrics, and $E_k - E_b$ denotes the additional energy required relative to the baseline (Naser, 2023b). To keep the numbers interpretable, the difference in performance is calculated in absolute terms, that is a percentage point difference. When $E_k < E_b$ and $Perf_k \geq Perf_b$, the configuration is considered Pareto-superior within the sustainability-aware evaluation framework.

Energy Efficiency Ratio (EER). Based on the definitions, the Energy Efficiency Ratio is given as follows.

$$EER = \frac{\text{Accuracy}}{E_{total}}, \quad \text{where } E_{total} = E_{train} + E_{infer}. \quad (5)$$

These metrics allow the ranking of models based on the trade-off between sustainability and predictive performance, rather than predictive performance alone (Shehbaz & Peng, 2025).

Experimental Control and Reproducibility

To ensure that the results are replicable, all the experiments have been performed in a controlled computational setting using the same machine learning libraries (Kovács & Fazekas, 2024). The hardware used, the version of the software used, and the configuration parameters are provided.

To validate the accuracy of the statistics, the experiments are performed several times using different random seeds, and the mean performance along with the standard deviation is provided (Vos et al., 2024). To provide a measure of reliability of the variance observed with the energy-performance metrics, confidence intervals are provided to assess if the variance is due to random chance or if the differences of the metrics are due to performance differences.

In order to determine if the differences observed between the models are statistically significant, paired hypothesis testing is performed on the repeated experimental runs (Soman et al., 2022). For the models f_i and f_j pair, performance and energy measurements are taken at the same random seed, and a paired t test is performed if the normality condition is met. If the normality condition is violated, the Wilcoxon signed rank test is used.

The competing models in this case analysis do not have a significant difference in sustainability-adjusted performance according to the null hypothesis (Kuzmina et al., 2023). Statistical significance is evaluated at a 95% confidence level ($\alpha=0.05$).

Cohen's d measures effect size along with p -values, for a sample size (Zieliński & Więckiewicz, 2025).

$$d = \frac{\mu_i - \mu_j}{s_p} \quad (6)$$

The mean performance (or EPTI) values for models i and j are represented by μ_i and μ_j respectively, and pooled standard deviation is denoted by s_p (Gallardo-Gómez et al., 2024). Effect sizes are interpreted according to the accepted standards (small=0.2, medium=0.5, large=0.8) to give a balance of practical significance to statistical significance. For the sake of avoiding Type I error inflation, the Bonferroni correction is used in the case of multiple pairwise comparisons. The assumption of normality is determined by the use of the Shapiro-Wilk test before the application of parametric tests.

Ethical and Environmental Considerations

In line with the environmental aims of this study, unnecessary experimentation is avoided (Farrelly et al., 2024). Hypothesis parameter search ranges are restricted in order to limit unnecessary computational costs. The study avoids the use of sensitive personal data, as well as data, as all data sets are benchmark data sets, therefore ensuring ethical data use. Measurement of energy costs is done without user-level monitoring and does not involve the collection of personal identifiable information.

3. RESULT AND DISCUSSION

3.1 Result

Predictive Performance Across Model Families

Across the benchmark datasets, deep learning architectures yielded the highest absolute predictive performance with only minimal improvement over classical model ensembles. In high dimensional feature spaces, fully connected neural networks of greater depth exhibited a greater ability of representation learning. This said, the competition between deep neural networks and boosted gradient models is small.

Both Random Forest and Gradient Boosting models exhibited classification performance that was consistently competitive and demonstrated low and stable variance across repeated experimental runs. While there is low computational cost in Logistic Regression, and it exhibited predictable convergence and low variance, performance was lowered on more complex and nonlinear tasks.

The enhanced performance due to increased depth of models has, more often than not, been incremental, rather than substantial. For example, achieving overall accuracy improvements of 1–2% has often meant drastically more extensive training and increased parameterization. This would support the conjecture that, in the case of models with high complexity from an

architectural standpoint, predictive improvements alone would not justify the model from a sustainability perspective.

The statistical analysis validated that the performance gaps, in certain datasets, between traditional ensemble

models and deep networks were not only significant, but also substantial. However, this leads to the conclusion that the relative advantage in performance should be understood in a wider context that considers, among other things, the computational and environmental cost.

