

CRISP-ML(QS): Embedding Sustainability in AI-ML Development Lifecycle

Sahil Garg

Decision Science Area,
Indian Institute of Management Lucknow, Lucknow, Uttar Pradesh, India.
Corresponding author: sahil.garg@iiml.org
ORCID: <https://orcid.org/0009-0002-3346-5093>

Garima Mittal

Decision Science Area,
Indian Institute of Management Lucknow, Lucknow, Uttar Pradesh, India.
E-mail: garima@iiml.ac.in
ORCID: <https://orcid.org/0000-0003-1861-6930>

(Received on February 5, 2026; Revised on May 3, 2026 & June 6, 2026; Accepted on July 15, 2026)

Abstract

Artificial Intelligence (AI) in contemporary industry and society is omnipresent, with its applications permeating across enterprise functions such as manufacturing, supply chain and research & development, and enterprise goals such as efficiency, innovation, and sustainability. However, an important question that has been raised, quite often recently, is: Is AI environmentally sustainable? This study proposes an extension to the Cross-Industry Standard Process for Machine Learning with Quality Assurance (CRISP-ML(Q)), to include 'sustainability' into Artificial Intelligence and Machine Learning (AI-ML) development lifecycle. This paper conducts a systematic literature review with 171 papers to understand the current state of sustainable AI research and synthesizes these studies onto the stages of the model development lifecycle – Business and Data Understanding, Data Preparation, Model Building, Evaluation, Deployment and Monitoring and Maintenance. The proposed framework, Cross-Industry Standard Process for Machine Learning with Quality and Sustainability assurance CRISP-ML(QS), highlights that design choices in various stages of the AI model development lifecycle can have a significant impact on the carbon footprint and energy consumption of the resulting AI system. The framework addresses the environmental dimension of sustainability and focuses on the direct environmental impacts of the AI development lifecycle. Energy-aware model development lifecycle, by utilizing smaller datasets, optimal number of features, algorithms with low energy profiles, efficient neural architecture search, pruning, compression and making efficient model design choices such as modelling frameworks, coding language, batch size, optimizers, training location, can lead to a significant reduction in energy consumption. Energy-aware hyperparameter tuning, deployment, and inference of models also contribute to reducing the environmental impact of AI.

Keywords- Green AI, Sustainability of AI, Energy consumption, Carbon footprint, AI-ML model development lifecycle.

1. Introduction

Artificial Intelligence operates through technologies, machines, and infrastructures (such as data centers) that consume scarce resources during their production, usage, and disposal. It increases energy consumption and hence creates environmental problems (Brevini, 2020). AI now performs at par with, or better than, humans across various tasks, such as computer vision and Natural Language Processing (NLP). However, the excellence in image recognition and NLP has been heavily reliant on substantial increases in computational power. This trend is economically and environmentally unsustainable (Thompson et al., 2021).

Rizvee et al. (2023) compared the CO₂ emissions footprint of daily life activities against training an AI model, and the difference is shocking: training a single AI model emits 5x more CO₂ than an average US car in its entire lifetime, as shown in **Figure 1**.

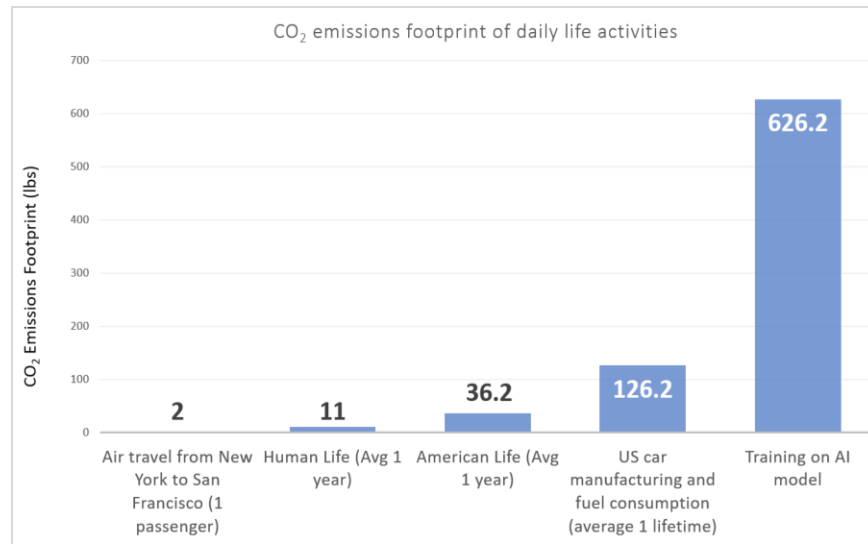


Figure 1. CO₂ emissions footprint of daily life activities vs training an AI model (adapted from Rizvee et al., 2023).

To exacerbate, the size of AI models has been on an exponential rise, especially models for NLP tasks (Bender et al., 2021). Training large language models such as GPT-3 requires substantial computational power, resulting in a notable environmental impact. The model with 175 billion parameters used 10,000 GPUs and emitted 552 metric tons of carbon dioxide (Färber & Lamprecht, 2022).

Sustainability is commonly understood by the Environmental, Social, and Governance (ESG) framework. ESG evaluates sustainability through three distinct lenses:

- **Environmental:** Assessing impact on the natural ecosystem, including carbon intensity, resource depletion, and waste management protocols.
- **Social:** Examining the impact on human capital and communities, focusing on labour standards, diversity and inclusion, and consumer protection.
- **Governance:** Auditing the internal systems of oversight, encompassing board structure, executive compensation, and transparency.

While acknowledging that the Social and Governance pillars are equally important, this study is focused on the Environmental dimension of sustainability.

‘Environmental Sustainability of AI’, also known as Green AI, is the practice of ensuring that the development and use of AI are sustainable, taking into account factors such as carbon footprints and computational power (van Wynsberghe, 2021). Green AI aims to reduce the environmental impact of AI systems by developing energy-efficient algorithms and sustainable computing practices (Thiyagaraj et al., 2026). Green AI also enhances cost-effectiveness and promotes sustainability through energy-efficient models and industry practices (Chaudhary et al., 2025). Green AI is designed to reduce carbon emissions through multiple functionalities, aiming to contribute to a sustainable future (Sowrirajan et al., 2025). This concept differs from ‘AI for Sustainability’, which involves using AI to achieve sustainability goals, such as the United Nations Sustainable Development Goals (SDGs) (Pedersen, 2018).

The literature lists various frameworks and methods for reducing energy footprint. The Sustainability Criteria and Indicators for Artificial Intelligence Systems framework proposed by Rohde et al. (2024) outlines sixty-seven indicators and nineteen criteria of sustainability for evaluating sustainable AI and advocates for consideration of energy consumption during model development, selection of less complex models, and utilization of pretrained models to decrease energy usage. Similarly, while advocating measures to address the environmental impact of information retrieval research, Scells et al. (2022) recommend to Reduce, Reuse and Recycle. Reduce: Focuses on minimizing resource usage through efficient algorithms and avoiding unnecessary computations. Reuse: Encourages the use of pre-trained models, shared datasets, and pre-indexed collections to avoid redundant work. Recycle: Promotes repurposing existing data, models, and code for new tasks, reducing the need for developing everything from scratch. These recommendations apply not just to information retrieval but to other applications of AI as well.

Regulations and operationally efficient, effective frameworks lag behind urgent environmental needs; there is a need for a "sustainable by design" approach to AI (Perucica & Andjelkovic, 2022), which integrates sustainability principles from the outset. AI practitioners should proactively pursue sustainability and feel responsible for developing computationally efficient methods to mitigate the AI carbon footprint and promote environmentally responsible practices (Tamburrini, 2022).

There is a considerable body of research work available to assist practitioners in pursuing sustainability while designing and developing Artificial Intelligence and Machine Learning (AI-ML) solutions. But there exists a significant gap between Green AI research and its adoption in industry (Sikand et al., 2023). This paper seeks to bridge the gap by providing a development framework that thoroughly reviews current sustainable AI research and its synthesis on the model development lifecycle, making it convenient for practitioners to understand and implement.

1.1 Research Questions

The primary Research Questions (RQ) that this review endeavors to address are:

RQ1: *What is the present state of cutting-edge research in Sustainable/Green AI?*

RQ2: *How does this existing literature map to the model development lifecycle CRISP-ML(Q)?*

The first research question (RQ1) aims to examine the current landscape of Green AI research through a comprehensive review of existing literature. The second research question (RQ2) seeks to align the existing literature with the model development lifecycle, with an in-depth analysis of how it maps to the stages in the model development lifecycle. Through addressing these inquiries, the review aims to help practitioners understand the design choices throughout the model development lifecycle to make AI more environmentally sustainable.

The remaining sections of this paper are systematically arranged to facilitate clear and logical progression. In Section 2, we elaborate on the methodology utilized for our study. Section 3 is dedicated to presenting the proposed framework, CRISP-ML(QS), offering a clear mapping of sustainable AI research to the model development lifecycle. Section 4 elaborates on the conclusion of the study, limitations, and the scope of future research.

2. Methodology of Study

This section explains the methodological framework used to identify and evaluate the relevant literature for this review. It specifies the information sources, search strategy, inclusion criteria, and the selection and

data-collection process, applied to ensure the resulting synthesis is transparent and reproducible.

2.1 Information Sources and Search Strategy

The information for this study has been sourced from prominent academic databases - Scopus, ScienceDirect, and Google Scholar. The query is structured to identify articles with titles including terms associated with environmental sustainability ("green" or words starting with "sustainab") and terms related to artificial intelligence and its subsets (AI, ML, artificial intelligence, machine learning, deep learning). Relevant studies published and available up to 3rd February 2026 were extracted from Scopus, ScienceDirect, and Google Scholar. The utilization of diverse databases enhances the reliability and comprehensiveness of the information. The search criteria did not impose any limits on the year of publication, enabling the inclusion of both contemporary and earlier works. **Figure 2** shows the source, query, and count of search results.



Figure 2. Database, query, and count of search results.

2.2 Criteria for Inclusion and Exclusion of Research

The criteria for any research to be a part of this review are based on the relevance and quality of the research. The following are the various Inclusion Criteria (IC) and Exclusion Criteria (EC):

- IC1 - The research pertains to the environmental sustainability of AI/Green AI.
- IC2 - The research focuses on AI and ML development.
- IC3 - The research addresses the software level and the model development lifecycle level.
- EC1 - The research is either in the format of presentations, tutorials, etc.

Inclusion criteria guided the selection process, encompassing specific parameters: IC1 specified the focus on the environmental sustainability of AI, distinguishing studies of Green AI from studies in which AI is used for sustainability. IC2 mandated direct relevance to AI-ML and excluded studies related to IOT, cloud, etc. IC3 filtered onto studies related to the software level and model development lifecycle level, refining the scope to exclude research at the hardware level.

