Criteria for Sustainable Use of Artificial Intelligence

University of Turku
Department of Computing
Master of Science Thesis
Data Analytics
May 2026
Lauri Kivimäki

UNIVERSITY OF TURKU
Department of Computing

Lauri Kivimäki: Criteria for Sustainable Use of Artificial Intelligence

Master of Science Thesis, 72 p.
Data Analytics
May 2026

The field of artificial intelligence is growing in popularity, but the impact it has on the environment has been neglected in research and in practice. The trend of deep learning models growing massively in size and computational complexity, demanding large amounts of energy and specialized infrastructure such as data centers, is not reflected in results; the performance gain is generally marginal. Moreover models are often constructed with unclear parameters or expensive methods, presenting barriers to research. This approach is known as Red AI. As such there is much room for incorporating sustainability demands and practices within the field of AI itself, known as Green AI, motivated by both efficiency and the environment.
This work aims to advance Green AI by investigating the current state of the research field via a Systematic Literature Review. The SLR addresses both the lifecycle and specific details related to building and using deep learning models more sustainably. Also discussed are tools and methods for estimating the lifecycle costs and thus environmental impact of models and the hardware they require.
Based on the contents of papers in the SLR there are industry-wide opportunities for improving the efficiency of deep learning throughout its lifecycle. Incorporating sustainability in this manner would allow for cheaper, more efficient models while lowering the barrier to participating in state-of-the-art research. However, Green AI is a complex topic with limitations and further work is needed on issues such as universally applicable metrics which consider sustainability and co-operation between AI research and industry. To help achieve this, a collection of criteria is presented in this work. The collection is based on implementing Green AI techniques to solve Red AI issues. The collection provides a novel listing of methods, principles and techniques for building more sustainable artificial intelligence and works as a starting point for those interested in the topic.

Keywords: Artificial intelligence, Deep Learning, Sustainability, Green AI

Contents

1 Introduction

Artificial intelligence or AI is a very active topic of discussion and investment. An estimated 154 billion US dollars was spent on investments in AI technology in 2023 [1]. Leading AI companies have announced total expenditure of up to $700 billion in 2026 alone. It is undoubted that many practices can be made more sustainable by applying AI. However, discussion on how to make AI itself more sustainable is much less prevalent.

Investigating the figures involved with AI research quickly shows that the field is completely unsustainable and current trends cannot continue forever. Between 2012-2019 the computing power required for state-of-the-art deep learning (DL) results has increased by 300,000x [2]. This computation has direct and indirect environmental costs which must be addressed as soon as possible.

The state-of-the-art is tunnel visioned on scaling up models with the hope of improving performance, but thinking critically this means that the field risks stagnation as financial and environmental costs become excessive and progress becomes marginal [3], [4]. This happens since any given approach will reach a saturation point with regards to performance increases, after which improvement requires super-linear growth in model size.

The concept of Red AI, which "buys" better results at the cost of using massive computational resources, was based on an analysis of 60 AI papers presented at the most prestigious conferences that showed that the vast majority of papers (between

75% and 90% depending on the conference) prioritized accuracy over efficiency. [5] "ResNet-101, which is less than 1% more accurate than ResNet-50, takes one week to train on four Maxwell M40 GPUs. Similarly, while ResNet-152 is 0.05% more accurate than ResNet-101, it takes 3.5 additional days of training time compared to ResNet-101." [6]

Large-Language Models, also known as LLMs, are the prime example of models with massive budgetary requirements and associated costs. Between OpenAIs GPT-2 and GPT-4, model size increased by 1000x. The energy consumed for their training grew by approximately 2000x. It's estimated that the total energy cost of training and running ChatGPT is equivalent to the yearly emissions of 175 000 Danish citizens. [7] It is a topic of some controversy whether the associated benefits outweigh the costs. In any case, following simple Green AI practices would result in environmental benefits with at worst negligible performance loss [4].

Green AI hopes to deliver smaller models using quality data with less need for computational resources and correspondingly lower environmental impact. This is possible because focus is shifted from having the largest and best-performing model to developing efficient and reduced models that are capable of performing the required task and nothing more. [5]

1.1 Research goal

The research goal of this work is to investigate the sustainability issues of Deep Learning and to find ways to begin addressing them in the spirit of Green AI.

As the topic has relatively little focus compared to the primary topics of interest in the DL field, part of the goal is to find out how much knowledge currently exists. As such the following three research questions have been formed.

RQ1 What is the current state of the field regarding the computational complexity, size and energy consumption of deep learning models and what are the trends

of these metrics?

RQ2 Does the literature present estimates and measuring methods for the environmental impacts, energy consumption, or manufacturing carbon footprint of deep learning training and usage based on factors such as model size or architecture?

RQ3 Based on currently available research is it possible to develop big picture calculation and estimation criteria for optimizing deep learning for sustainability?

The method of research is a Systematic Literature Review performed using the PRISMA guideline, consisting of recently published papers in the field addressing this topic. Based on the SLR a set of actionable criteria will be formed to provide a baseline for practitioners in the industry. The SLR is further elaborated on in the 3rd chapter of this work.

1.2 Declaration of AI use

No generative AI has been used by the author in the making of this thesis.

1.3 Structure of thesis

Chapter 2, "Background" summarises the basics of the field and elaborates on why and how the current state of things came about. Also included is background on Green AI and a lifecycle view of sustainability which form the theoretical backing for this work.

Chapter 3, "Systematic Literature Review" begins the SLR. This chapter goes through the primary contributions, findings, and arguments for Green AI in the SLR, divided into categories based on the topic. This chapter is the core of this work, and informs the reader on the possibilities and limitations of current Green AI research.

Chapter 4, "Problems and trends of Deep Learning" briefly outlines the

argument against Red AI and the challenges and problems associated with current practices. This chapter is essentially the answer to RQ1. This serves as both motivation for Green AI and a guideline for what problems need addressing. Chapter 4 describes how the current state of the field is unsustainable and why action must be taken to remedy things. The Red AI problems taken together with the Green AI methods from the 3rd chapter form a basis for actionable criteria presented in the next chapter.

Chapter 5, "Criteria for sustainable use of AI" presents the criteria, which are formed from the SLR, meant to address problems pointed out in the SLR. They are the primary contribution of this work, with the intent that anyone looking to practice Green AI will be able to benefit from implementing them or to motivate further research. These are meant to help both with research and industry practice.

2 Background

The field of artificial intelligence has existed since at least the early 20th century, but did not gain traction until relatively recently. This is mainly due to advances in the effectiveness of neural networks. Instead of focusing on achieving a human-level intelligence on a machine, it is more commonly used to describe machines improving their performance on a task without human oversight. The early inspiration for neural networks was the human brain and biological nervous systems. However they are not meant for simulating the brain and have notable differences in function.

The most relevant subset of artificial intelligence is machine learning (ML), which is concerned with developing statistical algorithms. ML models can be thought of as a (more or less complex) mathematical function built to learn to solve a specific problem. It is as simple as the input being processed into output even with complex models where no person understands the process. The most relevant ML model for this thesis is an artificial neural network (ANN).

Neural networks work by having one or more artificial neurons with connections to each other send signals. A neuron receives signal from other neurons before it, and sends signal to neurons after it in the network. What signal is outputted is computed by an activation function, which nowadays is typically nonlinear, which allows for solving complex problems. The strength of the signal is determined by a weight value for each connection, usually adjusted during training.

The object of study in this thesis is DL, a subfield of ML. Because of advance-

ments in DL models, neural networks are now leading the field of AI, and the largest LLMs are publically associated with the term AI. The neural network of a DL model consist of input and output layers, plus two or more hidden layers which are often sparsely connected. Hidden layers are located between the input and output layers, and they can't be accessed directly. This means specialized algorithms are required to update the model during training. While this means computational cost it also means that in theory a deep neural network (DNN) can learn to solve most problems since they can learn representations from complex data.

Neural networks learn from a dataset to solve whatever problem is at hand. Their ability to generalize to previously unseen data is validated on test data separate from training data. This dataset must be obtained in some way, and is usually processed in a pipeline to remove outliers or otherwise make it easier for the model to ingest. To improve performance models need to learn from more data, with diminishing returns. However, to execute training at a reasonable speed you then require a progressively larger model. [8]

Performance benchmarks on specific tasks are commonly used to compare performance between competing models. However, the older and more popular a benchmark is the less useful it can become. This is called benchmark saturation, as the performance of each line of state-of-the-art models reaches the top of a specific benchmark. This indicates that a task has effectively become too easy for it differentiate between model capabilities. Another problem is that performance on benchmarks does not directly translate to performance in practice [9].

The inherently opaque nature of DL means that humans cannot completely understand the workings of the models as they grow to tackle increasingly complex problems [9]. Larger models are increasingly hard to design effectively, and the network architecture is usually larger than necessary [10]. "DL is compute-intensive by design: The flexibility that lets it outperform expert models renders it expensive,

scaling faster than necessary“ [11].

DL model architectures allow for some layers that perform unique functions such as convolutional and attention layers, which are especially suited to specific problem domains. These include object recognition, natural language processing and generative AI. Generative AI methods using transformer models are one of the primary driving factors of present demand for data centers and computational power, with increasingly powerful hardware being developed purely for it [12].

The hyperparameter optimization governing the process of training a model is especially intensive computationally [13]. This optimization is important, because a poor set of hyperparameters can result in a completely useless model, but expensive because there is often little pre-existing knowledge regarding which set of hyperparameters will work best.

While the field has historically largely been viewed from a commercial and technical lens, the extreme financial and energy costs associated with high-performance computing and large datacenters needed to run large state-of-the-art models mean that sustainability perspectives have slowly started to gain ground. The question of sustainable artificial intelligence is part of a wider push toward more environmentally friendly research, which is already present in other fields. [12] In those other fields there have been previous European initiatives to provide criteria for sustainability, such as the Blue Angel [14], which partly inspired this work.

Life cycle assessesment or LCA is a standardized method for quantifying the environmental impact of a product, accounting for all costs throughout the lifecycle [7]. To improve the sustainability of AI, we must look at training, operation and embodied carbon footprint to account for the whole lifecycle, starting from data collection [8], [15]. This is the only way to avoid pitfalls such as optimising one part of the process in a way that makes the overall costs greater, and to keep in mind more obscure costs such as the emissions embodied in the hardware being used.

3 Systematic Literature Review

A systematic literature review was conducted based on the PRISMA guideline [16] to gather studies relevant to the research questions. Two search strings were used and the searches were performed on the Web of Science and Scopus databases. The same search strings were used on both platforms with modifications only for syntax:

SS1: ("neural network" OR "of deep learning") AND ("energy consumption" OR "energy efficiency" OR green OR sustainability OR "carbon footprint" OR "environment* impact")*

SS2: ("green artificial intelligence" OR "green ai") OR "environmental effect of ai" OR "ecological footprint" of ai OR "energy considerations" for deep learning*

Additionally, the search was restricted to publications from 2020 to early 2025 to focus on the current state of the research field.

Table 3.1: SLR exclusion criteria

Not in English
Not accessible
Not about deep learning
Article is not from the computer science field or related fields.