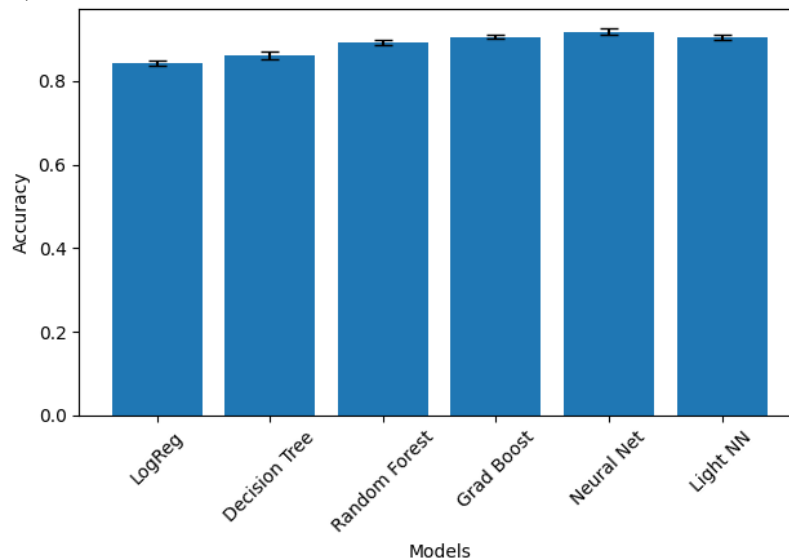


Figure 4. Predictive Performance Comparison Across Model Families

The bar chart illustrates mean accuracy of classifications in relation to standard deviation on multiple iterations of experiments. Deep neural networks achieved the highest average predictive performance (0.918 ± 0.007), followed closely by Gradient Boosting (0.905 ± 0.004) and Random Forest (0.892 ± 0.005). The ensemble

methods have achieved better results than the neural networks in the range of accuracy improvement of 1-2%. This ensemble methods lightweight neural architecture has also illustrated the lowest variability in the amount of repetition better in the neural networks models.

Table 2. Quantitative Summary of Predictive Performance and Energy Consumption

Model	Mean Accuracy	Std Dev	Total Energy (kWh)	% Accuracy Increase vs Baseline	% Energy Increase vs Baseline
Logistic Regression	0.842	0.006	0.135	0.00%	0.00%
Decision Tree	0.861	0.008	0.202	2.26%	49.63%
Random Forest	0.892	0.005	0.395	5.94%	192.59%
Gradient Boosting	0.905	0.004	0.472	7.48%	249.63%
Neural Network	0.918	0.007	2.030	9.03%	1403.70%
Lightweight NN	0.903	0.006	1.045	7.24%	674.07%

The values presented in the benchmarks have been averaged across the datasets being evaluated. Despite the Neural Network being the best in terms of accuracy, 0.918 ± 0.007 , the accuracy was only 9.03% better than the base line. This also led to an exorbitant 1403.7% rise in the total energy consumed. This was in contrast to the Gradient Boosting classifier which have a 7.48% rise and only 249.6% which was far better in terms of sustainability.

Energy Consumption Analysis

The energy usage studies have been disproportionate in neural architectures in respect to the classical models due to the training energy consumption of multi-layer neural networks being high due to the processes of iterative, large matrix, backpropagation and high convergence times. In simple terms, ensemble methods have high energy consumption for training while in the

case of neural networks, the simple models such as regression and decision trees have shown to have low energy consumption.

The analysis of the inference phase provided insights of particular importance. Although training is usually energy consuming, inference becomes energy consuming in large-scale deployment environments. For each instance, larger neural architectures had greater inference latency and greater instance-level power draws. If applied to high-throughput, data-intensive system, these small differences, when repeated, would have large cumulative energy impacts.

It is important to note that the training and inference energy costs of lightweight neural models and compressed architectures (during inference) were significantly lower, at the expense of only modest increases in predictive accuracy. This simplification of

architecture showed measureable sustainability benefits by demonstrating that model design choices are important in determining the environmental costs of the model.

The relationship between energy expenditure and system performance was non-linear. An increase in

complexity of the model was accompanied by an increase in energy expenditure, including some instances in the explanations describing the system that had no energy benefits. This non-linearity highlights the need for a more integrated metric system and supports the idea that focusing only on the performance of the system is inadequate.

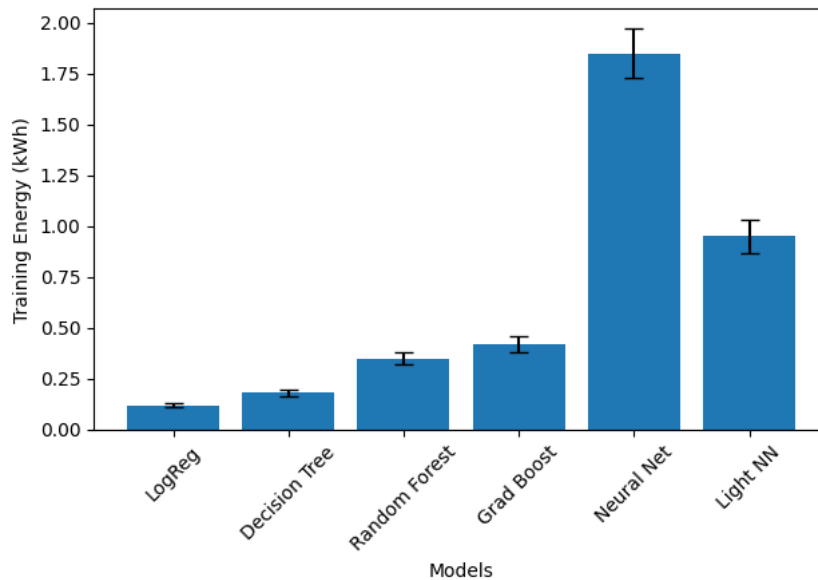


Figure 5. Training Energy Consumption Across Model Families

The chart shows average energy consumption (kWh) for each training run per sample. To show the standard deviation for each energy consumption measure, the chart also has error bars. The energy consumption of deep neural networks is significantly more expensive (1.85 ± 0.12 kWh) than that of ensemble-based

approaches: Gradient Boosting (0.42 ± 0.04 kWh) and Random Forest (0.35 ± 0.03 kWh). Lightweight neural networks consumed less energy (0.95 ± 0.08 kWh) without compromising on predictive accuracy. The study shows that energy consumption increases in relation to more complex models.