Conversely, exclusion criteria were implemented to maintain rigor: EC1 excluded specific formats such as tutorials and presentations, thereby streamlining the selection of scholarly works best suited for comprehensive analysis within the context of this research. Conference proceedings and book chapters were included in the study to ensure no piece of knowledge was missed. The judicious application of these criteria contributes to the quality of the study's information synthesis.

2.3 Process of Selection and Data Collection

During selection, we rigorously reviewed the search results against the inclusion and exclusion criteria for studies. Both reviewers conducted their assessments independently, and any disagreements or discrepancies were resolved through discussions and mutual agreement. After the initial round of applying ICs and ECs, we deployed backward citation chaining by examining the references cited in our key papers to find relevant literature. Citation chaining was done with a tool called ResearchRabbit (ResearchRabbit, 2025). We used ResearchRabbit to visualize citation relationships between papers and to explore additional relevant literature for this review. By adding the shortlisted papers to a collection on ResearchRabbit, we reviewed the recommendations by clicking on “Earlier Work”. **Figure 3** shows an illustrative paper graph from ResearchRabbit, which represents the papers already present in the collection as black-filled circles and those that are “not present in the collection but are similar” as white-filled circles.

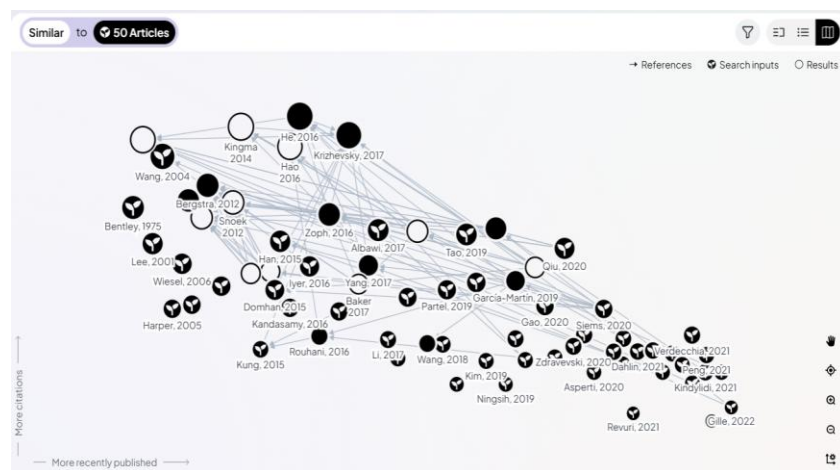


Figure 3. Research rabbit output of paper graph showing connections between the existing collection and other papers.

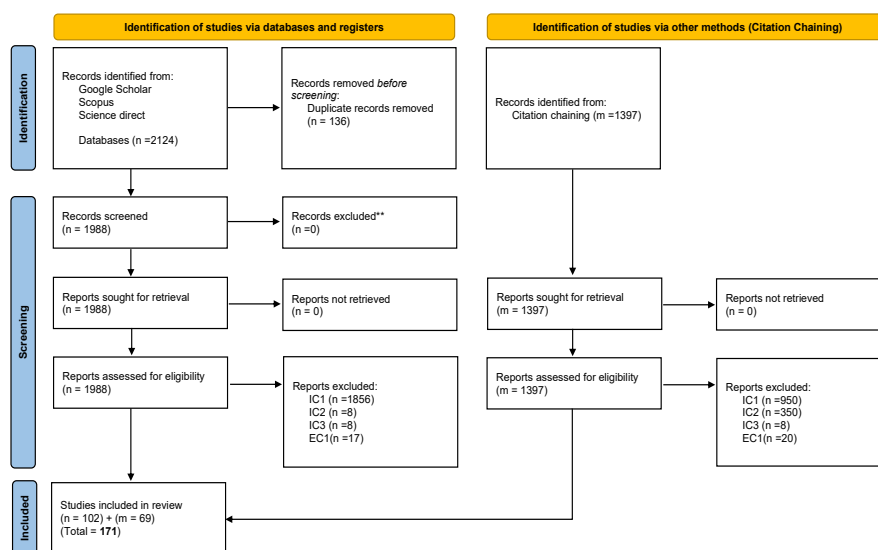


Figure 4. Process of selection and screening on inclusion and exclusion criteria (PRISMA (Page et al., 2021)).

Collectively, we were able to select 171 papers that satisfied the inclusion and exclusion criteria. The process of selection and screening on inclusion and exclusion criteria is summarized in **Figure 4**.

In the data collection phase, a meticulous study was undertaken to extract sustainable AI practices from the papers and map them to the model development lifecycle.

3. Proposed Framework: CRISP-ML(QS)

CRISP-ML(Q) (Studer et al., 2021) is an evolution of the classic CRISP-DM framework (Wirth & Hipp, 2000) and is a widely recognized and structured approach that provides a systematic and comprehensive methodology for machine learning processes. CRISP-ML(Q) framework divides the development lifecycle into six stages: Business and Data Understanding, Data Preparation, Modeling, Evaluation, Deployment and Monitoring and Maintenance. It emphasizes that quality assurance is embedded throughout the development lifecycle.

In this study, we propose CRISP-ML(QS) as an extension to the CRISP-ML(Q) framework with sustainability assurance embedded throughout the development lifecycle. Sustainability assurance methodology recommends integrating "Green AI" principles directly into the existing CRISP-ML(Q) structure. We propose embedding sustainability practices alongside every quality gate to ensure that the carbon footprint and energy efficiency are considered before progressing. The proposed CRISP-ML(QS) framework is presented in **Figure 5**.

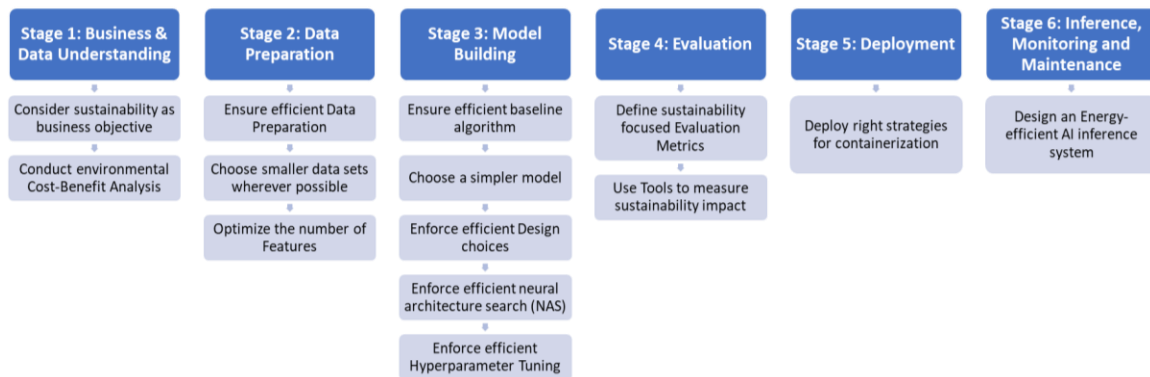


Figure 5. CRISP-ML(QS).

The following sections outline the details of the proposed CRISP-ML(QS) framework.

3.1 Stage 1: Business and Data Understanding

The Business Understanding stage involves the translation of organizational business objectives into a clearly defined AI-ML problem, establishing success criteria and assessing the required resources. This is followed by Data Understanding, which focuses on systematic acquisition, exploration, and quality verification of the data to ensure it is fit for the intended analytical purpose.

The subsequent sub-sections demonstrate how to integrate sustainable AI practices into these foundational stages by categorizing research that integrates environmental considerations into the Business and Data Understanding stage.

3.1.1 Consider Sustainability as a Business Objective

Researchers suggest multiple ways to pursue environmental ‘Sustainability of AI’, including policy and regulatory interventions. Dhar (2020) cites leaders from the Copenhagen Centre on Energy Efficiency to emphasize that environmental sustainability should be regarded as a key principle and objective in the pursuit of responsible AI. Similarly, on examining the CO₂ emissions by the computer vision model, Fu et al. (2021) propose adding "enforcement" as a pillar of ethical AI to ensure CO₂ considerations are incorporated into model development and use. Kindylidi & Cabral (2021) recommends that legislators, such as the EU, consider the impact of AI on sustainability while contemplating the AI regulation. Rohde et al. (2021) advocate comprehensive guidelines and regulation frameworks to ensure that AI development aligns with ecological objectives.

3.1.2 Environmental Cost-Benefit Analysis

An Environmental Cost-Benefit Analysis should be used to make sure that the "environmental cost," which is the energy and carbon used to train the model and keep it running, is less than the "environmental benefit," which is how much energy or carbon the AI system saves. It can help avoid building wasteful, "heavy" models and ensure that every project is environmentally sound.

3.2 Stage 2: Data Preparation

The Data Preparation stage encompasses all activities required to construct the final dataset from the initial raw data. This critical stage involves data selection, cleaning, feature engineering, and integration, ensuring that the resulting data structures are optimally formatted for the modeling algorithms. In a standard lifecycle, the primary objective is to maximize predictive performance and ensure high data quality. The subsequent sections illustrate how sustainable AI practices intersect with this stage.

3.2.1 Efficient Data Preparation

The growth of AI models and their training datasets has been exponential, especially for those used in natural language processing tasks. BERT (released in 2019) used a 16GB data set, which is minimal as compared to the 570GB dataset size of GPT-3 (released in 2020) (Bender et al., 2021). The trend continues with GPT-4, GPT-5 and later versions, which use an unprecedented volume of training tokens. The larger the data, the larger the computation cost.

Al-Jarrah et al. (2015) highlight various methods of sustainable data modeling and efficient learning, including Ensemble models, Local learning strategy, Semi-parametric approximation and Shallow learning. Salehi & Schmeink (2023) review data-centric technologies and highlight techniques, such as active learning, coresets, dataset distillation, knowledge transfer/sharing, curriculum learning, and data augmentation, that can help reduce the energy consumption of AI algorithms. The focus of sustainable data modelling is on fast and efficient processing of huge data volumes and on minimum cost of computation with maximum accuracy (Al-Jarrah et al., 2015).