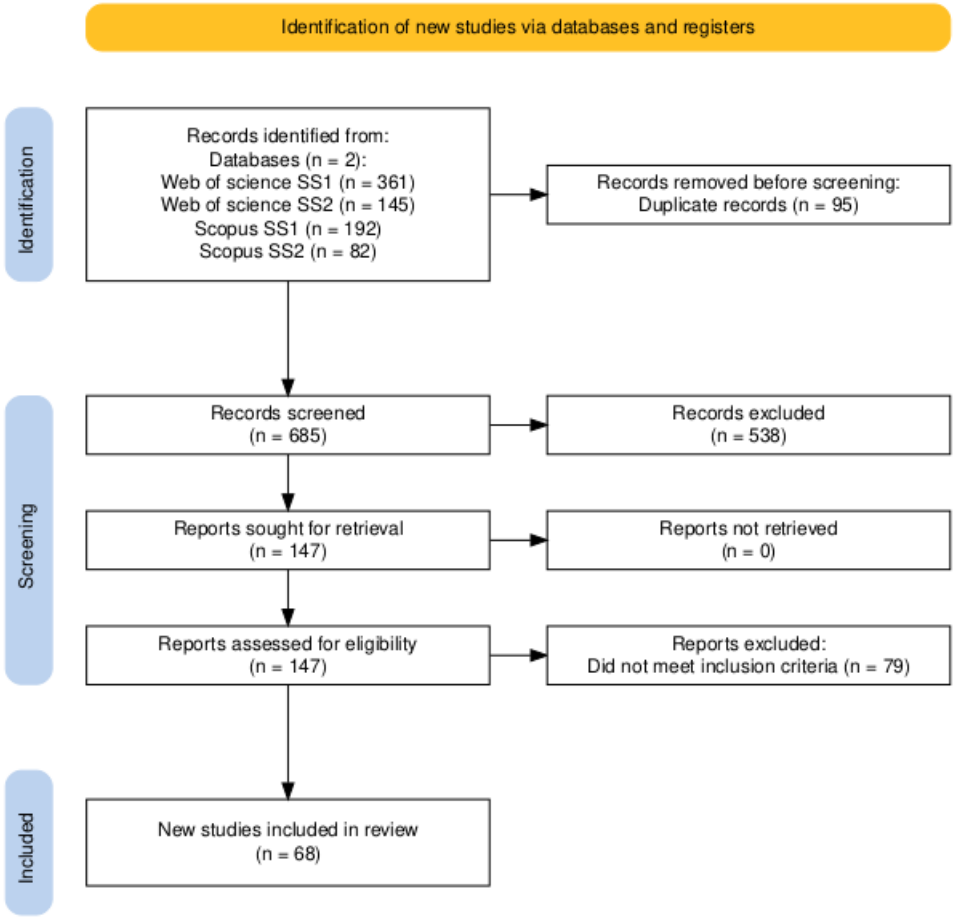
Table 3.2: SLR inclusion criteria

Title or abstract should be relevant to at least one research question
The publication addresses sustainability of deep learning in some manner

Whether a publication met these criteria was left to the sole judgment of the

author. This provides notable risk of bias when applying the criteria, but was unavoidable considering the solo nature of the work. The flow of the SLR is depicted in the following figure 3.1.

Figure 3.1: Flow diagram of the SLR



The SLR had the following results; Web of science 361 results for search string 1 and 145 results for search string 2. Scopus 192 results for search string 1 abd 82 results for search string 2. After removing 95 duplicates and applying the initial exclusion criteria the results amount for SS1 fell to 104 from Web of science and 15 from Scopus, with a notable number of false positive results. SS2 resulted in 79 results from Web of science and 44 from Scopus. Before screening, the remaining amount of publications was 147 articles (string 1 gave 56 and string 2 gave 91) "hits".

The remaining publications were retrieved to be examined for whether they meet

the inclusion criteria. After applying the criteria based on the full contents of each publication, the final count for papers included in the review is 68. The papers have been sorted into categories based on their perceived primary topic within the context of the SLR. However, most papers address multiple subjects on some level. Figure 3.2 visualizes how many papers are in each category and these categories will work as a basis for discussing the results of the SLR.

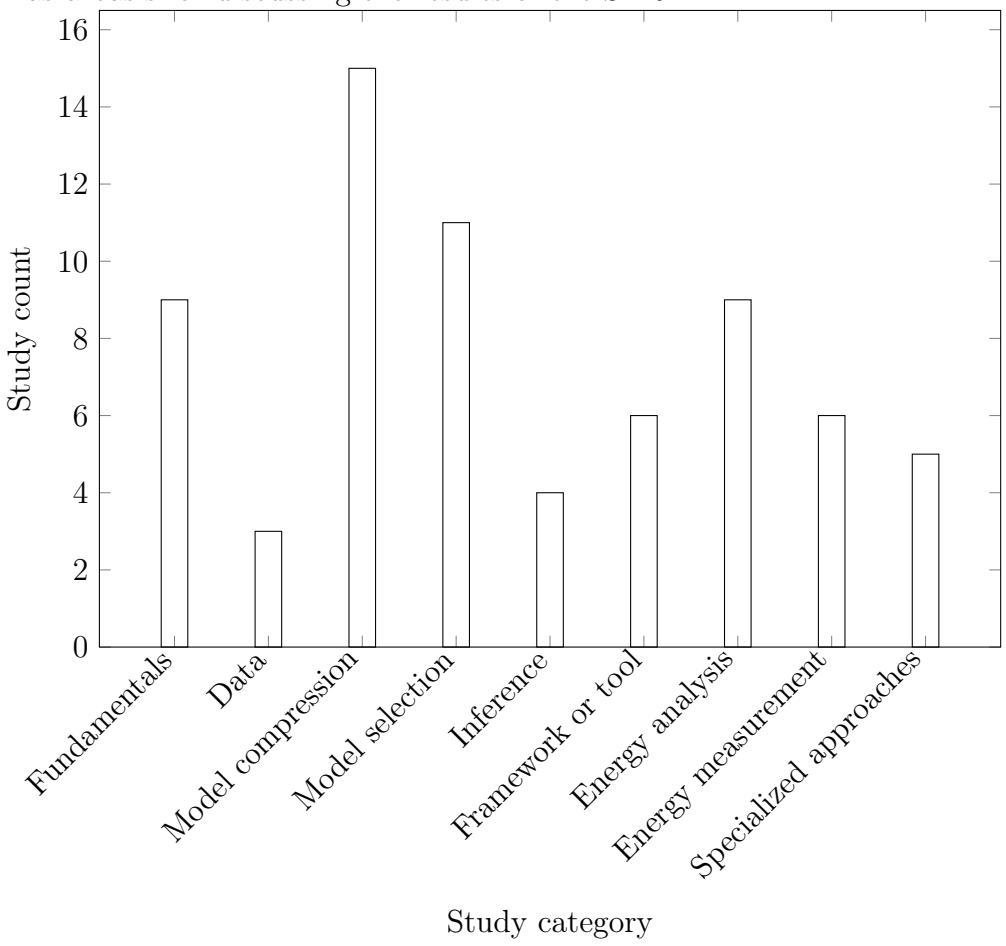


Figure 3.2: SLR paper categorization

A brief description of the categories and a list of their respective papers is presented in Table 3.3. Readers interested in further exploring a specific subtopic can use these references as a starting point.

Table 3.3: SLR category descriptions

Category	Description	References
Fundamentals	The DL field as viewed from a Green AI perspective	[1], [4], [5], [7], [12], [15], [17], [18], [19]
Data	The preparation and use of datasets	[2], [20], [21]
Model compression	Making models smaller while maintaining performance	[6], [10], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]
Model selection	The development and optimization of models	[3], [9], [35], [36], [37], [38], [39], [40], [41], [42], [43]
inference	The deployment and serving of models	[44], [45], [46], [47]
Framework or tool	Tools to help with evaluating and comparing models	[13], [48], [49], [50], [51], [52]
Energy analysis	Estimating energy costs and therefore carbon footprints of models	[8], [53], [54], [55], [56], [57], [58], [59], [60]
Energy measurement	Direct measurement of model metrics	[61], [62], [63], [64], [65], [66]
Specialized approaches	Ways to make models that differ from the traditional paradigm	[11], [67], [68], [69], [70]

3.1 Fundamentals of Green AI

Summary: *Deep learning models have been rapidly growing in size for many years, with much smaller gains in performance. This is in addition to inefficient or computationally expensive choices made during model development, such as exhaustive neural architecture search (NAS). Building and using the models also increasingly requires related resources such as new data centers and more training data on an unsustainable trajectory. Green AI as a paradigm is meant to address these prob-*

lems, and its essence is to provide novel results while considering the performance-efficiency tradeoff ignored by Red AI.

At present, with little regard for sustainability, deep learning state-of-the-art has reached a point where performance increases require exponentially greater increases in computational power. [12] These issues are expanded upon in the 4th chapter of this work. Green AI was born from the notion that focusing on optimizing efficiency and performance at the same time is more meaningful than chasing only performance increases.

Green AI as a field of research is characterized by still providing new results [17]. No matter the approach, Green AI means appreciating research effort that isn't reliant on massive datasets and training budgets [44] and it can be considered to encompass any method attempting to lessen the environmental impact of any point in an AI system's life cycle. One of the primary limitations of Green AI is that it is technologically demanding to perform. [7]

An important issue contributing to model growth is the grind to increase performance no matter the cost. Models constructed in the 2020s are still continuing to grow in size without corresponding benefits as there is always a saturation point for a given approach. According to Frey et. Al [3] "This "greedy optimization" with respect to a single metric (model accuracy), quickly leads to non-optimal, non-robust solutions".

Green AI posits that the computational costs required by state-of-the-art models are not necessary for progress in the field and that models can be developed to be much more efficient with minimal performance impact [12]. This is important as the proliferation of AI models for general use cases, such as self-driving cars and AI agents will further increase emissions and energy demand.

The energy consumption of machine learning in general is projected to be 30% of total energy consumption by 2030. [5] Moreover the computational cost and CO2e

emissions caused by AI models are not generally factored into the evaluation of the models by researchers, which is a hidden but massive environmental issue [69].

In some subfields and specific use-cases it is important that models require less computation. This includes the aforementioned self-driving cars, and more generally any type of embedded software that is meant to perform for long periods of time without supervision or with limited battery. These applications require high performance, low latency and good energy efficiency to be feasible. Larger models mean that the response time is longer and thus less examples can be processed. This greatly limits maximum performance for autonomous applications [13].

Cloud computing can be thought of as a way to address resource constraints, but can have privacy and latency issues [58]. There can also be issues with the cost and scale of data transfer prohibiting certain models from being feasible to use at all. From this point of view one benefit of the edge computing paradigm is that large-scale data transfer can be avoided and the data exists physically closer to the models.

There are also other problems with the data needed by models. Large models generally use massive amounts of data and acquiring it is becoming harder as the existing sources prove insufficient. The size of the training dataset is also a big part of the costs of developing a model due to the training process being repeated many times [8].

Approaches to Green AI can be primarily divided into being either model- or data-centric. Model-centric approaches make the model architecture smaller or more efficient, thus decreasing its energy consumption, running time and other related metrics. Compressing the model during or after training is one example of this. Data-centric approaches focus on either using the available data more efficiently (making learning faster) or selecting higher quality data to learn from while ignoring the bad. [2] Within the SLR model-centric approaches are clearly more popular but

in the big picture sense both aspects need to be optimized.

Boumendil et al. [58] discuss this classification more broadly. They describe that it is possible to focus efforts on optimizing the models algorithmically, by compressing them, by improving data use or by optimizing their infrastructure. Algorithmic approaches focus on manual or automatic design of more efficient network architectures. Compression describes methods of making the models smaller while preserving performance. Data optimization primarily refers to selecting only parts of the dataset. Finally, optimizing infrastructure focuses on the hardware and software used to run DL models. From this point of view special attention should be paid to data centers.

The massive computational costs and datasets of the largest DL models necessitate the use of AI data center infrastructure to house and run them. Data center optimization, which includes techniques such as dynamic load balancing, dynamic refrigeration adjustment, and optimization of resource allocation is relevant to most large scale AI-usage [5], [17].

In addition to the embodied carbon emissions within the infrastructure and hardware in a data center, the primary drivers of emissions are the efficiency of consuming energy and the carbon intensity of the location i.e. the mix of energy used. Cooling in data centers accounts for up to 40% of the total energy consumed [67]. In other words, large chunks of the energy being consumed do not directly contribute to the amount of work being done, which is massive overhead and should highlight the issue with Red AI.

For current DL purposes hyperscale data centers, also known as cloud data centers, are of the most interest. They inherently optimize resource consumption and are purpose built for large-scale computing. Their size allows them to be built for efficiency and to have higher power usage effectiveness (PUE) than on-premise data centers which often cannot implement the most efficient techniques due to scale

constraints [5]. They can also allow for flexibly performing computation in locations where renewable energy is currently available. However, lowering total energy consumption is still preferable.

In terms of optimizing model hardware, there are some general rules. For DL tasks a graphics processing unit (GPU) is more efficient than a central processing unit (CPU) due to their architectural differences. Then there are specialized hardware solutions for deep learning, such as tensor processing unit (TPU), optimized for low-precision calculations, performing them more efficiently than GPUs. [5] This approach is called hardware acceleration.

For hardware acceleration, processors do not exhibit uniform performance across manufacturers [61]. Knowledge of hardware benchmarks is needed to enable informed decision-making [7]. There are bottlenecks for specific GPUs and frameworks which mean that optimal load levels are not achieved during computation. A vast majority of model experimentation (over tens of thousands of training workflows) utilizes GPUs at only 30-50%, leaving room for utilization and efficiency improvements [15].