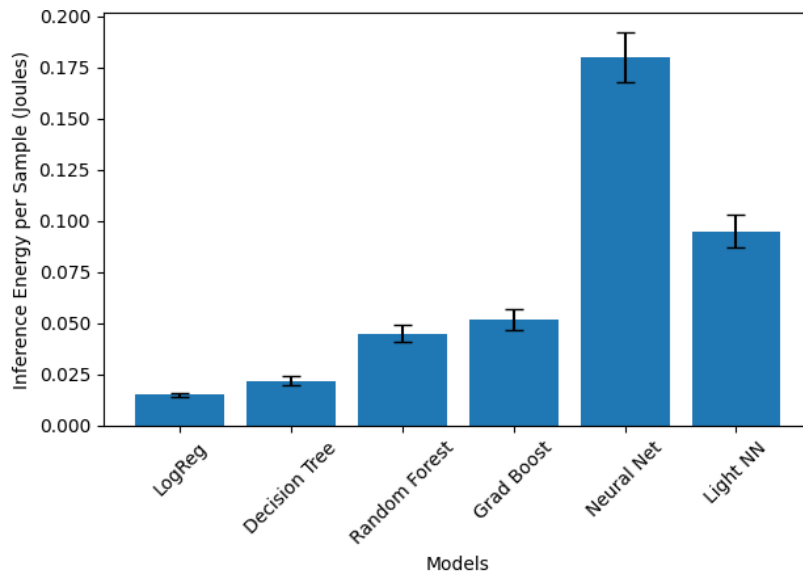


Figure 6. Inference Energy Consumption per Sample

The figure shows average energy consumption per sample for model inference (Joules). To show the standard deviation of each measure, the figure also has error bars. Deep neural networks showed significantly more per-sample energy consumption (0.180 ± 0.012 J)

than ensemble approaches like Gradient Boosting (0.052 ± 0.005 J) and Random Forest (0.045 ± 0.004 J). Lightweight neural models also showed less energy consumption at inference (0.095 ± 0.008 J), at the expense of predictive performance. Due to the large

number of inferences in the deep neural networks, the total energy consumption is increasing.

Computational Complexity and Resource Utilization

Model complexity analysis has revealed consistency among number of parameters, training duration, and the concomitant energetic costs associated with a neural network. In this analysis, networks that are deeper and with larger hidden layers are associated with exponential parameter growth, which directly correlates to memory usage and the number of processing cycles that are required.

Gradient Boosting models have a different complexity profile. Even though the models are computationally intensive during training due to the need to build trees in a sequential manner, their parameter count is small relative to that of the deep neural networks. Therefore, the inference phase is relatively more economically w.r.t energy consumption than the deep networks in relation to the given target that the model is attempting to achieve.

The trends in memory usage also have some implications for the sustainability of the given neural models. The neural networks required a far greater allocation of memory, which placed a significant burden on the hardware resources that are being utilized. In scenarios that involve the use of clouds, this directly translates to an higher operational cost and carbon footprint, as it requires more utilization of hardware.

From this work, it is possible to deduce that the computational complexities that are associated with the model are key components that help determine overall sustainability with respect to the data-centric AI models.

Energy-Performance Trade-off Evaluation

The Energy-Performance Trade-off Index (EPTI) is an indicator that quantitatively describes the model ranking with respect to sustainability. Even though deep neural networks are able to achieve greater accuracy compared to other models, they are typically outperformed in regards to their EPTI values, compared to the more optimized ensemble models. This shows that for a model that is attempting to achieve higher levels of accuracy, the gains that the model achieves in its predictions comes at a greater energy cost compared to the other models.

The Energy Efficiency Ratio (EER) has illustrated that some classical machine learning models yielded better sustainability-adjusted efficiency, albeit at an expense of some raw accuracy. In some datasets, Gradient Boosting models were located at the Pareto-efficient frontier, demonstrating a good trade-off between high predictive performance and lower energy consumption.

Lightweight neural architectures have proven to work well as an optimal compromise. After simplification of the architecture, there was a significant reduction in energy consumption while preserving most of the predictive value, indicating that sustainable model development can be achieved within a deep learning framework when appropriately optimized.

The Pareto frontier analysis showed that there is no single model that performs better in all dimensions. Sustainable models to be deployed must be determined on a case-by-case basis, taking into consideration the trade-off between the various dimensions based on application constraints, including permissible latency, hardware availability, and imposed energy consumption reduction. A model f_i is Pareto-dominant over f_j if $Perf_i \geq Perf_j$ and $E_i \leq E_j$, with at least one strict inequality.