3.2.2 Choose Smaller Datasets

Bender et al. (2021) suggest focusing resources on curating and carefully documenting datasets rather than indiscriminately gathering data from the web. Using smaller, more significant datasets for training can speed up the process and reduce Greenhouse Gas (GHG) emissions of AI systems (Rizvee et al., 2023). Verdecchia et al. (2022) indicate that utilizing smaller datasets significantly lowers energy consumption in machine learning models, with energy efficiency enhancements reaching as high as 92% upon dataset size reduction. Dhabe et al. (2021) devised a method to eliminate redundant data samples from training data and create smaller, more efficient ‘Bonsai data sets’. Redundancy is determined based on how similar a data

sample is to the average (mean) pattern of its class. Bonsai data sets have been shown to preserve classification and recognition accuracy. For recommendation engines, Eslami & Ghaderi (2023) introduced an incremental trust-aware matrix factorization method that requires learning only a small fraction of new data when a new customer (or data) arrives; the remaining data is leveraged from models developed in earlier stages using existing user-item records.

3.2.3 Optimize the Number of Features

The number of features has an even larger impact on energy consumption than the number of examples in the training data, as observed in decision tree implementations (Ferro et al., 2023). Verdecchia et al. (2022) show that in most cases, reducing the number of features significantly lowers energy usage with minimal impact on accuracy. For instance, by decreasing features, Random Forest can reduce energy use by up to 74.81% with only a 0.06% drop in F1-score. Nevertheless, this pattern is not consistent across all algorithms. Further, selecting features in KNN has little effect on energy use but results in a substantial decrease in performance, achieving only a 0.92% energy saving at the expense of a 98.05% reduction in F1-score. Similarly, in the case of large language models, fewer parameters do not always imply better efficiency. Zhou et al. (2020) found that despite having fewer parameters, the ALBERT model was less efficient than BERT due to its larger hidden size and more complex computations. However, having more parameters can lead to slower fine-tuning when using the same model.

Hence, finding the optimal number of features is important. Li et al. (2020) find an optimal range of granularity (between 60 and 80 features) in a decision tree that minimizes tree depth and computational cost, enhancing decision tree performance while reducing energy consumption. Asperti et al. (2021) provide a survey of how Variational Auto Encoders (VAEs) can be optimized for energy efficiency, to reduce the financial cost and carbon footprint of the generative techniques.

3.3 Stage 3: Energy-Aware Model Building

Model training is amongst the most important stages from a sustainability perspective. The training and development of models will account for a significant part of greenhouse gas emissions (Strubell et al., 2019), and this problem cannot be solved only at a hardware level. An ineffective machine learning model can result in significant overheating and high energy consumption, regardless of the hardware's efficiency.

The process of model training involves several steps. First, a suitable modeling technique is selected, and then a baseline algorithm is selected. In model training, a baseline algorithm is a simple model used as a reference point to evaluate the performance of more complex models. It provides a standard or benchmark against which more sophisticated models can be compared. The next steps involve experimental design (making design choices), algorithm evaluation, selection and validation.

3.3.1 Algorithm Selection: Select an Efficient Baseline Algorithm

Researcher suggests moving away from "accuracy at any cost." McIntosh et al. (2019) suggest that developers should select the optimal algorithm by analyzing application requirements and context, including the number of data attributes, dataset size, and whether the model will need updating, among other factors. Verdecchia et al. (2022) reveal that various algorithms exhibit distinct energy profiles. The disparity in energy usage can reach up to a 99.49% reduction, hence selecting an appropriate baseline algorithm can have a considerable impact on the consumption of energy.

AI practitioners have been focused on developing models that outperform the existing models (Eslami & Ghaderi, 2023). Hanafy et al. (2021) found that with larger models, the incremental gains in accuracy have been achieved at a significantly greater energy expense. The cost of computing for cutting-edge AI research

has surged by 300,000 times due to AI practitioners' focus on accuracy (Schwartz et al., 2019). The pursuit of accuracy beyond the cognitive requirements of the task, referred to as gratuitous generalization capability, is harmful to the environment (Spelda & Stritecky, 2020). For example, in automatic speech recognition, Gondi & Pratap (2021) found that while accuracy (word error rate) improves linearly, energy consumption increases exponentially. Similarly, Parcollet & Ravanelli (2021) showed that minor performance gains come at a substantial carbon cost: a state-of-the-art transformer model was observed to release half of the total CO₂ emitted during training just to achieve a 0.3 reduction in word error rate.

Hence, there is a trade-off between energy efficiency and predictive performance. Castellanos-Nieves & García-Forte (2023) observed a significant disparity in energy consumption and Carbon Dioxide Equivalent (CO₂e) emissions between complex and simpler models. Complex Models such as Gradient Boosting, Random Forest, and Support Vector Machines (SVMs) exhibit high Accuracy and F1 scores, but this is associated with higher energy consumption and CO₂e emissions as compared to simpler models such as Decision Trees, Naive Bayes, and Linear SVMs.

Complex models deliver higher predictive performance but also use more energy and emit more carbon dioxide equivalents. On the other hand, simpler models are more energy-efficient and emit lower levels of carbon, but have shown reduced predictive performance. However, Brownlee et al. (2021) demonstrated that the consumption of energy during training and inference can be significantly reduced at the cost of a minimal loss in accuracy. They also discovered that structural as well as non-structural parameters play a crucial role in the trade-off between predictive performance and energy consumption. Structural parameter like the size of hidden layers is key influencing factor.

3.3.2 Choose a Simpler Model

Comparing the carbon footprint of various models/algorithms alongside accuracy can help practitioners make informed decisions. Taddeo et al. (2021) recommend informing users about the environmental costs of AI models available in popular libraries and platforms, such as PyTorch, TensorFlow, and Hugging Face, by assessing and publishing their carbon footprints. Similarly, Henderson et al. (2020) recommend considering the trade-offs between energy consumption and performance before deploying energy-intensive models.

- Zhou et al. (2020) introduce HULK, a multi-task energy-efficiency benchmarking platform that evaluates the energy efficiency of pretrained NLP models by accounting for time and cost during both training and inference.
- Given the non-linear relationship between energy and performance, Hanafy et al. (2021) introduce a recommendation algorithm that can help practitioners select the best deep learning model that adheres to specific energy budget constraints while optimizing for both accuracy and energy efficiency. Cao et al. (2021) developed Interpretable Energy Prediction for Transformers (IrEne), a tool that predicts inference energy consumption for a diverse array of transformer-based NLP models. It can be used to evaluate different architectural options. In cases where accuracy is paramount for NLP tasks, Ang et al. (2022) found that enlarging the model size increases its energy efficiency as compared to extending the sequence length for summarization tasks.
- A Green AI model selection strategy is proposed by Ezenkwu et al. (2023), which utilizes a multi-criteria decision technique and includes computational time as a factor to identify the optimal model for final deployment.

In situations where comparison of models is not feasible, one must choose a simple model, as suggested by the literature.

- Kshirsagar et al. (2021) implemented You Only Look Once (YOLO) v3, Faster-Region based Convolutional Neural Network (F-RCNN), and Mask-Region based Convolutional Neural Network (Mask RCNN) for vehicle detection and found YOLOv3 to be most energy-efficient amongst them.
- Ang et al. (2022) observed that for question-answering tasks, smaller models achieve higher efficiency and accuracy as they allow for larger training batch sizes within a set resource budget.
- Naser (2023) suggests that a reduction in carbon emissions and energy consumption ranging from 23% to 99% can be achieved by using simpler machine learning models instead of more complex ones while still preserving a similar level of accuracy. Similar observations were made by Yu et al. (2022) for clinical laboratory data sets, on which XGBoost algorithm achieved 16%-24% lower power consumption levels as compared to a five-hidden-layer NN.

3.3.3 Efficient Design Choices

Deep Neural Networks (DNNs) are usually compared with the human brain, but the current AI systems are less efficient than the human brain, which operates on just 20 Watts due to differences in processing architectures and computational paradigms (Rajendran, 2023). Optimal performance in DNNs is achieved with architectures that include millions of parameters, which require high computational resources for training (Barlaud & Guyard, 2020).

Sze et al. (2017) provide a survey on recent work on enabling efficient processing of DNNs. According to them, efficient DNN processing involves a combination of computational transforms, dataflow optimizations, and leveraging advanced hardware capabilities to achieve significant improvements in both throughput and energy efficiency. Strubell et al. (2019) quantify the approximate financial and environmental costs of training a neural network model for NLP tasks with various successful algorithms and finds that large neural network models are not only financially costly (due to the cost of hardware and electricity) but also environmentally costly (due to the carbon footprint of hardware) to develop.

Gong et al. (2019) suggest Neural Architecture Search (NAS) and model quantization as two primary methods for developing compact and energy-efficient DNNs. NAS automates the creation of neural network structures to optimize both accuracy and energy efficiency, whereas model quantization uses arithmetic and low-precision representations to balance efficiency with accuracy.

Ikhlas et al. (2022) offer a comprehensive overview of implementation actions at the hardware, hardware-algorithm, algorithm, and data levels that should be taken to develop energy-efficient DNNs. At the algorithm level, it highlights: 1) techniques for reducing computational complexity, and 2) methods for model size compression and pruning.

The following are additional model training design choices, which influence the environmental sustainability of the AI-ML development lifecycle.

Framework: Empirical studies by Georgiou et al. (2022) show a statistically significant difference in the costs associated with the two deep learning frameworks: TensorFlow and PyTorch. It is observed that TensorFlow generally shows better energy and runtime efficiency during the training phase, especially for models in recommender systems and computer vision tasks. However, PyTorch performs better in the inference phase, particularly for NLP tasks. Yao et al. (2021) found Nvidia TensorRT be 1.53 times more energy-efficient than TensorFlow.

Coding Language: Python implementations consume more energy than C implementations (Ferro et al., 2023).

Batch size: Li et al. (2016a) suggest that an increase in the batch size can lead to a reduction in the consumption of energy per image, and this may continue till the hardware utilization saturates.

Optimizer: Hodak & Dholakia (2021) highlighted that the algorithmic improvements, such as the use of LARS (Layer-wise Adaptive Rate Scaling) optimizer instead of SGD (Stochastic Gradient Descent) with Momentum, significantly reduced the number of epochs required for AI training and helped in attaining faster training times by reducing the computational workload almost by half.

Cloud training: Geographic location is crucial for ML workload scheduling because the proportion of carbon-free energy and the resulting CO₂e emissions can vary by approximately 5 to 10 times, even within the same country and organization (Patterson et al., 2021). Dodge et al. (2022) suggests strategies to minimize emissions on cloud platforms, such as choosing servers from regions with lower carbon intensity, timing operations when carbon output is reduced, and temporarily shutting down instances during peak carbon intensity periods. They advocate transparency in reporting carbon emissions and highlight how strategic timing and location of server operations can greatly reduce the environmental impact of cloud-based AI. Carbon emissions vary significantly across different regions, and the thoughtful selection of server location can profoundly influence the overall carbon footprint of AI operations.

3.3.4 Efficient Neural Architecture Search

Efficient NAS has the potential to reduce GHG emissions of AI systems (Rizvee et al., 2023). By simplifying the layers of a neural network, the energy consumption of a deep learning pipeline can be reduced. Reducing network complexity can cut the overall energy consumption of training by half (Yarally et al., 2023).