Load balancing should be performed to ensure hardware is used at the most efficient levels [3]. Batch size also greatly affects performance, especially with regard to memory limits. As a general rule TPUs outperform other processors. [61]

Finally, there are some considerations for the software models run on. The most used DL frameworks, Tensorflow and PyTorch, are not suitable for resource constrained devices. Smaller versions of them have been developed as a response. Some works have also proposed building custom engines for each model, but the difficulty of this task restricts it to only simple network architectures. This is a topic with relatively little attention, but avenues for further research. [58]

3.2 Data

Summary: *Data preparation can demand large amounts of energy, and high-quality data should ideally be shared with others working within the same field. Judging whether data is good or bad is difficult, but allows the model builder to leave out irrelevant data from the pipeline. Red AI ignores this concern, whereas the data-centric approach to Green AI is focused on finding quality data and processing it more efficiently.*

The dataset and model are the key products of deep learning.The relatively overlooked data preparation phase, consumes in some cases an equal amount of energy as the training phase [8]. Data preparation will always precede model construction, which makes it a relevant point in the pipeline from a green AI perspective [2].

A data-centric approach to green AI means performing modifications on or reducing the size of the dataset used while keeping the model the same [2]. Performing these modification tasks, such as reshaping data into more efficient structures, does involve minor overhead but it is also commonly performed for other purposes [20].

Data quality bounds the results of the model, with a big data approach amplifying any quality problems. Quality here refers to whether or not the data accurately represents the real-world characteristics of what we are looking for. The presence of low quality data should be investigated and either fixed or removed, as it causes more issues the further it gets into the pipeline. [2]

Typically the data used for any given model is not of equal quality, which suggests that a part of the data can be left out of the model with only minor accuracy issues at worst [2], [20], [21], [58]. Big data trends, as part of Red AI, ignore this and focus on collecting as much data as possible, but for quality models it is more important to be able to collect good data rather than lots of it [2]. Low quality data has a direct impact on the training time of the model, as it must churn through the irrelevant data as well.

All datasets and models show the same trend: More data in the training process results in greater emissions. Reducing the amount of data by removing data points can result in a good performance/efficiency tradeoff. The size reduction of the dataset can be performed at random or "smartly", meaning that info about quality is needed. The primary motivation for randomness is that it doesn't need the overhead computation that a smart selection does. [2], [21]

Boumendil et al. [21] studied the impact of data selection smartly versus at random, which necessitates finding ways to compensate for accuracy loss. They propose using a dynamic selectivity ratio that uses proportionally more of the dataset in an early epoch to maximize accuracy gains, then proportionally shrinking the ratio in later epochs to maximize energy gains. Using this dynamic ratio with smart data selection, they report results of 0.86% accuracy loss for up to 59.55% energy gain. Notably, this approach also improves convergence in comparison to a fixed ratio. Random selection is roughly competitive if using a dynamic ratio. Notably the effectiveness of random methods arises from the fact that the model will go through the whole dataset eventually, but doesn't necessarily have to do it in every epoch. However the selected subset of data must be updated frequently and the ratio must be dynamic for good results. [21], [58]

Anselmo and Vitali [2], introduced a method of handling model datasets more efficiently. It involves exploring the relationship between the amount of the dataset used and the resulting performance of the model, in other words mapping the performance/efficiency tradeoff into a curve. These curves form the basis of a regression model that can predict the required amount of the dataset for a given performance threshold.

By using half of the original data, in some cases the reduction in performance can be as low as 2% while the carbon emissions drop by more than 50%. Doing data reduction smartly to affect performance less, i.e. examining the data to find which

parts of it are low-quality and prioritizing the removal of those parts gives better performance results than reducing data at random [21]. Also, smart removal is at least as good or better than removing nothing. [2]

Of course, being able to judge whether or data is good requires some analysis. One issue with this type of analysis is that many use cases involve the gathering or use of previously unused data, and insight about well-known and curated datasets is not strictly transferable to data of unclear quality. This is because properties inherent to the data can have significant impact on how easy or hard training is. One example is how separable classes are from each other based on properties in the data. [2]

Data-related tasks such as accumulation, labeling, storage, processing, and exploitation consume large amounts of resources if they are performed separately for each project and never reused. Conversely, reuse of data between research papers is more sustainable and thus sharing high-quality data is advisable for Green AI.

As the first steps, this type of analysis could be performed on widely-used and presumably high-quality datasets to address sustainability concerns. The natural language processing (NLP)-field commonly uses pre-existing datasets for tasks and investing some resources into comparing them from a sustainability point-of-view could allow researchers to evaluate model costs even before development.

3.3 Model compression

Summary: *With sufficiently large deep learning models some degree of model compression can be performed without any loss in performance. This is because a subset of the network is responsible for the performance of the model with the rest being irrelevant. Identifying this subset also helps while training, as it allows for early stopping. No method of compression is generally superior, and combining methods will usually save more.*

Deep learning networks are generally overparametrized, which means that there

are too many internal variables governing the behavior of the model than strictly necessary. [24][10] The size is by design, to help with training. This inevitably creates a threshold value for networks where energy consumption increases by a lot to improve performance only slightly. In other words it is possible to compress the model during or after training to remove unimportant parts while maintaining as much performance as needed [58].

A portion of the network can always be removed without impacting performance, but it will eventually result in performance decrease. In other words, there exists a subset of the model that captures the relevant properties. This idea is generally known as the Lottery Ticket Hypothesis [18]. Finding this subset can be performed with approaches such as constructing connected components that show which neurons contribute to the network [33], allowing for the Pruning of bad neurons.

Using algorithms for the search, good quality tickets can be identified by the 20th epoch for a 160-epoch training, enabling use of the Lottery Ticket hypothesis in practice. 'Choosing' the ticket and pruning the rest of the network as soon as the ticket is found achieves energy gains of 8.6x for a pruning ratio of 50% and 10.2x for a pruning ratio of 70% while maintaining accuracies above 92%. Additionally, it has been shown that tickets can be transferred between a dataset and optimizer within the same domain. [58]

Barlaud and Guyard [42], based on the Lottery Ticket hypothesis, created a Lottery optimizer. It uses Lasso constraints to shrink the network (by setting other weights to zero and freezing them) to a subnetwork, which is then trained from scratch. This consistently results in models with less memory and compute demand with a small tradeoff in accuracy depending on the chosen constraint region. The ticket search allows for compressing the neural network to reduce energy consumption while keeping very good prediction performances. In many cases model compression can accelerate inference or reduce computational cost without any relevant accuracy

decrease [26]. For example, a case study was able to prune 5% of weights with the model maintaining the same accuracy as the unpruned model [27].

There is no compression technique that is generally superior to others [18]. However, applying almost any compression strategy will result in a notable reduction in energy consumption. Combining strategies that are not mutually exclusive will usually result in additional reduction [13]. For example, performing both quantization and pruning is generally helpful [26], [69]. This can also impact the model in unpredictable ways, meaning that double-checking is necessary. In some cases compression can serve as a form of regularization, owing to the overparameterized networks [6]. Half-precision floating-point weight quantization, representing the weights with 16 bits, is also notable for not degrading model performance [13].

Some optimization techniques based on neural architecture search and knowledge distillation consistently result in the largest reduction in energy consumption, and learning from a foundation model generally results in cheaper and more accurate models [58], [69]. This benefits inference especially, as the student model is smaller than the teacher as a rule. However, the costs of the teacher model should be included within the total cost evaluation as the student model would be useless by itself [64].

In fields such as NLP, large language models being distilled for use in more specific tasks instead of each project developing its own model from scratch would result in greener AI. This requires relatively open access to such models, and is one reason that transparency in deep learning is helpful.

Pruning weights in an unstructured fashion can lead to problems. Unstructured pruning also causes additional storage overhead that is greater than the reduction in weight storage. [29] It also does not provide meaningful savings without additional work to realize the potential, such as specific accelerators that can take advantage of the sparsity [58]. It is not competitive with quantization. However, structured

approaches are largely beneficial with the impact depending on the approach used [29] . Structured pruning also works best on large models, which are commonly used [58].

Quantizing the model during training and training with mixed precision does not cause accuracy loss, but the smaller model will usually take longer to converge [28]. However, energy-aware quantization also demands adjusting to the computing environment and model internals [23][69]. In other words, to realize the gains of quantization the hardware must be able to perform low precision computation effectively [58]. Quantizing globally is generally better than doing it separately for each layer, as it results in a higher compression rate and similar or better accuracy [10].

Yang et al. developed a solution to quantizing a DNN more completely, encompassing all computation steps and operands, which allows for all operations to be bit-wise. Using INT8 representation, the method presents massive memory, speed and energy savings with competitive performance. While suffering a notable accuracy loss of around 3% compared to full-size networks, most notable is the fact that this approach is much more efficient than other quantization methods using floating point data. [32]

Given that quantization is an irreversible process, it is crucial to balance achieving maximum energy efficiency with maintaining satisfactory learning performance. In edge intelligence contexts, though the reduced precision of model gradients extends the training time, quantized stochastic gradient descent (SGD) decreases the communication energy consumption by 5.7× compared to the standard SGD. Quantization should be performed with the lowest possible bit-width while retaining accuracy. Jointly training a quantizer and the DNN gives good results. [69]

Balderas et al. [22] present a methodology for compressing time series forecasting models called GreeNNTSF. The methodology is semi-automatic by virtue of em-

ploying a pruning algorithm that can automatically discover a subnetwork, fitting the Lottery Ticket Hypothesis. A limitation of the methodology is that it performs no hyperparameter selection for the purpose of simplicity, and in cases where the original network is poorly designed the process will not lead to good results. The methodology results in smaller and more accurate networks on state-of-the-art datasets. However it does not address energy metrics such as execution time which means it can have blind spots.

Balderas et al. also published a similar paper [30] addressing the convolutional neural network (CNN) architecture. The main novelty of their technique as compared to regular pruning is that it focuses on the filters of convolutional layers, the defining feature of CNNs. A three-step calculation ranks the filters of each layer by importance, leaving out everything below the k-th percentile from the optimized model. The effectiveness of the technique was tested on CIFAR-10 and CIFAR-100 datasets with multiple pretrained models. The paper reports best in class performance as a pruning method for the models it was tested on, in some cases slightly improving accuracy. On the Imagenet dataset it still stayed competitive but was less dominant. In any case the fact that more efficient models in the vein of Green AI can be achieved with minimal accuracy losses is once again demonstrated.

Ben Letaifa and Rouas [31], while studying the effects of compression on conformer (convolutional transformer) and transformer models, determined that while smaller models are always more sensitive to pruning, the robustness of transformers is generally better than conformers when it comes to model compression. Large transformer models show performance increase when compressed, but this doesn't apply to conformers. This is likely due to the discrepancy in the network caused by pruning multi-headed attention layers but leaving the convolution layers connected to them untouched. Conformers also experience an increase in errors after quantization.

Xu et al. [34] examined the impact of implementing Green AI techniques on the

YOLOv5 object detection network. They replaced the backbone of the model with lightweight networks while also pruning the model to remove less relevant features.They find that performance drops caused by pruning can be recovered significantly by fine-tuning afterwards, and that "the fine-tuned models exhibit a mere 1.77% drop in average accuracy, yet benefit from a 43% reduction in size, a 20% decrease in GFLOP, and an 18% average acceleration in CPU inference speed." They highlight the importance of determining the appropriate performance-efficiency tradeoff.

Applying a green learning paradigm to computer vision tasks, Wu et al. [43] demonstrate that heavily pruned models can offer competitive results. They construct a transformer model that does not use any convolution, intending to make it smaller and cheaper to run. The proposed model is smaller and faster than other benchmark models, while exhibiting good performance. The main difference of the model is the removal of inefficient modules while substituting better designed ones to perform their function.

Another interesting approach is to build the model in such a way as to minimize the amount of costly computational operations performed. Multiplication is the most important and most costly computational operation performed by DNNs. There have been proposed approaches to reduce the amount of multiplications by replacing them with shifts and additions of bits. However, this approach is much less popular and less studied than general compression techniques. [58]

3.4 Model selection

Summary: *At the size and scope of current model research and development, manually selecting model parameters is highly impractical. Thus it is important to use heuristics and automated processes for model selection. Major factors are stopping training as early as possible to minimize resources used on poor configurations and*

the use of metrics that minimize environmental impact while maximizing accuracy.

Manually configuring machine learning algorithms is usually both time-consuming and inefficient. Human intuition is unsuited towards solving problems with multiple dimensions, which is part of the appeal and power of machine learning methods. Although methods for automated construction of neural networks are being developed, human expertise is still very relevant in ideating and selecting models for any deep learning task.