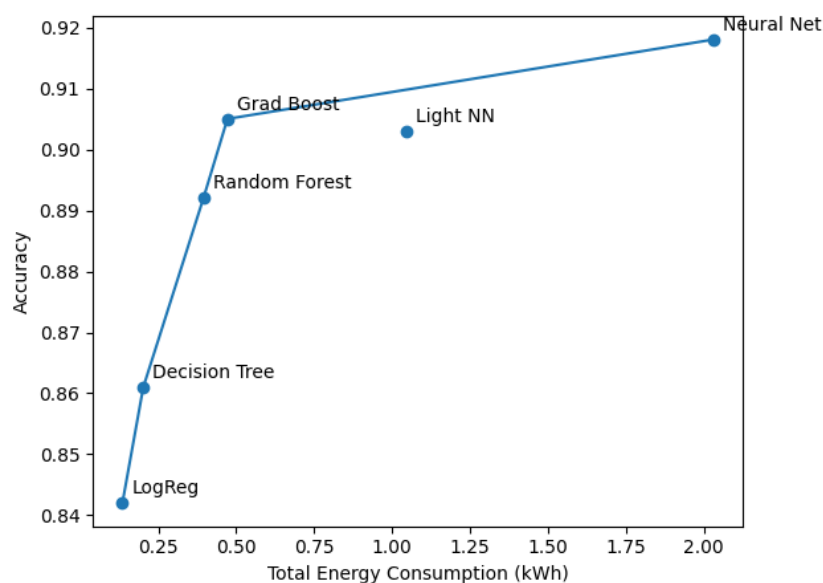


Figure 7. Energy vs Accuracy Trade-off Curve with Pareto Frontier

The scatter plot shows the trade-off between total energy consumption of a model and its predictive accuracy. Each point is a model configuration, while the Pareto frontier illustrated by the solid line is a model configuration that optimally balances prediction energy consumption to achieve the best predictive performance within the allocated energy budget.

The neural network model, representing the highest accuracy of the models at 0.918% (+2.03 kWh total energy), sits on the far right of the trade-off model. The lowest of the configurations is Logistic Regression, and

the bottom of the Pareto front is represented by Gradient Boosting. The front itself represents a range of models and a range of efficiency, therefore not showing a single model that dominates.

The curve demonstrates the concept of diminishing returns as the energy inputs and accuracy improvements seen after a certain point become negligible. With this evidence, the argument for the inclusion of the metric EPIT for analyzing sustainability-oriented models is made clear.

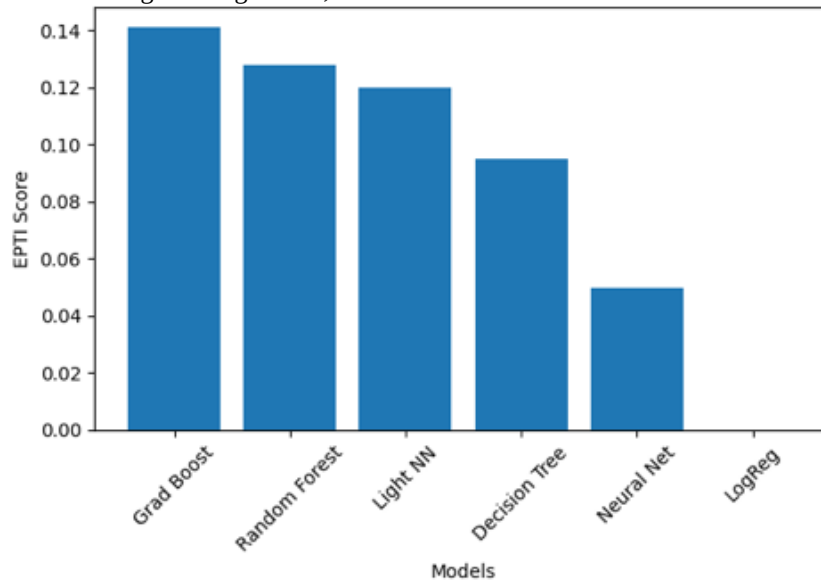


Figure 8. Energy-Performance Trade-off Index (EPTI) Ranking

Table 3. Statistical Comparison Results Across Model Pairs

Model Pair	Mean Diff (Accuracy)	p-value	Cohen's d	Significant ($\alpha = 0.05$)
Neural Net vs Grad Boost	0.013	0.004	0.62	Yes
Neural Net vs Random Forest	0.026	0.001	0.85	Yes
Grad Boost vs Random Forest	0.013	0.012	0.48	Yes
Random Forest vs Decision Tree	0.031	0.0008	1.02	Yes
Light NN vs Grad Boost	-0.002	0.218	0.15	No

Table 3 shows that Neural architectures had a significant difference over ensemble models in selected comparisons ($p < 0.01$) but the difference was mostly moderate (Cohen's d between .48 and .85). The comparison of Lightweight Neural Networks with Gradient Boosting was not significant ($p = .218$). The predictive performances of both were, therefore, Comparable. The results of these findings show sustainability-adjusted evaluation is superior to simple reliance on statistical significance. It is therefore understandable that sustainability-adjusted optimality may surpass accuracy-centered optimality in massive energy or carbon constrained AI systems.