DeLight, an automated framework, introduces energy as a design parameter in DNN training and execution, utilizing energy characterization to optimize network size according to physical resources. It demonstrates up to 100-fold energy improvement in various benchmarks compared to prior Deep Learning (DL) solutions (Rouhani et al., 2016). Research has shown that architecture search can help reduce energy consumption.

- Ayinala & Parhi (2013) represents a low-energy architecture for SVMs based on approximate computing, which, on average, consumes 31% less energy than conventional designs.
- While optimizing DNNs for practical applications, Canziani et al. (2016) built E-Net (Efficient-Network), which demonstrates superior parameter utilization compared to models like AlexNet (13x) and VGG (24x).
- To balance computational and memory efficiency with accuracy, Hasanpour et al. (2016) introduced SimpleNet. SimpleNet outperforms complex models such as VGGNet, ResNet, and GoogleNet on various well-known benchmarks, achieving this with 2 to 25 times fewer parameters and operations.
- Elsken et al. (2018) developed LEMONADE, an evolutionary algorithm designed for multi-objective architecture search, to search neural architectures which optimize both predictive performance and resource consumption. It approximates the entire Pareto front of architectures considering multiple objectives.
- Energy-aware Neural Architecture Optimization takes into account the energy cost of splitting different neurons and finds energy-efficient network architectures. Empirical tests show improved accuracy and energy efficiency compared to baseline methods and existing architectures (Wang et al., 2019).
- Zhang et al. (2017) developed a highly computation-efficient convolutional neural networks (CNNs) architecture, ShuffleNet, that achieves approximately thirteen times the speedup of AlexNet while maintaining similar accuracy levels.

- Liu et al. (2018) propose a method for discovering the structure of CNNs using a sequential model-based optimization (SMBO) strategy. This approach involves searching for structures in increasing order of complexity, while concurrently developing a surrogate model to guide the exploration of the structure space.
- Gong et al. (2019) introduced a novel methodology that combines NAS and quantization for end-to-end joint optimization. This method aims to find the best combinations of architectures and bit-widths to improve the accuracy of prediction and reduce the consumption of hardware energy, resulting in greater energy efficiency than current advanced quantization techniques and efficiency-aware NAS methods.
- Thompson et al. (2021) recommend finding high-performing small deep learning architectures using NAS and meta-learning.
- Fu et al. (2021) reveal that the cumulative CO₂ emissions from the widespread use of popular computer vision models like ResNet are substantial. Stamoulis et al. (2018) suggest that CNN architecture can be treated as hyper-parameters while optimizing for energy, accuracy, and communication demands of mobile devices. They show how to use Bayesian optimization to significantly enhance efficiency and reduce energy usage for mobile image classification by up to six times compared to existing best methods. Cacciatore et al. (2022) indicate that it is feasible to develop cost-efficient and high-performing CNNs. It examines three CNNs used for clinical applications; the two most effective networks demonstrate comparable accuracy, but the one designed according to Green AI principles requires only half the Floating-Point Operations of the other.

Pruning and compression are two primary techniques used to make models smaller, faster, and more energy-efficient without significantly sacrificing accuracy. The following section details both the techniques and methods for reducing computation complexity.

Pruning

Liu et al. (2017) introduce network slimming, a method that automatically identifies and prunes insignificant channels during training, resulting in a more compact and efficient network. It takes as input wide and deep network models and outputs thin and compact models with comparable accuracy, which are smaller, have a lower memory footprint, and incur lower computational cost; effectiveness has been demonstrated on models such as VGGNet, ResNet, and DenseNet.

Yang et al. (2016) developed an energy-aware pruning algorithm for CNNs. It reduces AlexNet's energy consumption by 3.7 times and GoogLeNet's by 1.6 times, while maintaining a top-5 accuracy loss of less than 1%. Li et al. (2016b) found that inference costs for VGG-16 can be reduced by up to 34% and for ResNet-110 by up to 38% on CIFAR10 by using pruning filters with a small impact on the output accuracy. Lin et al. (2019) present Structured Sparsity Regularization (SSR), a filter pruning technique that integrates two different structured sparsity regularizers into the original objective function to facilitate adaptive filter pruning. It is shown to achieve 2.1x CPU speedup and 1.6x CPU speedup of AlexNet and GoogLeNet on ImageNet.

Jeon & Kim (2018) demonstrate that traditional convolution operations can be broken down into simpler shift operations and pointwise convolutions. They introduce the Active Shift Layer (ASL), a shift operation that treats the shift amount as a learnable function defined by shift parameters. This method efficiently manages diverse convolutions, improving network accuracy while reducing network parameters and inference time.

Han et al. (2017) reduced the Long Short-Term Memory (LSTM) model size by 20x by using a load-balance-aware pruning method with negligible impact on prediction accuracy. This compression is achieved

through a combination of pruning and quantization: a 10-fold reduction via pruning and an additional 2-fold reduction through quantization. Jiang et al. (2019) propose a log-sum minimization method for pruning a trained network layer by layer, which enhances the compression ratio of the network while maintaining its prediction accuracy.

Compression

Jääskeläinen et al. (2022) suggest that there is a need to move away from the "bigger is better" mindset and to adopt more sustainable approaches. Some of the compression methods detailed in research are:

- One of the methods proposed to deal with the issue of high computational resources is proximal regularization methods (Barlaud & Guyard, 2020). Proximal regularization methods are time-consuming; to overcome this issue, Barlaud & Guyard (2020) introduce a constrained methodology and a general framework for the projection gradient approach. Testing on datasets demonstrates that this projection technique, employing the newly developed L (1,1) structured constraint, significantly reduces both memory usage and computational demand. Experimental results by Gille et al. (2023) also prove that memory and computational costs were notably reduced, while maintaining performance comparable to that of dense networks, with L (1,1) constraint.
- He et al. (2018) introduced AutoML for Model Compression (AMC), which uses reinforcement learning to automate the process of compression and achieves better compression than manual methods. Testing results show that AMC improved the compression ratio for ResNet-50 from $3.4\times$ to $5\times$ without accuracy loss.
- Zhang et al. (2018) proposed a procedure to simplify the structure that removes unnecessary neurons in any DNN layer based on the trained weights. The configurations generated by this process are Pareto-optimal in terms of energy consumption and accuracy. It reduces the cost by up to 33X in energy and 12X in memory with negligible loss in performance (accuracy). Bateni et al. (2019) introduce PredJoule, a framework that optimizes energy efficiency for DNN workloads on GPU-enabled automotive systems by employing a layer-based approach to manage timing and exploit each layer's unique characteristics. Experimental results confirm that PredJoule significantly enhances energy efficiency while maintaining predictable timing without missing deadlines.
- Thompson et al. (2021) recommend to reduce the complexity of computation by using Network Compression and Acceleration. Distillation is one of the techniques of network compression, which aims to transfer knowledge from a larger model to a smaller one. Teacher-student distillation was used by Anderson & Gómez-Rodríguez (2020) to enhance the efficiency of a dependency parser. When the model was distilled down to 20% of its original trainable parameters, the parser was found to be 1.19 times faster at inference time than the baseline model. Jooste et al. (2022) also demonstrated that sequence-level knowledge distillation can be employed to decrease model size and training duration while maintaining performance with minimal loss.
- Convolutional layers are the most resource-intensive in terms of computation (Yarally et al., 2023). Hence, the optimization of energy consumption in CNNs should begin with convolution layers. These layers consume 87% of the total energy, followed by fully connected layers that consume 10% of energy and then by pooling and ReLU layers, which consume 5% of energy (Li et al., 2016a). Zhou et al. (2016) show that the number of neurons can be reduced by adding sparse constraints into the objective function, which results in a smaller number of parameters and a lower memory footprint. Their experiment resulted in a CNN with only 30% of the original neurons without any effect on the top-1 classification accuracy. Alvarez & Salzmann (2016) introduced a method for a deep network that automatically determines the right number of neurons in each layer during training. This method maintains or even enhances the accuracy of the network by reducing the number of parameters by up to 80%.

- In federated learning, Xia et al. (2024) propose model compression to create sparse models with regular structures, significantly reducing the requirements for communication, storage, and computational resources. Kim et al. (2022) introduced a quantized federated learning framework capable of reducing energy consumption by as much as 53% when compared to a traditional federated learning model.

Reduce computation complexity

Multiply-Accumulate (MAC) operations are known to be a source of energy consumption in DNNs. Spiking Neural Networks (SNNs) are an alternative to DNNs, which showcase significant efficiency improvements through the replacement of MAC operations with simpler addition operations (García-Vico & Herrera, 2021). Further, Sorbaro et al. (2020) demonstrate that combining Synaptic Operations (SynOps) loss term and quantization-aware training effectively reduces energy consumption in SNNs with minimal impact on accuracy.

While building the model, DTs consume computation power for checking for optimum splits (García-Martín et al., 2021). Ferro et al. (2023) also found calculating node impurity to be the most energy-consuming function, and hence due to the presence of complex impurity calculation, classification tasks generally are more energy-consuming than regression. Ensemble models, such as Random Forest and XGBoost, also show significant energy consumption in ensemble processes (Ferro et al., 2023).

García-Martín et al. (2021) proposed an energy-efficiency focused Very Fast Decision Trees (VFDT) with *nmin adaptation* method. With this method, the energy consumption of the VFDT algorithm is reduced with only a minor effect on accuracy. *nmin adaptation* speeds up the model's development in branches where split confidence is higher and delays on splits where confidence is lower. This method reduces the high computations involved in split checks while maintaining similar accuracy. In the case of VFDT, García-Martín et al. (2017) identify that the labeling phase, especially with the Naive Bayes Adaptive (NBA) setting, is highly energy-intensive and by tuning parameters such as *nmin* (number of instances observed before splitting), τ (tie threshold), and δ (confidence level for splits), the consumption of energy can be reduced by up to 74.3%. Ferro et al. (2023) conclude that optimizing the parameters of decision tree algorithms can lead to substantial energy savings with only a minor impact on accuracy.

3.3.5 Efficient Hyperparameter Tuning

De Chavannes et al. (2021) examine how hyperparameters affect power consumption and model quality during the training of transformer language models. They suggest that multiple hyperparameters, such as the number of layers, attention heads, positional encodings, activation functions, dropout probabilities, and hidden layer sizes, directly affect the model's power consumption. The number of hidden layers, a high amount of attention heads, and the hidden layer size strongly impact the increase in energy consumption. Adding another piece of evidence, Gutiérrez et al. (2022) showed that the choice of solver in Logistic regression has an impact on energy consumption. LBFGS consumed 961.36 W/s whereas Newton-CG consumed 2,761.71 W/s (~3x) without major impact on accuracy.