Making informed choices when building models requires forethought and expertise from the human user to be effective because of the possible range of parameters. "For executing a given task and configuration, one is faced with a variety of options. Implementations can become arbitrarily complex." [52] For example some techniques are particularly sensitive to value choices when dealing with imbalanced datasets.

The need for humans is especially true for a shift towards green AI, as only the practitioners of AI can change the current state of the field. The use of tools to optimize for sustainability requires someone to use the tools. Additionally, the ability to accurately assess the impact of changes made requires expert knowledge at present [13].

Generally speaking the widespread use of automated machine-learning could provide a means for practitioners to construct models more efficiently, as they could simply share model parameters and architecture with each other instead of each researcher having to design and train their own model. An issue with the current use of automated machine learning is that there is no singular best-performing approach, so comparisons to state-of-the-art need to be done repeatedly with every repetition being computationally expensive [60].

There are ways to lessen the cost of this repetition. Layer freezing is a way of only changing some parts of the model during training. This is best applied to parts of the network that already perform well, which saves time and effort. Early stopping,

abandoning a training run before running it to completion if improvement stops, is also a proven method of saving energy while training, with minimal effect on accuracy. An auto-regressive algorithm can very accurately predict the accuracy and energy cost of the next epoch, which enables early abort of poor epochs. [36]

The problem space is massive even if selecting from a pre-existing model, and has to be performed within constraints of the problem being solved. Fine tuning each model to evaluate against the problem dataset wastes resources if it possible to instead select the correct models more efficiently. [50]

Solutions such as genetic algorithms can be used to effectively create and select ML models as a subset of automated machine learning [35]. Since a genetic algorithm mimics some useful properties of natural evolutionary processes it is one of the most effective ways of optimizing for multiple objectives. This makes them a natural fit for incorporating a performance-efficiency tradeoff into the process.

Another notable way of saving on the development costs of models is the use open source knowledge and transparency regarding methods and results. Instead of developing competing models to accomplish tasks within a field, one large open-source model could be used as a foundation model that others can train their models with. This presents large savings as training a large model is generally expensive. [1] The effectiveness of foundation models is generally better than training new models from scratch [25].

A paper that creates such a benchmark model allowing others to save resources could be more impactful than a paper showing a marginal increase in performance. These models also serve the function of lowering the barrier to entry into the field, and allowing more researchers without access to large models to study expensive problems. Development of foundation models should be coordinated so that the same work is not done multiple times. [60]

According to Del Rey et al. [54] "Selecting the proper model architecture and

training environment can reduce energy consumption dramatically (up to 80.72%) at the cost of negligible decreases in correctness" . This will naturally vary between use cases, but shows the importance of choices made. Manually configuring models will be less feasible the larger the model is, and with increasingly large networks being constructed it will likely become mandatory to let algorithms perform model selection [9]. Kannan et al. [50] provide the example of performing image object detection, where a researcher must select from over 1200 models while also having to configure the training process and evaluate the models. This type of method still requires expert input to pick out a suitable hypothesis space and model type.

Hyperparameter search is the most expensive part of training a model as it essentially involves constructing a model multiple times [13]. The results of this search can mean the difference between average and state of the art model, which means a lot of resources are put into it. Thus it is an important part to optimize, with especial care given to methods of stopping training early because the easiest way to save work is to do nothing at all. In contrast running a poor configuration of hyperparameters results in wasted resources with practically no knowledge gained. Stochastic methods can be extremely helpful here. Random selection has very low overhead but is commonly seen as ineffective. However, skipping random batches can be highly effective at escaping saddle points due to the added noise and can achieve high accuracy. [21]

Castellanos-Nieves and Garcia-Forte [39] compared using HPO-strategies with Red and Green AI objectives. The Green AI task is to search for hyperparameters that minimize CO2e while maximizing accuracy. The CO2e is estimated as $emissionintensity \times PUE \times totalenergy$. The Red AI only maximizes accuracy. The results indicate a reduction of 28.7% in CO2e emissions when implementing the Green AI strategy, compared to the Red AI strategy. This improvement in sustainability is achieved with a minimal decrease of 0.51% in validation accuracy.

Moreover, they find that accuracy and CO2 emissions are positively correlated, but the relationship between accuracy and the optimizer is not straightforward.
For hyperparameter search, there is also a data-centric approach (as opposed to improving the search strategy). For example, starting with fewer-dimensional data which is argued to give roughly similar results but is faster. [2]
In terms of optimizing optimizer functions for Green AI, the most important part to consider is the performance-efficiency tradeoff. This means an optimizer that provides almost as accurate results while leading the network to convergence in fewer iterations will generally be superior. In other words, an algorithm that trains a slightly worse network much faster should be considered better than the alternative. Foglia et al. [37] present an optimizer called HalpernSGD that increases convergence rate compared to traditional SGD, so as to decrease the carbon footprint of the process without compromising accuracy. While the new method outperforms SGD in terms of speed, Adam remains the superior optimizer in terms of performance. Future work on optimizing the effiency of Adam with similar measures is of interest. More generally, this reinforces the common trend in the SLR indicating that current methods can be improved in terms of sustainability for little loss.

3.5 Inference

Summary: *Inference will begin to dominate the lifetime energy cost a deployed model over time. There are many choices involved with making a model available to end users and inference, which greatly influence the associated costs. In terms of Green AI inference is somewhat unexplored as the research field tends to focus on training instead of using models, but it is gaining attention especially through LLMs.*
Inference is a major part of deep learning emissions, with the widespread adoption of generative AI tools based on LLMs. However, it has garnered relatively little attention in the literature. One of the contributing factors could be that the industry,

which has relatively little participation in Green AI currently, is more focused on the deployment and inference of models, whereas researchers typically focus on the training of models. As increasing amounts of DL's energy cost shifts from training to inference it is imperative that this aspect is investigated from the Green AI perspective.

50% of the operational carbon cost of ML tasks can be attributed to inference [44]. The impact of inference is so massive because of commonly adopted use of ChatGPT and other LLMs. A ChatGPT-like application handling 11 million requests per hour is estimated to emit 12,800 tons of CO2 annually, making inference 25 times more carbon-intensive than training GPT-3 [68] The computational cost of performing inference is dominated by computing the affine transformations (dot products between matrices). This means that models based on the attention mechanism are quite costly computationally. [24]

Generative AI and other such models deployed on a large scale are intuitively much better to train in a costly manner if it means inference is cheaper. Optimizing both at the same time is preferable, if possible. One method of doing this is to train a large model then using knowledge distillation to create a smaller model for deployment.

Another extremely important part of the lifecycle with little research on sustainability is the serving of models. i.e. what decisions are made when deploying a model for inference. Also, the efficiency of performing inference itself has been investigated, but not the overhead surrounding it. This refers to resources spent without direct benefit. Models are usually served within applications as a service. This is quite simple for the user but can result in inefficiencies. [46]

Samsi et al. [62], investigating the cost of LLM interference in terms of benchmarks found that increasing model size results in increasing energy costs per second not corresponding with the improvement in inference throughput. The relationship is not

linear. This means that small amounts of dynamic power capping, limiting the power fed to the models in real time, can result in useful performance/efficiency tradeoffs. This measurement also corroborates the fact that GPUs are mostly utilized at low percentages. Also, load balancing during inference is generally inadequate, perhaps due to bottlenecks in the frameworks or processors. In any case it leads to energy waste.

In a paper by Yarally et al. [44] the impact of batching inputs during inference is examined, especially the effect on energy consumption and response times. This means that the model is not given inputs until a specific batch size of them has been reached. As inference does not have a pre-existing dataset to divide into batches, this approach means introducing a delay to the process of any given query. The important part is factoring in the idle time that the method enforces as a tradeoff. Networks respond to inference batching differently, which means there is no universal solution or recommendation to be provided. For some networks, there are notable gains in efficiency which should be taken advantage of. In other words, this should be considered another optimisation parameter that can be tweaked while constructing the model. [44]

Another approach is decomposing the matrices of an LLM while attempting to preserve their information. Ben Noach and Goldberg [25] show that such an approach allows them to reduce model parameters by a factor of 0.4x, increasing inference speed by 1.45x with negligible accuracy losses. According to their results, decomposing a pre-trained model is better than randomly initializing one, demonstrating the benefits of sharing models.

When deployed into production, models tend to forget information from older data as they learn new data. Training from scratch each time is unviable. The computational footprint of retraining is a research gap. An especially important question to address is whether or not optimizing after each training session results in more

or less energy consumption in total due to the additional costs associated with the process. In any case training incrementally is better than retraining from scratch in terms of training time and computational footprint. [47]

3.6 Evaluation frameworks and tools

Summary: *Deep learning as a field is so large it is hard to generally evaluate and compare approaches. The most important techniques to understand are those that are widespread and architecturally agnostic. Overly specific research can quickly get outdated. To help with this problem there have been many proposed frameworks to help systematically evaluate models and their results. Efforts to provide evaluation tools have usually met with issues, however work is still ongoing. Multiple tools and frameworks are presented and discussed in this subchapter, with a focus on evaluating performance-efficiency in the Green AI sense.*

A general problem with the evaluation of DL is the vast amount of models used to solve problems, complicated by the fact that the techniques used to train them and the hardware platforms they're deployed on also differ greatly. For example, training a NN classifier is trivial, but inference can be costly. Computer vision models can be costly to train but cheaper to deploy [52]. Is the comparison meaningful in the first place, as the models are often used for different purposes?

This means that it is hard to evaluate DL, and even harder to know how accurate your evaluation is in practice. Even with the relevant metrics in hand, how do you use them to formulate a general evaluation of a model's efficiency? The combination of these factors makes any sort of guidance quite important in lowering the barrier to trying to achieve Green AI.

Efforts have been made to create standardized tools for the field. However, an issue is that these tools generally report different results, typically underestimating energy metrics in comparison to measurements [5]. Discrepancies in measurements

themselves can be caused by factors such as using indirect measurement methods or extrapolating imprecisely from hardware measurements [13]. However, absolute accuracy is not nearly as important as achieving high correlation between predicted and actual consumption figures so that they can be used to aid decision making

There are some Green AI tools that have been mentioned or even used in many papers throughout the SLR, such as Carbontracker and Codecarbon. Many of the tools are accessible as open source, which means the reader should examine them for further info if interested. This chapter gives an overview of them, starting with more general frameworks and ending with measurement tools.

Frameworks in this context refer to structured guidance provided to assist users with the objective of evaluating and developing Green AI models. Such frameworks are important steps in bridging the gaps in knowledge and lowering the threshold to participation in Green AI initiatives. Tools are executables that do not require subjective judgment or knowledge to use. Online tools are more versatile and add less load, however they may be less precise and require the user to obtain accurate information to provide to the tool.

EnergyNet [13] is a framework usable by running scripts, which runs through a pipeline that involves training a baseline model, then applying different optimization techniques to it. The models are then deployed and performance metrics profiled, leading to an evaluation and hopefully finding the best model. "Best" in this case is defined by an optimization profile selected by the user based on their needs.

In a use case described in the EnergyNet paper, the framework was able to optimize a model by reducing energy consumption by 82.304% with just a 0.08% degradation in balanced accuracy. This was achieved mainly with NAS and knowledge distillation. If the optimization profile did not permit for any loss in performance, quantization still provided a 58.208% reduction in energy consumption. [13].

Gaissalabel [49] is a web-based application aiming to provide a rough energy effi-

ciency label for models, which provides a relatively simple way of comparing metrics between models. Using it requires the user to provide the data in some way, such as inputting values or providing them in a file generated by another tool.

GreenRunner [50] is a tool that aids with component selection i.e. select a model, select evaluation metrics, select training process, then evaluate and compare results. The approach is to train a LLM to hypothetically be able to produce near-optimal combinations with less total cost than if manual searching was performed. Because an LLM is used the user does not need as much expertise and can prompt with plain text. Clear limitations arise from the use of a LLM, such as the response to the prompt being non-deterministic and the potential for hallucination or other issues. However, the approach has potential to be more beneficial on average than a manual selection process.

FECoM [48] is a framework for measuring the energy consumption of DL APIs. It is relevant to developers interested in Green AI. Greenlight [51] is based on FECoM, providing a dataset containing energy information of TensorFlow API calls. The main improvement over FECoM is interface and automation features making it more accessible. Codegreen, the tool used to manage the dataset, puts a wrapper around each API call that starts and ends energy measurement, with the aim of getting precise energy consumption measurements. Before measurement it performs checks for temperature and energy stability to reduce noise.