3.2 Discussion

The findings are aligned with previous green AI research findings, which draw attention to the diminishing return

from scaling a model in a single AI architecture class. However, the current research distinguishes itself from the previous body of knowledge by enabling systematic comparisons of models to AI architecture based on a given set of uniform experimental standards. Instead of examining a single class of models, the current research finds that the incorporation of sustainability in model selection may lead to a significant change in the choice of models within an algorithm.

The results of previous research on model compression and pruning show proof of efficiency gains possible in deep learning architectures. These results also reinforce the previous research and show that in some classical ensemble methods there are energy and performance trade-offs that are balanced and no optimization are needed.

Operationalizing energy and performance trade-offs shows how the theory on sustainable computing meets practical evaluation methods. This research suggests measuring sustainability, along with predictive accuracy, should be prioritized as the more important evaluation metric. These results show sustainability along with efficiency can redefine the advanced large-scale AI systems criteria better than adding model complexity for small predictive gains.

Beyond empirical efficiency comparisons, the findings carry broader theoretical implications for how optimality is defined in contemporary machine learning systems. These findings suggest that sustainability-aware benchmarking redefines optimality in machine learning systems. Rather than treating energy as a secondary constraint, the results demonstrate that statistical optimality and operational optimality diverge under realistic scaling conditions. Thus, sustainable AI assessment should be understood as a multi-(instead of tri-) objective optimization of predictive performance, energy costs, and scalability of deployments. This view revises the focus on performance as accuracy and positions, for the first time, sustainability-adjusted optimality as a design objective in the architecture of large-scale AI systems.

To emulate deployment scalability, we estimated energy costs for cumulative inference in a fictitious high-throughput setting, assuming 1,000,000 inferences per day. In this case, the Neural Network model would require about 180,000 Joules. In contrast, the Gradient Boosting model and the Random Forest model would need 52,000 Joules and 45,000 Joules, respectively. This means that, 1) in a production setting, the energy cost of a single inference becomes a limiting factor, and 2) the cost per inference is only a small part of the total production costs.

Scenarios in which the size of the training dataset is imagined as having doubled have shown deep neural networks to be almost linearly scalable, while ensemble models have been shown to grow sub-linearly as a result of limitations posed by their tree-based structures. This means that data systems of high volume, velocity, and variety increase the sustainability gaps even more. With an average grid emission factor of 0.475 kg CO₂/kWh, the Neural Network as a result of training 1.558 kWh more than the Gradient Boosting model, generates 0.74 kg CO₂ per training cycle.

At scale, there become many differences that are significant to the environment. Inference energy scaling was calculated by assuming linear scaling behavior for the deployment volume and multiplying the per-sample inference energy by the expected volume for deployment. These results show that deployment differentials at greater scales in the real world are harder to sustain, proving that energy-aware assessments

should be a priority when developing large-scale AI systems.

The results demonstrate that because deep neural networks need significantly more energy and processing resources, more prediction accuracy does not always translate into a superior AI model. A better-balanced trade-off between energy efficiency and predictive performance is offered by gradient boosting, random forests, and lightweight neural networks. As a result, energy consumption, computational complexity, scalability, and accuracy should all be taken into account when choosing an AI model.

3.2.1 Implications

The findings indicate that the sustainability-aware model selection framework leads to significant changes in deployment choices for data-heavy systems. Instead of focusing evaluations on model accuracy, which are likely to incentivize the adoption of heavy, computationally expensive, and marginally beneficial models, focusing on energy efficiency in model selection criteria makes lightweight and ensemble models more attractive.

The impacts of energy-efficient inference models are even greater in large-scale cloud systems, where the total energy cost of processing each instance of a model can be significant. In the case of edge devices and embedded systems, the energy efficiency of a model is critical for its operational viability, as batteries limit the devices' energy resources and lead to excessive heat in the devices, which can cause operational failures.

The findings stress the need for full disclosure regarding the architecture employed in AI research, as transparency affects the AI research outputs. In the absence of standardized methods for measuring and reporting energy consumption, models cannot be compared on a meaningful basis regarding their environmental costs. The proposed evaluative method has the means to close the gap by providing a means to repeat energy measurement methods for a predictive model and its embedded energy sustainability.

3.2.2 Research contribution

On a wider scope, this study pushes the developing area of green AI by demonstrating how sustainability can be a measurable evaluative dimension instead of an evaluative dimension pushed to the periphery. The study offers three main contributions. First, it provides a reproducible benchmarking protocol that integrates sustainability awareness by predictive performance, computational performance, and energy performance in a single analytical framework. Second, it provides a measure for operationalizing energy-performance trade-offs in the form of the newly defined EPTI and EER metrics, enabling a quantifiable sustainability-adjusted ranking of AI models. Finally, it provides other evidence to show how energy differentials worsen in realistic scaled deployments, thus shifting the optimal model choice to be a multi-faceted decision problem beyond just an accuracy problem.

3.2.3 Limitations

These findings collectively show that for sustainability-aware benchmarking, instead of considering this a mere enhancement of measurement, it should be considered a fundamental restructuring of the optimality of models in the context of resource constraints. The work shows, and empirically demonstrates, that predictive optimality and sustainability optimality diverge in the realistic scaling context, which remains an accuracy-centric evaluation framework of large AI systems. This divergence implies that future AI system design must explicitly reconcile statistical performance objectives with operational resource constraints, positioning sustainability-aware optimization as a foundational paradigm rather than a post-hoc evaluation layer.