Strubell et al. (2019) suggest that training time and sensitivity to hyperparameters of the algorithms should be reported, and computationally efficient algorithms and hardware should be prioritized. The efficiency of models involves a trade-off between accuracy, time to solution, and energy consumption (Yokoyama et al., 2022). They present an architecture optimization and hyperparameter optimization methodology implemented through Genetic Algorithms (GA). Other approaches to lessen energy usage in hyperparameter optimization include:

- Bergstra & Bengio (2012) proved that Random search can yield models of equal or greater quality with a substantially reduced computational demand as compared to grid search. Adding further, Yarally et al. (2023) showed that Bayesian optimization is the most energy-efficient hyperparameter optimization strategy amongst Bayesian optimization, random optimization, and grid search.
- The utilization of different information sources is seen as an important strategy to reduce computational time and energy consumed in hyperparameter optimization. (Candelieri et al., 2021).
- Candelieri et al. (2021) extended the Gaussian Process framework for HPO and proposed AGP-MISO - an Augmented Gaussian Process method exploiting multiple information sources, which solved HPO task with significantly lesser computation time and hence lesser energy usage.
- Chowdhury et al. (2022) present a novel Greedy Hyperparameter Optimization (GHO) algorithm, which adjusts each hyperparameter individually while holding the others constant. This approach aims to find a locally optimal solution that is likely close to the global optimum. The results show that GHO is more than 3 times more energy-efficient and twice as fast as Bayesian Optimization (BO) and Random Search (RS).

3.4 Stage 4: Evaluation

Strubell et al. (2019) quantify the approximate financial and environmental costs involved in training neural network models for NLP tasks with different successful algorithms. A similar analysis is required when evaluating models. Models should not be evaluated solely on accuracy, but also on their energy consumption.

A significant rift exists between the research on sustainable AI and its implementation in industry. One of the reasons is the lack of clear metrics or methods that quantify the "green" in the lifecycle of AI-based systems (Sikand et al., 2023). Färber & Lamprecht, (2022) developed a Green AI ontology for modelling the energy consumption of AI models. The ontology models not only metrics and tools, but also software settings and hardware settings.

García-Martín et al. (2019) provide a survey of methods that estimate the consumption of energy by machine learning (ML) solutions. It provides an overview of state-of-the-art methods for estimating power and consumption of energy, along with recommendations of a method to be used for specific scenarios based on goal, dataset and type of learning related to the ML solution. The "greenness" of AI is often equated with the energy consumption of the system, though estimating this energy use is difficult. García-Martín et al. (2019) tackle the challenge of estimating energy consumption by reviewing major approaches from computer architecture and adapting these methods for use in machine learning applications.

3.4.1 Define Sustainability-Focused Evaluation Metrics

- Schwartz et al. (2019) propose that efficiency should be considered an important Model valuation metric, in addition to accuracy and other related metrics. Verdecchia et al. (2022) also suggests that energy metrics are essential for choosing the most suitable model for a machine learning project.
- Dalgren & Lundeg (2019) proposed five efficiency metrics - Model Size Efficiency, Inference Time Efficiency, Inference Energy Efficiency, Training Time Efficiency and Training Energy Efficiency, for evaluating machine learning algorithms based on energy consumption, time, and model size in relation to the accuracy of the model.
- Lenherr et al. (2021) define "Recognition Efficiency" (RE) metric to comprehensively evaluate accuracy, complexity, and energy consumption and "Training efficiency" to calculate the effectiveness of a deep learning model in using computational and energy resources during the training phase.

- Jurj et al. (2020) introduce four metrics for deep learning: Accuracy Per Consumption (APC) and Accuracy Per Energy Cost (APEC) for inference, and Time To Closest APC (TTCAPC) and Time To Closest APEC (TTCAPEC) for training. These metrics evaluate the model's accuracy while accounting for energy consumption, energy cost, and the time required to achieve the specified training level.
- Sikand et al. (2023) introduce the Green AI quotient, a new metric and a methodology, based on energy-efficient AI research and practices to evaluate the "greenness" of any AI-based system and its development process.

3.4.2 Use Tools to Measure Sustainability Impact

Ikhlass et al. (2022) provide a comprehensive overview of energy and power measurement tools used across various fields of advanced computing architecture, including hardware, software, interfaces and simulators. Bannour et al. (2021) review tools that quantify energy consumption and CO₂ emissions in NLP techniques. It describes the capabilities of each measurement tool and assesses the performance of six tools—Carbon Tracker (CT), Experiment-Impact Tracker (EIT), Green Algorithms, ML CO₂ Impact, Energy Usage, and Cumulator (CMLTRs)—in named-entity recognition experiments conducted on diverse computational platforms. Similarly, Bouza et al. (2023) provides an in-depth comparison of the capabilities, technical requirements, and accuracy of seven tools—Green Algorithms, Code Carbon (CC), Eco2AI (E2), CT, EIT, ML CO₂, and CMLTRs—for measuring the carbon emissions and energy consumption of deep learning models during training.

Software-based energy estimates frequently show significant deviations from hardware power measurements, with an average discrepancy of 20%. Such discrepancies can result in wrong decisions regarding model optimization and deployment. Additionally, the accuracy of energy measurements varies with different hardware configurations, indicating that a universal approach to energy measurement is not sufficient. The research emphasizes the necessity for more precise models to estimate energy that takes into consideration hardware variations and the complex interplay between power consumption and resource usage (Cao et al., 2020).

Choosing the right tool for energy estimation is important, as significant variability was observed among the tools in terms of energy consumption estimates due to differences in measurement focus (process vs. machine), idle power consumption accounting, and usage factors (Bouza et al., 2023). Depending on the context, the following are tools that can be used for energy estimation.

- Yang et al. (2017) developed an online tool (available at energyestimation.mit.edu) to estimate energy consumption layer-by-layer in a DNN, taking into account the shape configuration, bit widths, and values of filter weights and activations. The total energy consumption of the network is the sum of the energy used by each layer.
- Henderson et al. (2020) introduced EIT, a tool that simplifies the consistent reporting and tracking of energy, computing, and emission metrics related to algorithms. They also created a leaderboard of reinforcement learning algorithms that are energy efficient to encourage research on energy-efficient methods.
- The Java Energy Profiler and Optimizer (JEPO) is designed for the Java programming language and helps profile and optimize machine learning source code by automatically measuring energy consumption at the method level. It provides energy-efficient recommendations for handling data types, strings, operators, exceptions, objects, control statements and arrays in Java. When this technique is used to optimize WEKA, the machine learning software, it has shown up to improve energy consumption by 14.46% with only 0.48% drop in accuracy (Kumar et al., 2020).

- Carbon tracker, developed by Anthony et al. (2020), is a tool that monitors and forecasts the energy usage and carbon footprint of training deep learning models, helping practitioners become more conscious of their environmental impact.
- Machine Learning Emissions Calculator estimates the carbon impact of ML, considering factors like the type of energy used, cloud provider, hardware, and training time (Luccioni et al., 2020).
- Lannelongue et al. (2021) developed an online tool named Green Algorithms, designed to help practitioners estimate and report the carbon footprint of their computational processes.
- EnergyVis, a dynamic tool, has been developed to track the energy usage of ML models. It provides multiple coordinated views, allowing practitioners to interactively monitor, view, and compare the energy consumption and carbon footprint metrics (in kWh and CO₂) of their models. Additionally, EnergyVis enables users to explore various hardware options and deployment locations that could potentially reduce their carbon footprints (Shaikh et al., 2021).
- Eco2AI, an open-source Python library, calculates energy consumption by monitoring GPU, CPU, and RAM usage during model training and differentiates itself by focusing on precise regional emissions accounting and process-specific energy consumption (Budenny et al., 2022).

3.5 Stage 5: Deployment: Deploy Right Strategies for Containerization

Yosuf et al. (2021) emphasize that deploying services near data sources using fog computing can significantly decrease both latency and power consumption as compared to deployment in remote cloud data centers. They demonstrate a mixed integer linear programming model to search for the deployment location of DNN inference models in a Cloud Fog Network (CFN) setup. This model enhances energy efficiency by effectively balancing the trade-offs between networking power consumption and processing.

Hampau et al. (2022) analyze the effect of three different strategies for containerization of AI —ONNX Runtime, WebAssembly, and Docker—on CPU usage, time to execute, energy consumption, and memory usage for computer-vision tasks on edge devices. They determine that ONNX Runtime, when used independently of Docker, has the most energy consumption and execution time efficiency.

3.6 Stage 6: Inference, Monitoring and Maintenance: Energy-efficient AI Inference System

Patterson et al. (2021) demonstrated that, in large-scale production ML deployments, it is essential to minimize total emissions from both the training and inference stages, instead of concentrating only on the training stage.

Wu et al. (2022) observed that the distribution of power capacity, at Meta, for AI infrastructures across the three primary phases, namely experimentation, training, and inference, is approximately 10:20:70, respectively, highlighting that inference consumes significantly high power.

Yarally et al. (2023) analyzed how batching of input affects the consumption of energy and response times of five fully trained, most recent neural networks for computer vision. They discovered that larger batch sizes generally improve energy efficiency across networks, though the ideal batch size differs for each network.

An energy-efficient edge AI inference system was created by Yang et al. (2020) through the combination of inference task selection and transmit beamforming vector optimization. This approach aims to reduce both computation power consumption and downlink transmission power consumption. Hashemi et al. (2017) examined different bit precisions in neural networks, with floating-point and fixed-point and powers of binary and two. They found that precision quantization can substantially reduce both memory footprint

and energy consumption. Using mixed precision (FP16 and INT8) significantly improves energy efficiency (3.36 times on average) with only a minimal reduction in accuracy (0.1%) (Yao et al., 2021).

Gondi & Pratap (2021) showed that mobile optimization and quantization techniques of PyTorch can reduce the model size and computational requirements of transformer-based ASR (Automatic Speech Recognition) systems on edge devices, while maintaining accuracy levels similar to those of server-based inference. When optimized using the TensorFlow Lite platform, deep architectures consume significantly fewer resources, including CPU, memory, and energy, compared to their original models (Bursa et al., 2022).

AutoScale is an adaptive, lightweight execution-scaling engine for edge-based deep learning inference, powered by a custom reinforcement learning algorithm. It dynamically determines the most energy-efficient inference target by evaluating neural network characteristics and available systems in a collaborative cloud-edge environment and adjusts to runtime changes. AutoScale achieves an average energy efficiency improvement of 9.8x over baseline mobile CPUs and 1.6x over cloud offloading, while maintaining real-time performance and accuracy standards (Kim & Wu, 2020).