Fischer et al. published a framework and an associated tool [52] inspired by EU energy labeling for appliances. The framework establishes a formula for thinking about AI tasks. The formula describes the set of all possible configurations of a system C $= tasktypeT \times datasetD \times modelM$(including hyperparameters). Environment E is the tuple (A,S) consisting of computing architecture A defined by hardware component combination and software S implementing models and tasks. In practice the framework measures metrics during use, aggregates them from logs, projects them

onto an index scale, rates them by metric and then derives a total rating based on that. The user can control the weights used in calculating scores, so that the scales can be adapted to different use cases.

The paper notes that assessment results should be provided at different levels of technical knowledge. Color coded labels at beginner level. At intermediate level log summaries with values, scores and derived ratings, Full output and log files for experts who want to reproduce or otherwise analyze results. State of the art results couldn't be reproduced because unambiguous info on hyperparameters couldn't be found. This is another problem present in the research field greatly complicating both replication of results and their analysis from a Green AI perspective.

MetaQuRe [38] is a tool intended to automate model selection with a regression model. It is not possible to choose a best performing algorithm without taking into account the hardware and software it will be executed on. Moreover, resource efficiency should be included as a performance metric. Such a trade-off can lead to a situation where no singular algorithm is the best, but instead a set of algorithms providing trade-offs emerges. Instead of always performing the costly measurement of each metric per algorithm per problem, a proxy value is estimated from meta-features. These are learned by meta-learning models using an evaluation history of algorithms as training data.

Index scaling helps the meta-learner as opposed to training with real-valued measures. In the study it was decided to train individual learners for each algorithm, but it is theoretically possible to train a universal meta-learner as well. However, individual models provide more interpretability, which could allow for easier communication with less knowledgeable users. The paper proposes a methodological framework that selects algorithms based on priorities set by the user, somewhat similarly to Energynet and Greerunner.

The framework contains a dataset containing performance data from algorithms,

providing a basis for using the framework. The dataset index scales metrics to enable meaningful comparison. A limitation is that mostly simple algorithms are involved in the dataset, leaving the applicability of this method to large DL model sizes and architectures as a question mark.

Eco2AI [55] is an open source tool, designed to evaluate CO2e during training by measuring power draw and converting it through estimates. It works on Nvidia GPUs only and uses psutil for CPU measurement. The CO2e estimate is based on data and estimations for emission intensity, PUE etc. The tool doesn't do anything beyond the tracking, which can be a barrier to beginners who don't know what to do with the numbers.

In a 2023 paper by Bouza et al. [59] aiming to compare measurement tools, the same experiment was performed with 7 different tools. As each tool is meant to measure power consumption incurred while training DL models, the paper discusses them only in relation to that function. The paper highlights an issue with many of the tools making assumptions based on arbitrary values in case some information is unknown or not provided.

In the case of using TDP to calculate CPU power consumption, the tool Green-algorithms assumes 100% usage if the usage factor is not known, while the tool Codecarbon assumes 50%. Some tools consider power consumption of RAM based on the amount of available memory, while others look at the amount of memory allocated by the process. [59]. These may or may not be correct, but the fact that each tool takes different approaches and makes different assumptions is an issue in itself.

One aspect to consider with the tools provided is the issue of training costs vs inference costs. Given that models deployed in the industry have their lifetime energy consumption dominated by inference, it is important to have tools for evaluating that phase as well. Another issue highlighted is that many tools require extensive

knowledge of the execution environment or admin privileges to provide accurate results, yet approaches such as virtualization and cloud-based computation are quite popular and may heavily limit access to such resources. [59]

Given that is possible to extrapolate the energy consumption of a training phase from the values observed on only few epochs, we could measure the consumption of the firsts epochs, and then estimate the consumption of the total training. In this way, the consumption corresponding to the measurement will be slightly lower. [59]

3.7 Energy consumption and carbon footprint analysis

Summary: *A big picture approach is the most important thing to keep in mind for analyzing environmental impact. Analysis refers to calculation and estimation methods regarding these metrics and gathers the knowledge necessary before meaningful measurement can take place. For Green AI the important part is determining which parts of the lifecycle are safe to cut back on without incurring unforeseen consequences.*

Understanding the carbon footprint of DL training will play a paramount role in allowing people to develop more carbon-efficient models and hardware, making the emissions more transparent, and choosing renewable energy where possible. When optimizing for both energy consumption and performance it is important to focus on the big picture of the model.

A small improvement in one part of the lifecycle is not helpful if it results in issues in another part. For example, while research models generally do not see use and thus have little to no carbon cost for inference, the lifetime emissions for industry models generally become dominated by the cost of inference. These models can be deployed to a huge number of independent devices. Depending on the use case trading costly

training for fast inference can result in less total emissions.

In principle you can estimate the savings caused by deploying a model before executing the process. The number of parameters can be used as a reference efficiency metric, another possibility is the amount of floating point operations or FLOP required. These carry some issues but do correlate with energy consumption [27]. It is also extremely difficult to accurately estimate models independently of the hardware they will be running on, which often varies greatly. Evaluating models independently of the task they are performing is also practically impossible in a fair manner.

The popularity of AI also means that non-expert practitioners are training and deploying models with probably inefficient methods, meaning there is a need for generalizable techniques to make green AI [3]. When trying to apply DL to a new problem domain, especially in fields without a state-of-the-art or previous approaches, the first step is evaluating competing architectures. If the dataset changes along the way this comparison may need to be performed again. [7] All this illustrates why being able to evaluate the utility-to-cost ratio before starting a DL project would be very helpful.

Typically, throughout the lifetime of AI models, 50% of their carbon cost lies in the embodied carbon footprint of the hardware used to develop these models. However, the vast majority of training workflows under-utilizes GPUs at 30% –50% of their full capacity. [17]

A benchmark calculation for the carbon footprint of a qualitatively good model for autonomous driving was over 35 tonnes of CO2e caused by training a total of 4,789 models [41]. In fiscal terms, the electricity cost for the full development of an NLP model was estimated to be $9,870 [20]. Noteworthy is the fact that this does not represent the cost of the largest and most capable models.

T. Yarally et al. [40], after measuring the energy consumption of convolutional, linear and ReLU layers found that convolutional layers are the most expensive. S.S.

Acmali et al. [27] discovered that Bayesian optimisation is the most effective strategy for hyperparameter tuning. Findings such as these are important stepping stones on the path toward a well-understood green AI pipeline for deep learning.

Even while using renewable energy, doing any amount of work is not free. Using any computational resources will always cause wear and tear, so doing nothing will be greener than doing something [60]. In the cases where there will be a net benefit, something should be done. The challenge is being able to reliably estimate the costs and benefits beforehand.

There is a lack of generally accepted metrics for discussing and estimating the environmental impacts of deep learning. Many of the proposed metrics are also somewhat controversial due to issues with their accuracy with regard to energy consumption. A baseline for making estimates is usually the energy consumption or FLOP of the model, but it isn't strictly accurate [4]. This is because the metrics fail to capture the big picture, with FLOP in particular being hardware-agnostic.

One important notion is that static metrics such as PUE should be replaced with dynamic metrics that allow for real-time comparison of data centers [8]. Static metrics are generally poor as many factors can change over time, such as the percentage of renewable energy currently available to a data center. Focus should generally be on the amount of work performed.

Illustrating the complexity of this issue, Tornede et al. [60] state that "Running ML tasks is, in general, not deterministic under real-world conditions. The choice of the actual (deep learning) model, the compute center and the compute unit can influence the carbon footprint of a work by a factor of 1000, and that simply choosing the right compute center location can result in a factor of up to 10".

The more computation is shifted from the cloud to edge devices, the more important the energy consumption of the models themselves becomes. The average and peak power requirements are lower while utilizing the CPU compared to the GPU, which

can make it a more appealing in contexts with limited power availability. [66]

A Green AI SLR [17] found that most papers reporting resource savings typically attained at least 50% savings. If these savings could be implemented in a wider scale, the industry-wide savings would be massive.

Despite AI being a major source of energy consumption, equating its environmental impact to just energy is naive. The embodied emissions of the system can be a dominant factor for the carbon footprint of a system. Few investigate how other design decisions like selecting the training environment, development framework, and the like impact the environmental sustainability of DL [8]. Power and energy consumption estimation methods can be adapted to specific ML scenarios based on three key characteristics: target, dataset size, and training type (offline or online learning) [12].

Methods should be developed with the inherent consideration of their footprint. Ali et al. [57], analyzing the CO2 footprint of the training phase attest that the accuracy of the model is not in line with the caused emissions, and that the emission patterns cannot be generalized for each class of model. The problem of quantifying the environmental footprint of scientific research is extremely complex and may constitute a custom paper on its own.

According to Patterson et al. [53] the size (in terms of parameters) of a DNN is not as important as the activation rate. A sparsely activated DNN can consume less energy than a dense but smaller one while maintaining performance in terms of accuracy. However, a large but sparse DNN has greater system requirements and thus its embodied carbon footprint grows [17]. This highlights the pitfalls of trying to optimize a single part of the whole.

3.8 Energy consumption and carbon footprint measurement

Summary: *We must be able to measure something to be able to track and properly optimize it. However, the DL field is lacking in generally accepted and useful metrics to measure. There are two main obstacles. Firstly the field is somewhat new but very expansive meaning that most proposed metrics fail in some cases. Secondly there has not been enough attention regarding tracking the sustainability of DL models and their research, which manifests as a lack of available data to base research on.*

One of the most limiting factors of measurement is that no commonly accepted metric or measuring method is yet accepted within the field [58]. Especially difficult is that Green AI needs an universal metric to describe the performance-efficiency tradeoff applicable that applies to all approaches. [11] Discussion is still ongoing within the papers found in the SLR, with notable issues being some popular metrics such as the required FLOP. Another issue is that there is no systematic approach to measuring energy consumption, and no real benchmarks or datasets for evaluating energy costs.

Some studies might ignore factors causing noisy measurements or measure with a less accurate method [48]. Further, some works make claims of energy savings without measurement or verification of those claims [58]. The tools used should be accurate and the measurements repeatable. Commonly, both the metrics and methods vary from work to work.

A large portion of machine learning model experimentation only utilise their GPUs at 30–50%. This shows that an important step for Green AI engineering is to monitor and optimise GPU acceleration in pipelines [17]. There are thresholds of effectiveness for GPU load. A very high load results in additional strain and thus energy consumption, whereas for low usage merely the GPU being turned on

essentially wastes energy.

Among the issues being debated is whether to use a hardware power monitor or an energy profiler for the measurement. There is a strong correlation between the energy reports of wattmeters and profilers. Even though profilers are less accurate they are relatively effective at providing meaningful values for comparison [56]. They are also cheaper and easier to get started with. The power meter is intuitive, measurement software is easy but limited in application, and proxy metrics often present issues with accuracy [58].

According to Georgiou et al. [63] who measured the differences between DL frameworks, there is a statistically significant difference in their energy costs. Tensorflow is overall cheaper for training models, and PyTorch for inference. Thus it is in the interests of Green AI to use both depending on the context.

Alizadeh, N. and Castor, F. [45], performing similar measurements on runtime infrastructure, determined that inference time and energy consumption are generally strongly correlated, but performance and energy efficiency are difficult to predict exactly.

Tensorflow is slower in inference, due to suboptimal GPU utilization, reinforcing Georgiou et al. [63] The best performing framework changes based on priorities, meaning that developers should match their selection based on the objective. ONNX converted-models deliver better performance on all frameworks. For execution providers, CUDA is less efficient than TensorRT due to GPU utilization. [45]

This highlights the need both for holistic investigation of the DL processes and tools, but also the fact that for the research to be useful in practice the developers of AI models need to care about this information. At present "we observe that there is little involvement of the industry (23%) and that most studies revolve around laboratory experiments" [17]. In terms of Green AI, a metric that can correlate the

environmental impact of a model's lifecycle costs to its usefulness is also of interest [7].

There was a high precision measurement done by Caspart et al. [65] to improve awareness of energy efficiency in typical deep learning applications. They mention multiple obstacles toward measurement. Measuring large computer clusters, where much of DL work is performed, is infeasible. In contrast investigating the power draw of a single device and linearly scaling to estimate cluster performance is inaccurate as it neglects the surrounding environment connecting each device. Additionally, detailed information about power draw is typically only available with root level access. The larger a cluster is the less practical it is to measure with external power meters but using internal tools does not provide the full picture. Extrapolation from just runtime is inaccurate so measurement has to be done somehow.