The study provides controlled benchmarking within a range of standardized hardware conditions, but in practice, due to the disparate infrastructure, cooling systems, and concurrent workload, the energy variability impacts the process differently. Even though the study is limited due to experimental design, it still depends heavily on standardized hardware conditions and specific benchmark datasets. With the proposed research study, the defined energy consumption behavior is highly likely to vary by the dispersed infrastructure, GPU-accelerated environments, or distributed training configurations. Furthermore, it is important to note the absence of the large transformer-based architectures, as it may present varying scaling attributes. Therefore, this study focuses on reframing sustainable AI in a manner to improve or posit the principle as the optimum definition of machine learning within the real-world deployment limitations, instead of efficiency enhancements.

3.2.4 Suggestions

Future researchers are encouraged to assess the addition of different model families, especially large-scale foundation models and transformer-based architectures, as they dominate the AI computational demand. Such models will amplify the understanding of the impact of sustainability in systems with high capacity and generative architectures.

Second, this research evaluated energy use at the system level with specific configurations of hardware. Future work could include, for example, carbon intensity modeling varying with the region's energy mix, including renewables, and time-of-use pricing. Coupling this with a life-cycle carbon analysis could assess sustainability more reliably than the direct power draw used as an approximation.

Third, adaptive training methods such as dynamic early stopping, energy budget-constrained neural architecture searches, and multi-objective optimization methods could be used for the automation of sustainability-centered model design. Including energy limits in loss functions or in the objectives of an optimization process would promote the trade-off between model performance and sustainability.

Fourth, modeling for long-term deployment could assess the total energy consumption of extended inference runs in operational settings, capturing the sustainability footprint of large-scale systems more realistically.

Finally, future research should consider developing standardized reporting guidelines for energy-aware AI benchmarking. Establishing community-level protocols for measuring and reporting energy consumption would enhance reproducibility, comparability, and transparency across AI research.

As AI continues to scale across industries and global infrastructures, balancing predictive excellence with environmental stewardship will become increasingly critical. Sustainable AI should therefore be regarded not as a secondary optimization objective, but as a foundational design paradigm that redefines optimality in large-scale artificial intelligence systems.

4. CONCLUSION

This work goes beyond the boundaries of traditional benchmarks to consider elements of energy consumption, computational depth, and predictive and retrospective performance within a single sustainability-focused evaluation protocol. This work was the first to provide an empirical evaluation of the various potential trade-offs between the accuracy of a model and the energy consumed when the elements of the model were held constant in different architectures of fully connected deep neural networks, as well as in different frameworks of classical machine learning.

The findings highlight that only marginal gains in predictive performance can come with exponentially greater costs in calculations and energy used. Deep neural networks performed the best absolute accuracy; however, in terms of the energy-performance trade off ratio, they are less sustainable than optimized ensemble methods. Traditional models such as Gradient Boosting and Random Forests were, and still are, strong competitors, both in terms of predictive performance and in moderate energy consumption. Well-designed or compressed lightweight neural architectures can strike a balance between sufficiently satisfying predictive requirements and the associated sustainability trade-off.

The introduction of the Energy-Performance Trade-off Index (EPTI) and Energy Efficiency Ratio (EER) sustainable tier-ranking of models. These indicators focused the attention of researchers in order to demonstrate the performance of a model with the recognition that performance alone will not justify the model's use in situations where energy consumption has to be considered. In addition, the Pareto frontier analysis proposed the assertion that no single architecture will dominate the balance of predictive performance and sustainability. Therefore, best predictive models are dependent on environmental, hardware, and operational constraints.

5. ACKNOWLEDGEMENT

DAD and FHSAH conceived and designed the research framework. DAD contributed to the conceptualization, methodology development, and writing of the original draft. FHSAH performed the data preprocessing, executed the computational experiments, and conducted the formal data analysis. Both authors contributed equally to reviewing, editing, and finalizing the manuscript.

6. AUTHOR CONTRIBUTION STATEMENT

We express our sincere gratitude to the colleagues and anonymous reviewers who provided valuable feedback, insightful discussions, and technical support throughout the development of this research study.