In sum, the results detailed in this section demonstrate that various strategies yield quantifiable reductions in AI's total environmental footprint. These findings reinforce the need to embed sustainability into the AI-ML development lifecycle. The following section synthesizes these insights into a conclusion and recommendations for future research.

4. Conclusion

This study proposes a CRISP-ML(QS) framework to embed sustainable practices into the AI-ML development lifecycle. Sustainability assurance is woven into every stage of the model development lifecycle. Paper highlights research as evidence to the fact that significant improvements can be made to energy consumption of an AI system by utilizing smaller datasets, optimal number of features, algorithms with low energy profiles, efficient neural architecture search, pruning, compression and making efficient model design choices such as modelling frameworks, coding language, batch size, optimizers, training location. Energy-aware hyperparameter tuning, deployment and inference of models also contribute to the reduction in the environmental impact of AI.

The framework currently focuses only on the direct environmental impacts of the AI development lifecycle, leading to an underestimation of the true environmental impact of AI services. Emissions originating from the production and end-of-life stages are either overlooked or incompletely considered (Bannour et al., 2021). As part of future research, researchers should adopt Life Cycle Assessment (LCA) to evaluate the full scope of an AI service's environmental impact to encompass the entire lifecycle of a service, from production through use to disposal (Ligozat et al., 2022).

The framework is also limited to the environmental dimension and excludes the broader Social and Governance dimensions of the ESG framework. Consequently, the scope of future research involves operationalizing these omitted pillars to develop a multi-dimensional perspective. Such an approach will enable a holistic and comprehensive paradigm for sustainable AI development.

Conflicts of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors would like to thank the editor and anonymous reviewers for their comments that help improve the quality of this work.

AI Disclosure

During the preparation of this work the author(s) used generative AI in order to improve the language of the article. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

- Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., & Taha, K. (2015). Efficient machine learning for big data: a review. *Big Data Research*, 2(3), 87-93. <https://doi.org/10.1016/j.bdr.2015.04.001>
- Alvarez, J.M., & Salzmann, M. (2016). Learning the number of neurons in deep networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2270-2278). Curran Associates Inc. New York, USA. <https://doi.org/10.48550/arxiv.1611.06321>
- Anderson, M., & Gómez-Rodríguez, C. (2020). Distilling neural networks for greener and faster dependency parsing. In *Proceedings of the 16th International Conference on Parsing Technologies and The IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies* (pp. 2-13). <https://doi.org/10.18653/v1/2020.iwpt-1.2>
- Ang, P., Dhingra, B., & Wills, L.W. (2022). Characterizing the efficiency vs. accuracy trade-off for long-context NLP models. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP* (pp. 113-121). <https://doi.org/10.18653/v1/2022.nlppower-1.12>
- Anthony, L.F.W., Kanding, B., & Selvan, R. (2020). Carbontracker: tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*. <https://doi.org/10.48550/arxiv.2007.03051>
- Asperti, A., Evangelista, D., & Piccolomini, E.L. (2021). A survey on variational autoencoders from a green AI perspective. *SN Computer Science*, 2(4), 301. <https://doi.org/10.1007/s42979-021-00702-9>
- Ayinala, M., & Parhi, K.K. (2013). Low-energy architectures for support vector machine computation. In *2013 Asilomar Conference on Signals, Systems and Computers* (pp. 2167-2171). IEEE. USA. <https://doi.org/10.1109/acssc.2013.6810693>
- Bannour, N., Ghannay, S., Névél, A., & Ligozat, A.L. (2021). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing* (pp. 11-21). ACL. Virtual event. <https://doi.org/10.18653/v1/2021.sustainlp-1.2>
- Barlaud, M., & Guyard, F. (2020). Learning sparse deep neural networks using efficient structured projections on convex constraints for green AI. In *2020 25th International Conference on Pattern Recognition* (pp. 1566-1573). IEEE. Milan, Italy. <https://doi.org/10.1109/icpr48806.2021.9412162>
- Bateni, S., Zhou, H., Zhu, Y., & Liu, C. (2019). PredJoule: a timing-predictable energy optimization framework for deep neural networks. In *2018 IEEE Real-Time Systems Symposium* (pp. 107-118). IEEE. Tennessee, USA. <https://doi.org/10.1109/rtss.2018.00020>
- Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). ACM. New York, USA. <https://doi.org/10.1145/3442188.3445922>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305. <https://dl.acm.org/doi/10.5555/2188385.2188395>
- Bouza, L., Bugeau, A., & Lannelongue, L. (2023). How to estimate carbon footprint when training deep learning models? a guide and review. *Environmental Research Communications*, 5(11), 115014. <https://doi.org/10.1088/2515-7620/acf81b>

- Brevini, B. (2020). Black boxes, not green: mythologizing artificial intelligence and omitting the environment. *Big Data & Society*, 7(2), 2053951720935141. <https://doi.org/10.1177/2053951720935141>
- Brownlee, A.E., Adair, J., Haraldsson, S.O., & Jabbo, J. (2021). Exploring the accuracy–energy trade-off in machine learning. In *2021 IEEE/ACM International Workshop on Genetic Improvement* (pp. 11-18). IEEE. Madrid, Spain. <https://doi.org/10.1109/gi52543.2021.00011>
- Budenny, S.A., Lazarev, V.D., Zakharenko, N.N., Korovin, A.N., Plosskaya, O.A., Dimitrov, D.V., Akhripkin, V.S., Pavlov, I.V., Oseledets, I.V., Barsola, I.S., Egorov, I.V., Kosterina, A.A., & Zhukov, L.E. (2022). Eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI. *Doklady Mathematics*, 106, S118-S128. <https://doi.org/10.1134/s1064562422060230>
- Bursa, S.O., Incel, O.D., & Alptekin, G.I. (2022). Transforming deep learning models for resource-efficient activity recognition on mobile devices. In *2022 5th Conference on Cloud and Internet of Things* (pp. 83-89). IEEE. Marrakech, Morocco. <https://doi.org/10.1109/ciot53061.2022.9766512>
- Cacciatore, A., Migliorelli, L., Berardini, D., Tiribelli, S., Pigliapoco, S., & Moccia, S. (2022). Some ethical remarks on deep learning-based movements monitoring for preterm infants: green AI or red AI? In *2022 International Conference on Image Analysis and Processing* (pp. 165-175). Springer. Lecce, Italy. https://dl.acm.org/doi/abs/10.1007/978-3-031-13324-4_15
- Candelieri, A., Perego, R., & Archetti, F. (2021). Green machine learning via augmented Gaussian processes and multi-information source optimization. *Soft Computing*, 25(19), 12591-12603. <https://doi.org/10.1007/s00500-021-05684-7>
- Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*. <https://doi.org/10.48550/arxiv.1605.07678>
- Cao, Q., Balasubramanian, A., & Balasubramanian, N. (2020). Towards accurate and reliable energy measurement of NLP models. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing* (pp. 141-148). ACL. <https://doi.org/10.18653/v1/2020.sustainlp-1.19>
- Cao, Q., Lal, Y.K., Trivedi, H., Balasubramanian, A., & Balasubramanian, N. (2021). IrEne: interpretable energy prediction for transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 2145-2157). ACL. <https://doi.org/10.18653/v1/2021.acl-long.167>
- Castellanos-Nieves, D., & García-Forte, L. (2023). Improving automated machine-learning systems through green AI. *Applied Sciences*, 13(20), 11583. <https://doi.org/10.3390/app132011583>
- Chaudhary, D., Sharma, R., Verma, A., Srivastava, P., Gupta, R., & Bhati, R. (2025). Energy-efficient artificial intelligence: leveraging green technologies for a sustainable digital future. In *2025 International Conference on Engineering, Technology & Management* (pp. 1-6). IEEE. New York, USA. <https://doi.org/10.1109/icetm63734.2025.11051895>
- Chowdhury, A.A., Hossen, M.A., Azam, M.A., & Rahman, M.H. (2022). Deepqgho: quantized greedy hyperparameter optimization in deep neural networks for on-the-fly learning. *IEEE Access*, 10, 6407-6416.
- Dalgren, A., & Lundeg, Y. (2019). GreenML: a methodology for fair evaluation of machine learning algorithms with respect to resource consumption. *Semantic Scholar*. <https://api.semanticscholar.org/corpusid:202707656>
- de Chavannes, L.H.P., Kongsbak, M.G.K., Rantza, T., & Derczynski, L. (2021). Hyperparameter power impact in transformer language model training. In *2021 Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing* (pp. 96-118). ACL. <https://doi.org/10.18653/v1/2021.sustainlp-1.12>
- Dhabe, P., Mirani, P., Chugwani, R., & Gandewar, S. (2021). Data set reduction to improve computing efficiency and energy consumption in healthcare domain. In: Eilu, E., Baguma, R., Pettersson, J.S., Bhutkar, G.D. (eds) *Digital Literacy and Socio-Cultural Acceptance of ICT in Developing Countries*. Springer, Cham. pp. 53-64. https://doi.org/10.1007/978-3-030-61089-0_4

- Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2(8), 423-425. <https://doi.org/10.1038/s42256-020-0219-9>
- Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A.S., Smith, N.A., DeCario, N., & Buchanan, W. (2022). Measuring the carbon intensity of AI in cloud instances. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (pp. 1877-1894). ACM. New York, USA. <https://doi.org/10.1145/3531146.3533234>
- Elsken, T., Metzen, J.H., & Hutter, F. (2018). Efficient multi-objective neural architecture search via lamarckian evolution. *arXiv preprint arXiv:1804.09081*. <https://doi.org/10.48550/arxiv.1804.09081>
- Eslami, G., & Ghaderi, F. (2023). Incremental trust-aware matrix factorization for recommender systems: towards Green AI. *Applied Intelligence*, 53(10), 12599-12612. <https://doi.org/10.1007/s10489-022-04150-7>
- Ezenkwu, C.P., Stephen, B.U.A., Affiah, I., & Daniel, B. (2023). A green AI model selection strategy for computer-aided mpox detection. In *2023 IEEE AFRICON* (pp. 1-6). IEEE. Nairobi, Kenya. <https://doi.org/10.1109/africon55910.2023.10293707>
- Färber, M., & Lamprecht, D. (2022). The green AI ontology: an ontology for modeling the energy consumption of AI models. In *2022 21st International Semantic Web Conference*. CEUR.
- Ferro, M., Silva, G.D., de Paula, F.B., Vieira, V., & Schulze, B. (2023). Towards a sustainable artificial intelligence: a case study of energy efficiency in decision tree algorithms. *Concurrency and Computation: Practice and Experience*, 35(17), e6815. <https://doi.org/10.1002/cpe.6815>
- Fu, A., Hosseini, M.S., & Plataniotis, K.N. (2021). Reconsidering co2 emissions from computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2311-2317). <https://doi.org/10.48550/arxiv.2104.08702>
- Garcia-Martin, E., Lavesson, N., & Grahm, H. (2017). Identification of energy hotspots: a case study of the very fast decision tree. In *2017 International Conference on Green, Pervasive, and Cloud Computing* (pp. 267-281). Springer. https://doi.org/10.1007/978-3-319-57186-7_21
- García-Martín, E., Lavesson, N., Grahm, H., Casalicchio, E., & Boeva, V. (2021). Energy-aware very fast decision tree. *International Journal of Data Science and Analytics*, 11(2), 105-126. <https://doi.org/10.1007/s41060-021-00246-4>
- García-Martín, E., Rodrigues, C.F., Riley, G., & Grahm, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134, 75-88. <https://doi.org/10.1016/j.jpdc.2019.07.007>
- García-Vico, Á.M., & Herrera, F. (2021). A preliminary analysis on software frameworks for the development of spiking neural networks. In *2021 International Conference on Hybrid Artificial Intelligence Systems* (pp. 564-575). Springer. Switzerland. https://doi.org/10.1007/978-3-030-86271-8_47
- Georgiou, S., Kechagia, M., Sharma, T., Sarro, F., & Zou, Y. (2022). Green AI: do deep learning frameworks have different costs? In *Proceedings of the 44th International Conference on Software Engineering* (pp. 1082-1094). ACM. New York, USA. <https://doi.org/10.1145/3510003.3510221>
- Gille, C., Guyard, F., Antonini, M., & Barlaud, M. (2023). Learning sparse auto-encoders for green AI image coding. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1-5). IEEE. Rhodes, Greece. <https://doi.org/10.1109/icassp49357.2023.10096638>
- Gondi, S., & Pratap, V. (2021). Performance and efficiency evaluation of ASR inference on the edge. *Sustainability*, 13(22), 12392. <https://doi.org/10.3390/su132212392>
- Gong, C., Jiang, Z., Wang, D., Lin, Y., Liu, Q., & Pan, D.Z. (2019). Mixed precision neural architecture search for energy efficient deep learning. In *2019 IEEE/ACM International Conference on Computer-Aided Design* (pp. 1-7). IEEE. Colorado, USA. <https://doi.org/10.1109/iccad45719.2019.8942147>

- Gutiérrez, M., Moraga, M.Á., & García, F. (2022). Analysing the energy impact of different optimisations for machine learning models. In *2022 International Conference on ICT for Sustainability* (pp. 46-52). IEEE. Plovdiv, Bulgaria. <https://doi.org/10.1109/ict4s55073.2022.00016>
- Hampau, R.M., Kaptein, M., Van Emden, R., Rost, T., & Malavolta, I. (2022). An empirical study on the performance and energy consumption of AI containerization strategies for computer-vision tasks on the edge. In *2022 Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (pp. 50-59). ACM. New York, USA. <https://doi.org/10.1145/3530019.3530025>
- Han, S., Kang, J., Mao, H., Hu, Y., Li, X., Li, Y., Xie, D., Luo, H., Yao, S., Wang, Y., Yang, H., & Dally, W.B.J. (2017). ESE: efficient speech recognition engine with sparse LSTM on FPGA. In *2017 Proceedings of the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (pp. 75-84). ACM. New York, USA. <https://doi.org/10.1145/3020078.3021745>
- Hanafy, W.A., Molom-Ochir, T., & Shenoy, R. (2021). Design considerations for energy-efficient inference on edge devices. In *2021 Proceedings of the 12th ACM International Conference on Future Energy Systems* (pp. 302-308). ACM. New York, USA. <https://doi.org/10.1145/3447555.3465326>
- Hasanpour, S.H., Rouhani, M., Fayyaz, M., & Sabokrou, M. (2016). Lets keep it simple, using simple architectures to outperform deeper and more complex architectures. *arXiv preprint arXiv:1608.06037*. <https://doi.org/10.48550/arxiv.1608.06037>
- Hashemi, S., Anthony, N., Tann, H., Bahar, R.I., & Reda, S. (2017). Understanding the impact of precision quantization on the accuracy and energy of neural networks. In *2017 Design, Automation & Test in Europe Conference & Exhibition* (pp. 1474-1479). IEEE. Lausanne, Switzerland. <https://doi.org/10.23919/date.2017.7927224>
- He, Y., Lin, J., Liu, Z., Wang, H., Li, L.J., & Han, S. (2018). AMC: AutoML for model compression and acceleration on mobile devices. In *2018 Proceedings of the European Conference on Computer Vision* (pp. 784-800). Springer. Munich, Germany. https://doi.org/10.1007/978-3-030-01234-2_48
- Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248), 1-43.
- Hodak, M., & Dholakia, A. (2021). Recent efficiency gains in deep learning: performance, power, and sustainability. In *2021 IEEE International Conference on Big Data* (pp. 2040-2045). IEEE. Florida, USA. <https://doi.org/10.1109/bigdata52589.2021.9671762>
- Ikhlasse, H., Benjamin, D., Vincent, C., & Hicham, M. (2022). Recent implications towards sustainable and energy efficient AI and big data implementations in cloud-fog systems: a newsworthy inquiry. *Journal of King Saud University-Computer and Information Sciences*, 34(10), 8867-8887. <https://doi.org/10.1016/j.jksuci.2021.11.002>
- Jääskeläinen, P., Pargman, D., & Holzapfel, A. (2022). On the environmental sustainability of Ai art (s). In *8th Workshop on Computing within Limits* (pp. 1-9). ACM. New York, USA.
- Jeon, Y., & Kim, J. (2018). Constructing fast network through deconstruction of convolution. *arXiv preprint arXiv:1806.07370*. <https://doi.org/10.48550/arxiv.1806.07370>
- Jiang, T., Yang, X., Shi, Y., & Wang, H. (2019). Layer-wise deep neural network pruning via iteratively reweighted optimization. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5606-5610). IEEE. Brighton, UK. <https://doi.org/10.1109/icassp.2019.8682464>
- Jooste, W., Haque, R., & Way, A. (2022). Knowledge distillation: a method for making neural machine translation more efficient. *Information*, 13(2), 88. <https://doi.org/10.3390/info13020088>
- Jurj, S.L., Opritoiu, F., & Vladutiu, M. (2020). Environmentally-friendly metrics for evaluating the performance of deep learning models and systems. In *2020 International Conference on Neural Information Processing* (pp. 232-244). Springer. Bangkok, Thailand. https://doi.org/10.1007/978-3-030-63836-8_20

- Kim, M., Saad, W., Mozaffari, M., & Debbah, M. (2022). On the tradeoff between energy, precision, and accuracy in federated quantized neural networks. In *2022 IEEE International Conference on Communications* (pp. 2194-2199). IEEE. Seoul, South Korea. <https://doi.org/10.1109/icc45855.2022.9838362>
- Kim, Y.G., & Wu, C.J. (2020). Autoscale: energy efficiency optimization for stochastic edge inference using reinforcement learning. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 1082-1096). IEEE. Athens, Greece. <https://doi.org/10.1109/micro50266.2020.00090>
- Kindylidi, I., & Cabral, T.S. (2021). Sustainability of AI: The case of provision of information to consumers. *Sustainability*, 13(21), 12064. <https://doi.org/10.3390/su132112064>
- Kshirsagar, M., More, T., Lahoti, R., Adgaonkar, S., Jain, S., Ryan, C., & Kshirsagar, V. (2021). GREE-COCO: green artificial intelligence powered cost pricing models for congestion control. In *2021 Proceedings of the 13th International Conference on Agents and Artificial Intelligence* (Vol. 2, pp. 916-923). SciTePress. <https://doi.org/10.5220/0010261209160923>
- Kumar, M., Zhang, X., Liu, L., Wang, Y., & Shi, W. (2020). Energy-efficient machine learning on the edges. In *2020 IEEE International Parallel and Distributed Processing Symposium Workshops* (pp. 912-921). IEEE. LA, USA. <https://doi.org/10.1109/ipdpsw50202.2020.00153>
- Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: quantifying the carbon footprint of computation. *Advanced Science*, 8(12), 2100707. <https://doi.org/10.1002/advs.202100707>
- Lenherr, N., Pawlitzek, R., & Michel, B. (2021). New universal sustainability metrics to assess edge intelligence. *Sustainable Computing: Informatics and Systems*, 31, 100580. <https://doi.org/10.1016/j.suscom.2021.100580>
- Li, D., Chen, X., Becchi, M., & Zong, Z. (2016a). Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In *2016 IEEE International Conferences on Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications* (pp. 477-484). IEEE. Georgia, USA. <https://doi.org/10.1109/bdcloud-socialcom-sustaincom.2016.76>
- Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H.P. (2016b). Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*. <https://doi.org/10.48550/arxiv.1608.08710>
- Li, T., Ma, L., Liu, Z., & Liang, K. (2020). Economic granularity interval in decision tree algorithm standardization from an open innovation perspective: towards a platform for sustainable matching. *Journal of Open Innovation: Technology, Market, and Complexity*, 6(4), 149. <https://doi.org/10.3390/joitmc6040149>
- Ligozat, A.L., Lefevre, J., Bugeau, A., & Combaz, J. (2022). Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability*, 14(9), 5172. <https://doi.org/10.3390/su14095172>
- Lin, S., Ji, R., Li, Y., Deng, C., & Li, X. (2019). Toward compact convnets via structure-sparsity regularized filter pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2), 574-588. <https://doi.org/10.1109/tnnls.2019.2906563>
- Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.J., Fei-Fei, L., Yuille, A., Huang, J., & Murphy, K. (2018). Progressive neural architecture search. In *2018 Proceedings of the European Conference on Computer Vision* (pp. 19-34). Springer. Munich, Germany. https://doi.org/10.1007/978-3-030-01246-5_2
- Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In *2017 Proceedings of the IEEE International Conference on Computer Vision* (pp. 2736-2744). IEEE. Venice, Italy. <https://doi.org/10.1109/iccv.2017.298>
- Luccioni, A., Lacoste, A., & Schmidt, V. (2020). Estimating carbon emissions of artificial intelligence [opinion]. *IEEE Technology and Society Magazine*, 39(2), 48-51. <https://doi.org/10.1109/mts.2020.2991496>
- McIntosh, A., Hassan, S., & Hindle, A. (2019). What can android mobile app developers do about the energy consumption of machine learning?. *Empirical Software Engineering*, 24(2), 562-601. <https://doi.org/10.1007/s10664-018-9629-2>