One proposed way to measure energy consumption is by sampling the power consumption of the processor at training time repeatedly and taking the average over all processed samples. Excluding components such as system memory and storage as they are harder to take into account and account for a fraction of the total energy consumption. Additionally making the assumption that all devices and data centers are connected to the local grid at their physical location, as exact information is not publicly available. [67]

3.9 Specialized approaches to DL and Green AI

Summary: *Investigating edge cases and research that differs from the paradigm is especially important to Green AI, as novel findings and potential paradigm shifts will hopefully enable performance improvements with less cost for the environment. Approaches such as edge intelligence are especially helpful for Green AI as inherent resource constraints force innovation and efficiency at every part of the lifecycle.*

As the focus of this work is on DL generally, for specialized approaches the dicussion

will be placed on how they compare to traditional DL and what insights for Green AI can be garnered from them. A specialized approach here means either investigating a specific model architecture or differentiation from the paradigm of centralized learning with the models housed in AI data centers. Within the SLR the approach with the most focus is edge intelligence. Edge intelligence means deploying models physically closer to the user than with data centers.

The energy consumption of neural networks on edge devices has received attention even before Green AI due to the field's inherent hardware restrictions, especially due to the high resource costs associated with DL. There can be cases where the model should be deployed to a device with less computational power than the hardware it was trained on, with limited battery life. These batteries can be drained quickly by computationally intensive algorithms, and also enforce thermal limitations so the device won't overheat. In practice this is comparable to what Green AI as a research field is about as edge AI requires multi-objective optimization. Most important is just-enough intelligence, as opposed to maximal performance state-of-the-art approaches that are extremely expensive in comparison. [69]

Edge devices are approaching competive accuracy to the cloud, but consume much less energy. This is because hardware accelerators for edge devices have advanced enough that cloud computing is mostly unnecessary with proper otpimization. Red AI has neglected this opportunity which means even readily obtainable efficiency gains have not been realized. [11]

One of the most investigated recent advances in distributed learning for Green AI is Federated Learning or FL [7] . It is primarily a response to the privacy and environmental concerns caused by data centers. However, its environmental impact remains relatively unexplored [67]. In FL, multiple models are trained in a distributed fashion, then a central server aggregates them. The models must be aggregated as locally training causes the models to diverge over time. This can also

be used so that sensitive data is not uploaded to a data center. In any case, this process has inherent energy costs primarily due to communication overhead.

Each device in FL requires less power but the total is usually greater than centralized learning, however in the same order of magnitude. For lightweight workloads FL can reach target performance levels with little energy cost. Emissions can be roughly similar on smaller models. More realistic i.e. complex data makes FL take longer and cause more emissions. More complex optimization strategies can help here, especially if they require less communication rounds. Marginal carbon emission for additional accuracy gains is increasing exponentially. [67]

Hardware advancements make FL theoretically more efficient than centralized learning as it does not require cooling and has less overhead in that sense. However, the networking costs of transferring the models to and from the server to local devices is the major bottleneck [11]. This issue can be remedied by designing more intelligent data gathering and transferring techniques [69].

For federated edge learning, removing irrelevant and potentially adversarial data samples accelerates model convergence, reducing both the computation energy for local training and the communication energy for model exchanging. Second, using only the most informative data samples in local training reduces the computation overhead in each training iteration. Evaluating the quality of data samples is hard as the data is stored locally. [69]

A somewhat similar approach to FL is ensemble learning, where a pool of models combines their predictions to improve accuracy. The challenge is selecting which models to use on which problem to balance energy consumption with performance as using all the models all the time is strictly inefficient [68].

Implementing such selection strategies for models in production, Nijkamp et al. [68] were able to cut relative CPU usage massively while barely impacting performance. They enhanced standard selection methods with an energy-awareness metric to make

the selection process optimize for multiple objectives. The tool Datadog was used to extract the metrics used for this assessment, using CPU usage as a proxy metric, as energy consumption is not measured by the tool.

Asperti et al. [70] attempted to compare state of the art variational autoencoders or VAE:s to each other with emphasis on Green AI. These are a specific type of Generative AI model. The metric used in the study is FLOP which has issues as a metric. They are abstract and machine-independent which does not reflect what happens in practice. The study also finds that the relationship between execution time and FLOP seems random. Significantly the researchers were also unable to properly compare one the more advanced model architectures, as it required more computational facilities they didn't have access to.

4 Problems and trends of deep learning

This section higlights the problems with the current state of the field as presented in the SLR, providing arguments against it. Red AI has emerged as a term to describe the mentality of prioritizing performance improvements over all else [5]. This is characterized by constantly increasing model size which also requires more data and computation. While Red AI has resulted in much advancement in the field of research,such a myopic view must not continue in the long term.

90% of ACL papers, 80% of NeurIPS papers, and 75% of CVPR papers target improvements in accuracy [9]. With focus on improving only the performance metrics, Red AI research directly depends on exploitation of computing resources with marginal performance improvements [5]. In an attempt to further optimize the performance of deep learning networks, complexity and thus computational cost is often added to either the network or the training process [4].

A large part of the cost of AI is that much of the work is done with a brute-force approach in practice. This includes the data being gathered and the search for the best performing architecture and hyperparameters. This combines with the redundancy of the network to create the massive overhead costs that state-of-the-art models exhibit. [19]

The largest DL models are so large they are practically infeasible to run with-

out hardware acceleration and associated infrastructure [9]. Such hardware can be energy-intensive. The high computational cost also leads to massive energy use, and a corresponding embodied carbon footprint through the equipment used for AI. Excessive resources are being used on practically every part of the lifecycle. As the model architectures approach their performance limits, the energy consumption required to achieve further performance improvement increases super-linearly [69]. Using expensive methods such as Neural Architecture Search to train a Transformer has roughly equivalent CO2 emissions to the lifetime emissions of 5 average cars in the United States [25]. Another study reports the same scenario as emitting about 57 times the CO2-equivalent of the average human per year. These figures do not provide a meaningful point of comparison without additional analysis but serve to illustrate the need for Green AI. The size increase of models is also one of the primary drivers causing the amount and size of data centers to increase. The energy consumption of data centers has been forecasted to possibly reach 1,000 TWh in 2026, which is comparable to the energy consumption of Japan [7].

Training GPT-3 resulted in the emission of 502 tons of CO2, nearly 100 times what an average human emits in a year. In terms of energy it required 1287 MWh, equivalent to 121 homes in the US. Since the subsequent version, GPT-4, was trained on 570 times more parameters than GPT-3, it must have required even more energy. The environmental cost is not restricted to training, as using these systems also has a cost. As an example, GPT-3 was accessed 590 million times in January 2023, leading to energy consumption equivalent to that of 175,000 persons. ChatGPT queries using GPT-4 consume 260.42 MWh per day. [5]

When deployed, larger models consume more energy. Google and Microsoft may receive exponentially more questions each day as chatbots and picture generators gain popularity and as they integrate AI language models into their search engines. NVIDIA's new AI servers are expected to consume more energy than Argentina

and Sweden by 2027, with an annual consumption of about 85.4 TW-hours [1]. Deploying models effectively can be a real issue for Red AI as energy efficiency is an important factor for real world applicability [3].

DL model training is preceded by data preparation which impacts the rest of the task and is part of every task. The most power hungry phase of model construction is hyperparameter search as it consists of training multiple models. "Previous studies have suggested this to increase energy consumption by a factor of roughly 2000×: Strubell et al. show that, while training one of their natural language processing models has an electricity cost of $5, the electricity cost of performing the full R&D required to develop that model is estimated to be $9,870." [20]

Understanding the carbon footprint of ML training will play a paramount role in allowing people to develop more carbon-efficient models and hardware, making the emissions more transparent, and choosing renewable energy where possible. However, using renewable energy is not an excuse to waste energy. If all the renewable energy is used by one application then someone else will use the energy coming from fossil fuels [7].

4.1 Trends leading to red AI

As previously stated, from 2012-2019 the computational cost for training the largest models grew over 300,000x [2][67]. This is roughly equivalent to a 3.4 month doubling period for the cost of training. Yarally et al. [44], while comparing the progress of a series of models published from 2014-2022 on the same dataset, found that accuracy increased by 35 percentage points while energy consumption increased by 135 percentage points. Meanwhile, the size of the models increased by over 30x, from 60 million to 2.1 billion network parameters [69]. As another example, recommendation model sizes have increased by 20× between 2019 and 2021 at Meta [15].

Figure 4.1 works as an illustration of these growth trends, displaying the 3.4 month doubling period [12]. However, in reality these values did not rise continuously and instead discrete jumps happened as new models were developed and published. The rise in computational cost completely dwarfs all other metrics.

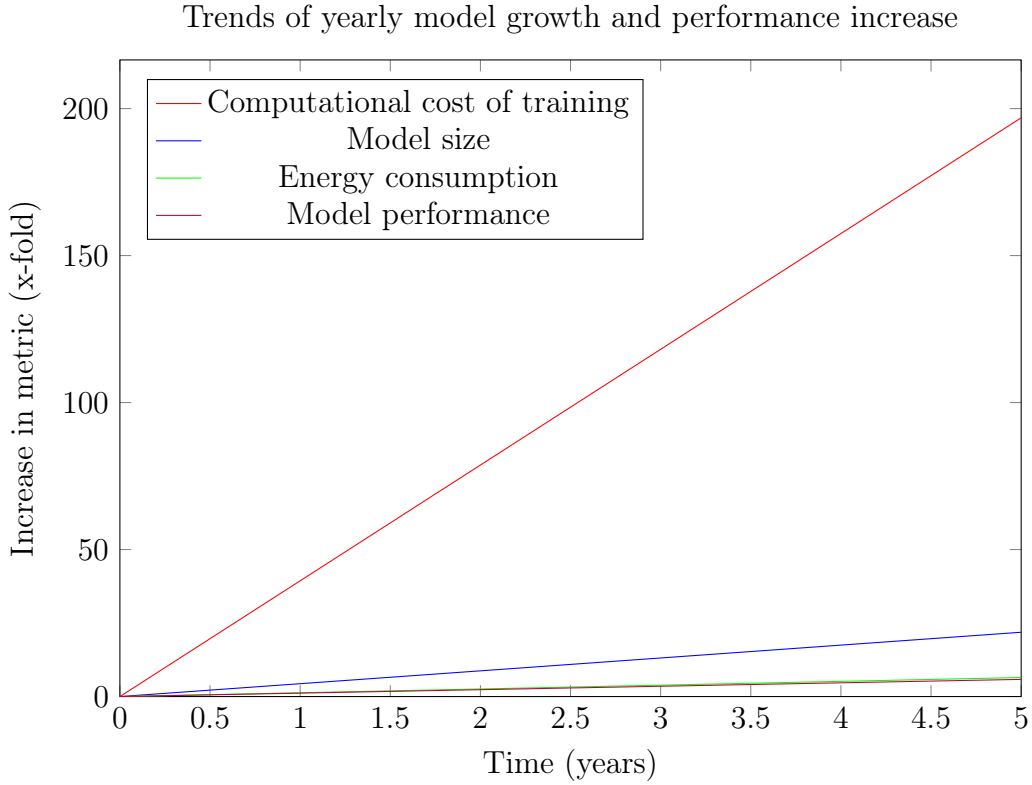


Figure 4.1: Red AI growth trend

General interest in the field of deep learning has contributed to the extreme pace at which the field is trying to advance, partly due to the amount of money it now has access to in comparison to the past. This pace inherently means that much work is done unnecessarily and without consideration, which compounds with the fact that state of the art models are exceedingly large and laborious to create. There is still much that is not understood about the inner workings of deep learning networks and how to construct them well.

There is a disconnect with AI research and the industry. Verdecchia et al. [17] found that during and after 2020 a spike in Green AI publications happened. However, the industry still has little involvement in the field. In addition, as researchers usually

focus on the training of models and the industry on inference their priorities will not always match. As long as Green AI is not a primary concern for the industry, even simple and easily accessible performance gains will not be realized.

Larger models by nature require more data to learn from. However, with currently available datasets and methods it is a fact that the amount of quality data available is nowhere near sufficient for the most complex models. This fact necessitates the use of methods such as reusing the same training data, using low quality data, or manufacturing new data to train increasingly large models.

Wu et al. [15] discussed the data trends for models at Meta: “The amount of data for recommendation use cases has roughly doubled, leading to 3.2 times increase in data ingestion bandwidth demand 2019-2021”. For some tasks, the energy cost of data transfer is higher than that of training. This data volume increase leads to a super-linear model size growth trend, where a search engine model growing 1000x larger improves the performance metric AUC by 0.030.