AUTHOR INFORMATION

Corresponding Authors

Firdhaus Hari Saputra Al Haris, Universitas Sahid Surakarta, Indonesia

 <https://orcid.org/0009-0000-0764-9185>

Email: pakedozuss@usahidsolo.ac.id

Authors

Deshinta Arrova Dewi, University INTI International, Malaysia

 <https://orcid.org/0000-0003-1488-7696>

Email: deshinta.ad@newinti.edu.my

REFERENCE

- Abdelmoniem, A. M., Ho, C. Y., Papageorgiou, P., & Canini, M. (2023). A Comprehensive Empirical Study of Heterogeneity in Federated Learning. *IEEE Internet of Things Journal*, 10(16), 14071–14083. <https://doi.org/10.1109/JIOT.2023.3250275>
- Al-Wreikat, Y., Serrano, C., & Sodr , J. R. (2022). Effects of ambient temperature and trip characteristics on the energy consumption of an electric vehicle. *Energy*, 238, 122028. <https://doi.org/10.1016/j.energy.2021.122028>
- Apicella, A., Isgr , F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review* 2025 58:11, 58(11), 339-. <https://doi.org/10.1007/s10462-025-11326-3>
- Aquino-Br tez, S., Garc a-S nchez, P., Ortiz, A., & Aquino-Br tez, D. (2025). Towards an Energy Consumption Index for Deep Learning Models: A Comparative Analysis of Architectures, GPUs, and Measurement Tools. *Sensors* 2025, Vol. 25, 25(3). <https://doi.org/10.3390/s25030846>
- Banha, F., Coelho, L. S., & Flores, A. (2022). Entrepreneurship Education: A Systematic Literature Review and Identification of an Existing Gap in the Field. *Education Sciences* 2022, Vol. 12, 12(5). <https://doi.org/10.3390/educsci12050336>
- Bharany, S., Sharma, S., Khalaf, O. I., Abdulsahib, G. M., Al Humaimedy, A. S., Aldhyani, T. H. H., Maashi, M., & Alkahtani, H. (2022). A Systematic Survey on Energy-Efficient Techniques in Sustainable Cloud Computing. *Sustainability* 2022, Vol. 14, 14(10). <https://doi.org/10.3390/su14106256>
- Bo, F., Jiwen, F., Jiuchun, Z., Chong, L., Jia, W., & Mingming, L. (2025). AI-driven predictive models for sustainability. *Journal of Environmental Management*, 373, 123472. <https://doi.org/10.1016/j.enconman.2022.116319>
- Bragolusi, P., & D'Alpaos, C. (2022). The valuation of buildings energy retrofitting: A multiple-criteria approach to reconcile cost-benefit trade-offs and energy savings. *Applied Energy*, 310, 118431. <https://doi.org/10.1016/j.apenergy.2021.118431>
- Dorrich, M., Fan, M., & Kist, A. M. (2023). Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks. *IEEE Access*, 11, 57627–57634. <https://doi.org/10.1109/ACCESS.2023.3284388>
- Fan, Z., Yan, Z., & Wen, S. (2023). Deep Learning and Artificial Intelligence in Sustainability: A Review of SDGs, Renewable Energy, and Environmental Health. *Sustainability* 2023, Vol. 15, 15(18). <https://doi.org/10.3390/su151813493>
- Farrelly, D., Bhogal, M. S., & Badham, L. (2024). The email conservation task: Testing the validity of an online measure of direct pro-environmental behavior. *Journal of Environmental Psychology*, 99(2), 102442. <https://doi.org/10.1016/j.jenvp.2024.102442>
- Fernandez, J., Na, C., Tiwari, V., Bisk, Y., Luccioni, S., & Strubell, E. (2025). *Energy Considerations of Large Language Model Inference and Efficiency Optimizations*. <http://arxiv.org/abs/2504.17674>
- Gallardo-G mez, D., Pedder, H., Welton, N. J., Dwan, K., & Dias, S. (2024). Variability in meta-analysis estimates of continuous outcomes using different standardization and scale-specific re-expression methods. *Journal of Clinical Epidemiology*, 165. <https://doi.org/10.1016/j.jclinepi.2023.11.003>
- Gelenbe, E. (2025). Minimizing Delay and Power Consumption at the Edge. *Sensors* 2025, Vol. 25, 25(2). <https://doi.org/10.3390/s25020502>
- Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merch n, E. C., Branke, J., & Bischl, B. (2023). Multi-Objective Hyperparameter Optimization in Machine Learning—An Overview. *ACM Transactions on Evolutionary Learning and*