- Naser, M.Z. (2023). 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>
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *The BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Parcollet, T., & Ravanelli, M. (2021). The energy and carbon footprint of training end-to-end speech recognizers. In *2021 Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 4583-4587). ISCA. Brno, Czech Republic. <https://doi.org/10.21437/Interspeech.2021-456>
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*. <https://doi.org/10.48550/arxiv.2104.10350>
- Pedersen, C.S. (2018). The UN sustainable development goals (SDGs) are a great gift to business!. *Procedia CIRP*, 69, 21-24. <https://doi.org/10.1016/j.procir.2018.01.003>
- Perucica, N., & Andjelkovic, K. (2022). Is the future of AI sustainable? a case study of the European Union. *Transforming Government: People, Process and Policy*, 16(3), 347-358. <https://doi.org/10.1108/tg-06-2021-0106>
- Rajendran, B. (2023). Building a smart and green AI. *The Electrochemical Society Interface*, 32(1), 47-48. <https://doi.org/10.1149/2.f09231if>
- ResearchRabbit. (2025). ResearchRabbit (Version 2025-11-18) [Search tool]. <https://app.researchrabbit.ai/>
- Rizvee, M.M., Rahman, M.H., Chakraborty, P., & Shomaji, S. (2023). Understanding the innovations required for a green & secure artificial intelligence paradigm. In *2023 IEEE 16th Dallas Circuits and Systems Conference* (pp. 1-6). IEEE. Texas, USA. <https://doi.org/10.1109/dcas57389.2023.10130257>
- Rohde, F., Gossen, M., Wagner, J., & Santarius, T. (2021). Sustainability challenges of artificial intelligence and policy implications. *Ökologisches Wirtschaften-Fachzeitschrift*, 36(01), 36-40. <https://doi.org/10.14512/oewo360136>
- Rohde, F., Wagner, J., Meyer, A., Reinhard, P., Voss, M., Petschow, U., & Mollen, A. (2024). Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for Artificial Intelligence systems. *Current Opinion in Environmental Sustainability*, 66, 101411. <https://doi.org/10.1016/j.cosust.2023.101411>
- Rouhani, B.D., Mirhoseini, A., & Koushanfar, F. (2016). Delight: adding energy dimension to deep neural networks. In *Proceedings of the 2016 International Symposium on Low Power Electronics and Design* (pp. 112-117). ACM. New York, USA. <https://doi.org/10.1145/2934583.2934599>
- Salehi, S., & Schmeink, A. (2023). Data-centric green artificial intelligence: a survey. *IEEE Transactions on Artificial Intelligence*, 5(5), 1973-1989. <https://doi.org/10.1109/tai.2023.3315272>
- Scells, H., Zhuang, S., & Zuccon, G. (2022). Reduce, reuse, recycle: green information retrieval research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2825-2837). ACM. New York, USA. <https://doi.org/10.1145/3477495.3531766>
- Schwartz, R., Dodge, J., Smith, N.A., & Etzioni, O. (2019). Green AI. *arXiv preprint arXiv:1907.10597*. <https://doi.org/10.48550/arxiv.1907.10597>
- Shaikh, O., Saad-Falcon, J., Wright, A.P., Das, N., Freitas, S., Asensio, O., & Chau, D.H. (2021). EnergyVis: interactively tracking and exploring energy consumption for ML models. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). ACM. New York, USA. <https://doi.org/10.1145/3411763.3451780>

- Sikand, S., Sharma, V.S., Kaulgud, V., & Podder, S. (2023, September). Green AI quotient: assessing greenness of AI-based software and the way forward. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering* (pp. 1828-1833). IEEE. Luxembourg. <https://doi.org/10.1109/ase56229.2023.00115>
- Sorbaro, M., Liu, Q., Bortone, M., & Sheik, S. (2020). Optimizing the energy consumption of spiking neural networks for neuromorphic applications. *Frontiers in Neuroscience*, 14, 662. <https://doi.org/10.3389/fnins.2020.00662>
- Sowrirajan, S.R., Reddy, P.K.K., & Poojitha, P.S. (2025). Green AI: a comprehensive approach to carbon emission reduction. In *2025 International Conference on Inventive Computation Technologies* (pp. 1067-1071). IEEE. Kirtipur, Nepal. <https://doi.org/10.1109/icit64420.2025.11004680>
- Spelda, P., & Stritecky, V. (2020). The future of human-artificial intelligence nexus and its environmental costs. *Futures*, 117, 102531. <https://doi.org/10.1016/j.futures.2020.102531>
- Stamoulis, D., Chin, T.W., Prakash, A.K., Fang, H., Sajja, S., Bogнар, M., & Marculescu, D. (2018). Designing adaptive neural networks for energy-constrained image classification. In *Proceedings of the International Conference on Computer-Aided Design* (pp. 1-8). ACM. New York, USA. <https://doi.org/10.1145/3240765.3240796>
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650). ACL. Florence, Italy. <https://doi.org/10.18653/v1/P19-1355>
- Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.R. (2021). Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(2), 392-413. <https://doi.org/10.3390/make3020020>
- Sze, V., Chen, Y.H., Yang, T.J., & Emer, J.S. (2017). Efficient processing of deep neural networks: a tutorial and survey. *Proceedings of the IEEE*, 105(12), 2295-2329. <https://doi.org/10.1109/jproc.2017.2761740>
- Taddeo, M., Tsamados, A., Cowls, J., & Floridi, L. (2021). Artificial intelligence and the climate emergency: opportunities, challenges, and recommendations. *One Earth*, 4(6), 776-779. <https://doi.org/10.1016/j.oneear.2021.05.018>
- Tamburrini, G. (2022). The AI carbon footprint and responsibilities of AI scientists. *Philosophies*, 7(1), 4. <https://doi.org/10.3390/philosophies7010004>
- Thiyagaraj, P.B., Priya J.D., & Kavitha, S. (2026). Green AI: energy-efficient artificial intelligence for sustainable computing. *International Journal of Scientific Innovation in Engineering*, 3(2), 636-641. <https://doi.org/10.70849/ijsci03022679082>
- Thompson, N.C., Greenewald, K., Lee, K., & Manso, G.F. (2021). The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10(2). <https://doi.org/10.48550/arxiv.2007.05558>
- Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3), 213-218. <https://doi.org/10.1007/s43681-021-00043-6>
- Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., & Hotellier, E. (2022). Data-centric green AI an exploratory empirical study. In *2022 International Conference on ICT for Sustainability* (pp. 35-45). IEEE. Plovdiv, Bulgaria. <https://doi.org/10.1109/ict4s55073.2022.00015>
- Wang, D., Li, M., Wu, L., Chandra, V., & Liu, Q. (2019). Energy-aware neural architecture optimization with fast splitting steepest descent. *arXiv preprint arXiv:1910.03103*. <https://doi.org/10.48550/arxiv.1910.03103>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: towards a standard process model for data mining. In *2000 Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (Vol. 1, pp. 29-39). PAC. England, UK.

- Wu, C.J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F.A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.H.S., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., & Hazelwood, K. (2022). Sustainable AI: environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4, 795-813.
- Xia, C., Tsang, D.H., & Lau, V.K. (2024). structured bayesian federated learning for green AI: a decentralized model compression using turbo-vbi-based approach. *IEEE Internet of Things Journal*, 11(7), 12783-12798.
- Yang, T.J., Chen, Y.H., & Sze, V. (2016). Designing energy-efficient convolutional neural networks using energy-aware pruning. *arXiv preprint arXiv:1611.05128*. <http://arxiv.org/abs/1611.05128>
- Yang, T.J., Chen, Y.H., Emer, J., & Sze, V. (2017). A method to estimate the energy consumption of deep neural networks. In *2017 51st Asilomar Conference on Signals, Systems, and Computers* (pp. 1916-1920). IEEE. California, USA. <https://doi.org/10.1109/acssc.2017.8335698>
- Yang, X., Hua, S., Shi, Y., Wang, H., Zhang, J., & Letaief, K.B. (2020). Sparse optimization for green edge AI inference. *Journal of Communications and Information Networks*, 5(1), 1-15.
- Yao, C., Liu, W., Tang, W., Guo, J., Hu, S., Lu, Y., & Jiang, W. (2021). Evaluating and analyzing the energy efficiency of CNN inference on high-performance GPU. *Concurrency and Computation: Practice and Experience*, 33(6), e6064. <https://doi.org/10.1002/cpe.6064>
- Yarally, T., Cruz, L., Feitosa, D., Sallou, J., & Van Deursen, A. (2023). Uncovering energy-efficient practices in deep learning training: preliminary steps towards green AI. In *2023 IEEE/ACM 2nd International Conference on AI Engineering—Software Engineering for AI* (pp. 25-36). IEEE. Melbourne, Australia.
- Yokoyama, A.M., Ferro, M., & Schulze, B. (2022). A multi-objective hyperparameter optimization for machine learning using genetic algorithms: a green AI centric approach. In *2022 Ibero-American Conference on Artificial Intelligence* (pp. 133-144). Springer. Cartagena de Indias, Colombia.
- Yosuf, B.A., Mohamed, S.H., Alenazi, M.M., El-Gorashi, T.E., & Elmirghani, J.M. (2021). Energy-efficient AI over a virtualized cloud fog network. In *2021 Proceedings of the 12th ACM International Conference on Future Energy Systems* (pp. 328-334). ACM. New York, USA. <https://doi.org/10.1145/3447555.3465378>
- Yu, J.R., Chen, C.H., Huang, T.W., Lu, J.J., Chung, C.R., Lin, T.W., Wu, M.H., Tseng, Y.J., & Wang, H.Y. (2022). Energy efficiency of inference algorithms for clinical laboratory data sets: green artificial intelligence study. *Journal of Medical Internet Research*, 24(1), e28036. <https://doi.org/10.2196/28036>
- Zhang, B., Davoodi, A., & Hu, Y.H. (2018). Exploring energy and accuracy tradeoff in structure simplification of trained deep neural networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(4), 836-848. <https://doi.org/10.1109/jetcas.2018.2833383>
- Zhang, X., Zhou, X., Lin, M., & Sun, J. (2017). Shufflenet: an extremely efficient convolutional neural network for mobile devices. *arXiv preprint arXiv:1707.01083*. <https://doi.org/10.48550/arXiv.1707.01083>
- Zhou, H., Alvarez, J.M., & Porikli, F. (2016). Less is more: towards compact cnns. In *2016 European Conference on Computer Vision* (pp. 662-677). Springer. Amsterdam, Netherlands. https://doi.org/10.1007/978-3-319-46493-0_40
- Zhou, X., Chen, Z., Jin, X., & Wang, W.Y. (2020). HULK: an energy efficiency benchmark platform for responsible natural language processing. *arXiv preprint arXiv:2002.05829*. <https://doi.org/10.48550/arXiv.2002.05829>