One issue of using low-quality or unvetted data is that with general statistical models, you would only want the model to learn from the good data which actually shows it info about the distribution of interest. All other data going through the pipeline is essentially wasted computation as it at best does nothing for the model and at worst makes it perform worse via overfitting. This phenomenon has been often summarized with the phrase "You put trash in, you get trash out". [2]

Additionally, with the vast amount of models available and being trained it is hard to select the correct model for the task, or to know if one is needed at all [50]. This can often result in practitioners choosing inefficient models, which is directly in opposition to the performance/efficiency tradeoff of green AI.

4.2 Performance issues

The main consequence of the usage of Red AI is based on the diminishing returns after increased computational cost over time. Schwartz et al. [12], who popularized the concept of Red AI, highlighted that "The relationship between model performance and model complexity (measured as number of parameters or inference time) has long been understood to be at best logarithmic; for a linear gain in performance, an exponentially larger model is required" a pace that far exceeds Moore's Law, which states that the amount of transistors in a microchip doubles every two years. Thus, the increase in demand happens 7x more often than the increase in supply, and will inevitably result in a problem as the hardware simply cannot run the software in a reasonable amount of time at some point.

Keeping in mind also the fact that deep learning networks are overparametrized by nature, an excessive amount of useless computation must be performed to get marginal gains with current methods. Major advances in the performance of state-of-the-art models are usually often only found by completely new approaches and paradigm shifts [2].

Some incidental problems caused by this trend of "needing more" in every aspect of the lifecycle are the hardware and energy required to operate the systems running the models. Larger models generally cost more energy when deployed. This is especially relevant to LLMs as a hot topic of research and use. NVIDIA's new AI servers are expected to consume more energy than Argentina and Sweden by 2027, with an annual consumption of about 85.4 TW-hours. [1] The scale of data center use and construction is rivaled only by the AI boom that necessitates it.

4.3 Research barriers

With the concerning trend of state of the art results in the field being siloed in corporations with large amounts of computational resources and capital, it means that research in the field faces multiple problems. The economic cost of AI means that key developments are in the hands of actors with the vast resources required. GPT-2 trained with 1.5 billion of parameters in 2019 cost an estimated 50,000 USD, while PaLM trained with 540 billion parameters in 2022 cost 8 million USD; thus, PaLM was 360 times bigger and 160 times more expensive. [5]

Specialized models can have even more extreme costs, such as AlphaGo, the best version of which required 1,920 CPUs and 280 GPUs to play a single game of Go, with an estimated cost to reproduce this experiment of $35,000,000. [12]

One issue arising from this is the lack of reproducibility with respect to the results being reported. As mentioned before a study found SOTA performances impossible to reproduce, because unambiguous information on hyperparameters is hard to find [52]. In addition to other problems Red AI also makes contributing to research harder, as the barrier to entry is much higher [2], [53].

It is impossible to reproduce the best BERT-large results or XLNet results using a single GPU, and models such as openGPT2 are too large to be used in production. The research field often changes quickly enough that overly specific research will be outdated. The use of massive data creates barriers for many researchers to reproducing the results of these models, and to training their own models on the same setup.

With the models and their computation being hosted on private data centers that do not share their information, it also becomes extremely hard to accurately measure the models and their statistics. In addition to this making any singular model impervious to accurate measurement, it also means large scale analysis and comparison is ineffectual, as it can't be backed up by more than estimates. Model results

should be made more transparent with regard to their carbon footprint, reflecting energy and water demands and allowing comparisons to see if AI is actually helping in solving issues [5]. The lack of standardized reporting processes and tools can also result in research bias, even if attempts to record metrics are made [4].

5 Criteria for sustainable use of AI

The collection of criteria presented in this chapter has been formed based on the author's judgment of the contents of the SLR. It is intended to provide a starting point for anyone interested in the implementation of Green AI, and also to highlight points of further research as some things are not possible to do without gathering more data.

When forming the criteria focus was put on principles and techniques that were actionable and should have immediate results. However some benefits of the most important criteria can only be realized with industry-wide involvement in Green AI. Of particular note in this regard is a call to action toward transparency of research methods and results in the AI field, to enable the comparison and replication of results.

As shown in the results of many papers in the SLR, sustainability gains can be realized with just a shift in mindset and priorities, or in other words taking into account the efficiency perspective. Even better would be the widespread adoption of "just enough" as a paradigm, which involves using AI models that can reach the required performance threshold for a problem and nothing more.

Badar et al. [4] state: "Simple standardized practices can lead to significant reduction in environmental impact... We also suggest the research community to take a step back and focus on the bigger picture when designing the networks, instead of targeting minute improvements with narrow applicability at the cost of simplicity,

versatility, and energy"

The criteria are presented as a list without strict categorization as environmental concerns necessitate a holistic approach instead of focusing on singular parts of the process. They are roughly organized in the same order that the SLR was discussed in. A particular criterion may be more or less abstract, and the difficulty of implementing or considering a particular criterion may vary. This aspect is discussed in the descriptive text provided for each criterion. The references provided as justification for each criterion either argue convincingly for such an approach or provide evidence of it working in practice.

The detailed descriptions of the criteria can be found in a table format. The following criteria are presented: Fit for Purpose (Table 5.1), Performance-efficiency tradeoff (Table 5.2), Generally optimal solutions (Table 5.3), Avoid reliance on intuition (Table 5.4), Prioritize efficiency instead of exhaustiveness (Table 5.5), Effective use of data (Table 5.6), Foundation models (Table 5.7), Multi-objective model selection (5.8), Lottery Ticket search (5.9), Early stopping (Table 5.10), Model compression (Table 5.11), Optimizing for inference (Table 5.12), Measure metrics of models (Table 5.13), Report metrics of models (Table 5.14), and Hardware acceleration (Table 5.15)

Table 5.1: Fit for Purpose

Name	**Fit for Purpose**
Description	Fitness in this context can be summarised as "is the tool we are going to build the best tool for the job" considering all possible factors, available resources etc. In other words the objective should be to find the cheapest method that is good enough for the task, instead of finding the absolute best performing method regardless of cost. Evaluate the project and what is known about it including how likely is it that developing a DL model will help. This can lead to a mountain of questions which are all relevant to attempt to answer before doing any work. Questions like these should be considered before starting to implement a DL model: How much will developing it cost, do you have access to the required data already or do you do need to obtain it from somewhere. Has someone already made a model performing a similar function you could access? What type of model would best fit, what parameters does it need to meet to be useful, what is the acceptable performance/efficiency tradeoff. Do you have the infrastructure to run the model or do you need to obtain it from somewhere? Where and how will the model be deployed? How complex of a model do you need to perform the task you want it to perform? Is a linear fit good enough? How long will the model be in use. Will it need to be retrained. Will there be additional projects that could use it as a foundational model
Justification	The easiest way to save energy and resources is to do nothing at all, which is why it's important to consider if creating a new DL model is the most effective way to achieve a given project's goals. Just because you have access to high performance computing does not mean you need it for every task. [1], [59], [60]
Relation to lifecycle	Project and lifecycle, parent criterion for most criteria related to the model itself
Limitations	Simple to understand but hard to do well. This criterion is somewhat open-ended, and there is no real standard or benchmark for it as of yet. However, any work done to improve fitness will be helpful.

Table 5.2: Performance-efficiency tradeoff

Name	**Performance-efficiency tradeoff**
Description	This tradeoff is the foundation of Green AI, and not considering it leads to Red AI. Ignoring the efficiency of the model can lead to extremely expensive configurations with almost irrelevant increases in performance, whereas neglecting performance can lead to very small and fast models that do nothing at all. This is why establishing a threshold for the acceptable tradeoff between these two values is important. Methods such as AutoML should be used to optimize for multiple objectives while selecting the model for a given task.
Justification	As depicted in many of the case study articles in the SLR, it is generally possible to achieve cheaper and smaller models with little to no impact on accuracy. The efficiency gains will be even greater if performance can be lowered. This means that it is better to get good enough performance cheaply rather than the best performance expensively . [12], [18], [26], [30], [34], [35], [36], [39], [40], [52], [56], [64], [68]
Relation to lifecycle	Encompasses every consideration that deals with the design, construction and performance of the model itself.
Limitations	Simple to implement, however no third-party way of validating that the optimal value is chosen. Choosing a good threshold for the tradeoff can be difficult depending on the goals of the project.

Table 5.3: Holistic optimization

Name	**Focus on finding generally optimal solutions, with improvement efforts focused on efficiency instead of performance**
Description	Optimize holistically (need metrics and analysis methods), so the model performs as well as possible in all metrics including sustainability instead of optimizing a single metric such as accuracy "greedily". With the proliferation of DL models into many aspects of life, it is prudent to make the models robust and generally optimal so that they perform as well as possible in practice. For example embedded or edge models will have issues with performance if they are not optimized for efficiency. More generally, as a motivator, major performance gains are usually realized only with paradigm shifts or notable advances in the field and as such pouring 1000x computational resources to realize a marginal gain in accuracy is simply not "worth it".
Justification	[4], [5], [7], [8], [12], [15], [18], [23], [26], [39], [40], [43], [52], [69]
Relation to lifecycle	Model lifecycle, performance-efficiency tradeoff
Limitations	Understanding what is and is not more optimal can require expertise based on the problem domain.

Table 5.4: Avoid intuition

Name	**Avoid reliance on human intuition**
Description	Including manual configuration of models, the larger the models are the harder it is for humans to be able to make informed decisions about model parameters. This process should be automated with techniques such as automated machine learning, which is much more efficiently able to optimize multi-dimensional problems. More generally, decisions should be made based on statistical analysis instead of trying to intuit correct solutions. Human expertise is still generally required to make informed decisions in the model construction process but should be secondary and based on data or expertise instead of at random.
Justification	[9], [13], [32], [35], [38], [39], [50], [60]
Relation to lifecycle	Model selection, training, compression
Limitations	Can result in problems if selecting algorithms or methods that are unsuitable for the task, as can happen if no human expert is involved whatsoever

Table 5.5: Efficient instead of exhaustive

Name	**Prioritize efficient instead of exhaustive**
Description	When building a model, instead of trying to run an exhaustive NAS, it is more efficient to use stochastic techniques such as Bayesian optimization which leads to good results while doing much less work since randomization is inherently fast. Large models with high performance can also be made by making network activations more sparse, making the model more efficient and less costly to run in practice while maintaining performance. Throughout the construction of the model it is possible to make these adjustments to do more with less.
Justification	[3], [4], [5], [12], [15], [21], [34], [37], [40], [41], [44], [53]
Relation to lifecycle	Model selection, architecture
Limitations	Being overly efficient while making decisions can result in missed opportunities, which means that decisions should be made based on statistical analysis of data and only while well-informed. Stochastic methods are an example of both a danger and an opportunity

Table 5.6: Effective use of data

Name	**Effective use of data**
Description	Data should be collected and selected with attention instead of hoping that "more data will be enough". This also includes not blindly gathering more data as a response to problems, instead one should attempt to coordinate with others in the problem domain to obtain good data. Poor quality data that does not "teach" the model anything should ideally be left out of the pipeline. For quality models good data is more important than lots of it. Data selection before and during training can result in massive savings without performance loss, as demonstrated in the SLR. Additionally, the dataset can be modified into a more efficient data structure or made smaller with minor overhead. The model can be trained on data with reduced dimensions as long as the resulting level of performance is good enough.
Justification	[2], [7], [18], [19], [20], [21], [54], [56]
Relation to lifecycle	Datasets and training
Limitations	Analyzing the quality of data requires work and knowledge from curated datasets does not directly transfer to use of previously unused data. Doing modifications on the dataset without knowing what you're doing can ultimately cause severe issues to the model.

Table 5.7: Foundation models

Name	**Foundation models**
Description	Knowledge distillation conveys additional knowledge that improves the performance of the model being trained. After a model such as an LLM is trained to perform at a high level, it is possible to train much smaller models to copy it with similar performance but at a much lower model size. These smaller models can be provided as more accessible open source versions of the larger model with similar performance. This procedure is usually effective if the foundation model is a large DNN, meaning that it is suited especially well to LLMs/generative models and more generally the growth trends of the AI field. Additionally, transfer learning is very effective on high-performance models which results in cheaper training for problem domains with limited training data.
Justification	Foundation models are generally proven to be effective both as a way to efficiently train high-performance models on new problem domains and a way to make smaller models for inference. [13], [64]
Relation to lifecycle	Model selection and training
Limitations	Requires either access to a foundation model or the ability to train such a model, which can be prohibitively costly if only needed for a single project. As such, this criterion cannot be universally applied.