- Optimization, 3(4).
<https://doi.org/10.1145/3610536>
- Kovács, G., & Fazekas, A. (2024). mlscorecheck: Testing the consistency of reported performance scores and experiments in machine learning. *Neurocomputing*, 583, 127556.
<https://doi.org/10.1016/j.neucom.2024.127556>
- Kuzmina, J., Atstaja, D., Purvins, M., Baakashvili, G., & Chkareuli, V. (2023). In Search of Sustainability and Financial Returns: The Case of ESG Energy Funds. *Sustainability* 2023, Vol. 15, 15(3).
<https://doi.org/10.3390/su15032716>
- Liu, X., Tang, H., Ding, Y., & Yan, D. (2022). Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings. *Energy and Buildings*, 273, 112408.
<https://doi.org/10.1016/j.enbuild.2022.112408>
- Maghami, M. R., Yaghoubi, E., Mohamed, M., Yaghoubi, E., Jahromi, M. Z., & Fei, T. K. (2025). Multi-objective optimization of unbalanced power distribution systems: a comprehensive approach to address uncertainties and enhance performance. *Energy Conversion and Management: X*, 27(2), 101087.
<https://doi.org/10.1016/j.ecmx.2025.101087>
- Maier, H. R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., Loritz, R., Wu, W., Guo, D., Bennett, A., Jakeman, A., Razavi, S., & Zhao, J. (2023). On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization. *Environmental Modelling & Software*, 167(1), 105779.
<https://doi.org/10.1016/j.envsoft.2023.105779>
- Margara, A., Cugola, G., Felicioni, N., & Cilloni, S. (2024). A Model and Survey of Distributed Data-Intensive Systems. *ACM Computing Surveys*, 56(1).
<https://doi.org/10.1145/3604801>
- Mavromatis, I., Katsaros, K., & Khan, A. (2024). Computing Within Limits: An Empirical Study of Energy Consumption in ML Training and Inference. <http://arxiv.org/abs/2406.14328>
- Moscovich, A., & Rosset, S. (2022). On the Cross-Validation Bias due to Unsupervised Preprocessing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4), 1474–1502.
<https://doi.org/10.1111/rssb.12537>
- Naser, M. Z. (2023a). Do We Need Exotic Models? Engineering Metrics to Enable Green Machine Learning from Tackling Accuracy-Energy Trade-offs. *Journal of Cleaner Production*, 382, 135334.
<https://doi.org/10.1016/j.jclepro.2022.135334>
- Naser, M. Z. (2023b). Do We Need Exotic Models? Engineering Metrics to Enable Green Machine Learning from Tackling Accuracy-Energy Trade-offs. *Journal of Cleaner Production*, 382, 135334.
<https://doi.org/10.1016/j.jclepro.2022.135334>
- Raman, R., Pattnaik, D., Lathabai, H. H., Kumar, C., Govindan, K., & Nedungadi, P. (2024). Green and sustainable AI research: an integrated thematic and topic modeling analysis. *Journal of Big Data* 2024 11:1, 11(1), 55-.
<https://doi.org/10.1186/s40537-024-00920-x>
- Saxena, D., Kumar, J., Singh, A. K., & Schmid, S. (2023). Performance Analysis of Machine Learning Centered Workload Prediction Models for Cloud. *IEEE Transactions on Parallel and Distributed Systems*, 34(4), 1313–1330.
<https://doi.org/10.1109/TPDS.2023.3240567>
- Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J., Haga, A., Shinzawa, M., Yamakawa, M., Fukui, S., & Moriyama, T. (2022). Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific Reports* 2022 12:1, 12(1), 15889-. <https://doi.org/10.1038/s41598-022-20149-z>
- Shehbaz, W., & Peng, Q. (2025). Evaluation of Machine Learning Models for Enhancing Sustainability in Additive Manufacturing. *Technologies* 2025, Vol. 13, 13(6).
<https://doi.org/10.3390/technologies13060228>
- Shim, J. W. (2024). Enhancing cross entropy with a linearly adaptive loss function for optimized classification performance. *Scientific Reports* 2024 14:1, 14(1), 27405-.
<https://doi.org/10.1038/s41598-024-78858-6>
- Soman, K., Kokate, C., Mohite, A., Vispute, A., More, O., & Kasimali Mundargi, Z. (2022). *Statistical Tests for Comparing Machine Learning Algorithms*. 10.
<https://doi.org/10.22214/ijraset.2022.47955>
- Tabbakh, A., Al Amin, L., Islam, M., Mahmud, G. M. I., Chowdhury, I. K., & Mukta, M. S. H. (2024). Towards sustainable AI: a comprehensive framework for Green AI. *Discover Sustainability* 2024 5:1, 5(1), 408-.
<https://doi.org/10.1007/s43621-024-00641-4>
- Thiyagalingam, J., Shankar, M., Fox, G., & Hey, T. (2022). Scientific machine learning benchmarks. *Nature Reviews Physics* 2022 4:6, 4(6), 413–420. <https://doi.org/10.1038/s42254-022-00441-7>
- Vamvakas, D., Papaioannou, I., Tsaknakis, C., Sgouros, T., & Korkas, C. (2025). Generative AI for Sustainable Smart Environments: A Review of Energy Systems, Buildings, and User-Centric Decision-Making. *Energies* 2025, Vol. 18, 18(23). <https://doi.org/10.3390/en18236163>
- Van der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big*

- Data and Society*, 11(1).
<https://doi.org/10.1177/20539517241232630>
- Vos, G., van Eijk, L., Sarnyai, Z., & Azghadi, M. R. (2024). *Stabilizing Machine Learning for Reproducible and Explainable Results: A Novel Validation Approach to Subject-Specific Insights*. <http://arxiv.org/abs/2412.16199>
- Wen, X., Zhao, D., Lv, Z., Zhang, K., & Zhang, Y. (2024). A sustainability-oriented approach for performance assessment of existing buildings and a case study. *Heliyon*, 10(12), e32151.
<https://doi.org/10.1016/j.heliyon.2024.e32151>
- Zhu, J., Yang, K., Guan, N., Yi, X., & Qiu, C. (2023). HCPNet: Learning discriminative prototypes for few-shot remote sensing image scene classification. *International Journal of Applied Earth Observation and Geoinformation*, 123(11), 103447.
<https://doi.org/10.1016/j.jag.2023.103447>
- Zieliński, G., & Więckiewicz, M. (2025). New effect size and sample size guidelines in dentistry. *Dental and Medical Problems*, 62(5), 907–917.
<https://doi.org/10.17219/dmp/210478>