Table 5.8: Efficient HPO

Name	**Multi-objective model selection**
Description	Hyperparameter selection and optimization is performed almost always with large enough DNNs, but the efficiency of the process is not always a priority. It is much more cost effective to for example use an optimizer that causes the network to converge earlier with a slight loss in performance in comparison to using an optimizer that takes forever to create the best network. It is naturally also better to select the faster optimizer with equal performance to a slower one, but this might not be possible if efficiency is not considered at all. Manually creating a large model is infeasible and the process should be automated, thus the automation should be made to pay attention to efficiency as well which means searching for parameters that maximize accuracy while minimizing CO2e. Techniques such as genetic algorithms are effective for this purpose.
Justification	[3], [7], [9], [11], [13], [29], [35], [38], [39]
Relation to lifecycle	Model selection, optimizer, early stopping
Limitations	Depending on method, may get stuck at saddle points, demand excessive computational resources or result in poor models altogether.

Table 5.9: Lottery Ticket search

Name	**Lottery Ticket search**
Description	When training a network, it is possible to identify the subnetwork that is responsible for the majority of the performance relatively fast. As everything outside of this subnetwork is essentially redundant to the model it should be eliminated to save resources. Some part of the model can be removed without impacting performance at all, then it will gradually start degrading performance. Based on the pruning ratio, the performance-efficiency tradeoff incurred can be adjusted.
Justification	The method has been demonstrated to be both effective and generally applicable [42], [62], [67], [69]
Relation to lifecycle	Model selection, training, model compression
Limitations	Care should be taken that no more pruning is done than necessary, and that the Ticket has actually been found.

Table 5.10: Early stopping

Name	**Early stopping**
Description	As little work should be performed as possible, so as soon as it becomes evident that a batch will not give good results training should be stopped for that batch. This technique can be extended to account for energy costs, in other words if we predict that continued training will result in small performance improvement for massive energy cost, we should stop. Skipping random batches can also result in high accuracy as the added noise can give better results than more structured optimizers. Essentially the challenge then becomes predicting the performance of each batch, which can be done relatively effectively and early stopping generally has little impact on accuracy.
Justification	HPO is the costliest part of training, and should be optimized as much as possible. Moreover, running a batch with a poor configuration of hyperparameters to completion is a waste of resources that will generate no value whatsoever. [3], [7], [36], [55], [60]
Relation to lifecycle	HPO, model selection, training
Limitations	As stopping is based on prediction there is a chance that stopping techniques will result in terminating good runs, however slight. This differs based on the methods used and will be more likely if using stochastic optimization functions.

Table 5.11: Model compression

Name	**Model compression**
Description	Large and sparse networks are the most efficient way to get good performance. Quantization and weight pruning are the most common ways of compressing a model, and especially quantization can result in compression without performance degradation. Quantizing during training means the model will typically take longer to converge, which can be an issue. Compression can also serve as regularization in some cases. Since the best compression methods vary case by case, no general recommendations can be made.
Justification	[3], [5], [7], [10], [22], [24], [25], [26], [27], [29], [30], [32], [33], [45], [46], [48], [50], [56], [63]
Relation to lifecycle	Training and inference
Limitations	This is a complex topic, and if done haphazardly can result in unpredictable consequences to model performance. Models can start to make more errors after poorly done compression.

Table 5.12: Optimizing for inference

Name	**Optimizing for inference (deployed model)**
Description	Models that will be deployed should prioritize lowering the cost of inference even at the expense of increasing training costs. This can be done by methods such as knowledge distillation to make the deployed model smaller, and serving the deployed model in such a way as to cost the least amount of overhead possible. Large models should be deployed on optimized data center infrastructure, preferably cloud data centers. Smaller models should be deployed on the edge if possible to avoid unnecessary data transfer and to reduce the required data ingestion bandwidth. Even in edge contexts the amount of data transferred should be minimized. More generally the model should be served to users in a such a way that it best matches the use case.
Justification	A model that is deployed for use will eventually see its environmental impact be dominated by the cost of inference. [5], [13], [15], [24], [28], [44], [46], [62]
Relation to lifecycle	Model construction, deployment
Limitations	It requires some knowledge to know how to make inference cheaper. and access to resources to implement those ways.

Table 5.13: Measure metrics of model

Name	**Measure or approximate metrics related to models**
Description	Use tools to track and analyze impact of model during development and after deployment, both the positive impact and associated costs (need relevant tools and metrics). Some metrics such as the costs associated with building a specific model should be thought of at every point. This allows the project to analyze the cost-benefit ratio of the model and whether or not its use actually has a net positive impact. Special focus should be given to the computational cost of inference if the model is intended for deployment.
Justification	Collecting relevant metrics allows a DL project to actually know how much the model is helping and how much it costs to use it, enabling much more meaningful decision making. [5], [7], [8], [17], [23], [40], [48], [53], [54], [55], [65], [66]
Relation to lifecycle	Deployment of model, avoid overhead
Limitations	At present the metrics and measurement methods available are somewhat lacking in terms of capturing the big picture of the lifecycle.

Table 5.14: Report model metrics

Name	**Report metrics of models, including sustainability metrics**
Description	Once metrics about the model have been obtained, it would the advance the field as a whole if they were shared and reported openly, which would serve as an example and a benchmark to others. Also, if the model is provided to others with associated metrics then more informed decisions can be made. This might not be possible current data, but hopefully will be in the future.
Justification	[1], [5], [8], [12], [40], [47], [49], [52], [53], [60], [61]
Relation to lifecycle	Measure metrics related to model use
Limitations	Issues can arise if the chosen metrics are ill-suited for the project context in some way. For example it is possible that no useful information is conveyed even though the metrics have been accurately tracked.

Table 5.15: Hardware acceleration

Name	**Use of fitting hardware, Hardware acceleration**
Description	As much computational work as possible should be performed on the hardware that makes that work the most efficient due to a lack of overhead processing cost. CPU performs best in low-energy contexts with small networks, which can happen for example with edge computing, since it has lower average and peak energy requirements than a GPU. The lower latency in comparison to GPUs also makes it a better choice in cases where serial operations bottleneck parallel operations. In general use the GPU is notably faster than the CPU in performing computation for a model, and even though it uses more energy than the CPU the speedup almost always results in net savings. The GPU is generally good with CNNs due to 2D convolution. However, its use should be optimized with regard to utilization levels since just turning it on costs more energy than a CPU. TPU usage: The TPU requires large amounts of energy but is extremely fast at low precision parallel computation and should be used when available for large models housed in data centers. There is also a edge model TPU which is typically preferred for edge contexts, if accessible.
Justification	Hardware acceleration is self-evidently effective and necessary for efficient AI use.
Relation to lifecycle	Related costs, data centers
Limitations	This is relatively simple, with a possible problem being a lack of access to specific hardware that would be required.

6 Discussion

6.1 Answers to research questions

To address the results of the SLR, it is necessary to answer the research questions.

RQ1. What is the current state of the field regarding the computational complexity, size and energy consumption of deep learning models and what are the trends of these metrics?

Based on the SLR, models are still continuing to expand with regard to all relevant metrics mentioned with no end in sight. There is much resource waste associated with this trend considering the resulting very limited performance increase. DL models are so large that part of them is superfluous as a rule, with energy consumption especially being excessively large due to the lack of attention placed on it by the industry. These trends are issues by themselves but also greatly limit the potential of approaches such as edge intelligence and autonomous agents.

All this highlights the usefulness of the Green AI paradigm. It can be argued that criteria such as those presented in this work are absolutely necessary to improve the sustainability of the AI field, as infinite scalability is not feasible or ultimately even possible. Furthermore, more focus should be placed on exploring alternative paradigms and approaches over trying to scale up existing solutions.

RQ2. Does the literature present estimates and measuring methods for the environmental impacts, energy consumption, or manufacturing car-

bon footprint of deep learning training and usage based on factors such as model size or architecture?

There are many competing and disputed tools and strategies aiming to achieve this, so at least within the bounds of the SLR there is no clear consensus for the best approach. The main issues are that model factors do not directly correspond to increases in cost metrics, and there is no universally agreed upon metric that would facilitate meaningful comparison of models across tasks. Arguably such metrics might not be possible and focus should be placed on finding metrics that at least are accurate within a specific domain.

However it is a subject of much discussion in the literature, and some highlights have been presented in this work. Estimation methods are currently the most accurate approach, as there is no holistic measurement. Thus all measurements within this area will lead to an estimate or an incomplete picture. For anyone looking to analyze their own models in this manner, the most important thing is to consider the correlation and at least degree of magnitude -level accuracy of the estimation methods in use.

RQ3. Based on currently available research is it possible to develop big picture calculation and estimation criteria for optimizing deep learning for sustainability?

There have been some big picture approaches aiming to promote Green AI in the DL field. Within the SLR they are generally limited to discussion of proposed methods and tools, falling somewhat short of establishing tangible criteria. However, accurate calculation is exceedingly hard within the bounds of the DL field as is, and such this work was not able to develop reliable criteria for that. The main issue is the dominance of Red AI trends and practitioners leading to a lack of shared knowledge. This manifests in the SLR through problems such as being unable to perform effective measurement, being unable to replicate models or their results due

to lack of hyperparameter sharing, and being unable to compare performance of certain models due to excessive computational requirements. However, estimates and more general principles to optimizing deep learning have been developed from proposals within the SLR and the result is a somewhat mixed answer to the research question.

The big picture of the lifecycle is much more important as a starting point for achieving Green AI within the field than focusing on details, as specific model optimization techniques already in use share the interests of Green AI on a smaller scale. As such the criteria developed and presented within the work aim to establish good principles for approaching DL more sustainably in place of more concrete suggestions with restricted applicability.

6.2 Threats to validity

The solo nature of the work and the SLR make it possible that there are blind spots within the results. These could be caused either by errors in judgment or gaps in research. However there is no conflict of interest.

6.3 Practical implications

Based on the SLR it is both prudent and possible to perform DL research and practice more sustainably. The primary restriction for Green AI research at this point is that the industry has not adopted its principles and thus research is restricted, especially with regard to analyzing the impact of implementing Green AI on the closed source models most commonly used by the public.

The criteria presented should be useful to existing and new practitioners of deep learning, or organisations looking to incorporate artificial intelligence into their operations. However, their general nature and limitations mean that for most large

subdomains specialized advice could be more helpful. Such advise could not be presented in this work due to limitations of scope.

6.4 Further work

As the scope of this work is restricted to deep learning, an immediate avenue for future research is expanding the scope to examine alternative approaches to machine learning such as neuromorphic computing which is inherently energy-efficient. The human brain is an especially good source of inspiration, as it's naturally built to cut corners and identify patterns to save resources. Furthermore, more concrete research for finding new criteria, especially those able to calculate metrics for models, and for applying the ones developed in this work into the lifecycle of a DL model to ascertain the value of the criteria is highly advisable.

7 Conclusion

AI models are here to stay regardless of the form they will take in the future. As with any technology that can potentially improve the world around it, steps should be taken to investigate the impact of using said technology to ensure a net positive effect. Currently multiple trends are contributing to unnecessary expenditure of energy and an excessive need for more resources. Arising from these trends is a concern about the environmental impact of the field, which is beginning to be addressed on a research level by Green AI initiatives. This work aims to highlight and present solutions to remedy the harmful trends of the AI field.

The SLR performed in this work is meant to investigate the current state of the field. The SLR touches on all parts of the AI model lifecycle on some level, allowing for a big picture view of the situation. As a general notion, while facing many obstacles and a lack of industry support, Green AI research has been able to contribute novel findings including many opportunities to decrease the costs associated with AI.

The fact that so many papers in the SLR find an opportunity to improve efficiency without sacrificing performance is indicative of systemic potential for saving resources when Green AI is implemented in practice, and the criteria presented in this work are a step towards that implementation. Fundamentally the research field should shift towards considering a performance-efficiency tradeoff and exploring new approaches and paradigms instead of using as many resources as needed to marginally increase performance on benchmarks.

To hopefully advance the field towards sustainability, this thesis provides a set of criteria that should be approachable to all readers and applicable to all DL practice. It is clear that the theory for implementing more sustainable deep learning is there, but whether industry practice will shift towards sustainability is not. It is the opinion of the author that for many reasons including fiscal and environmental ones but also to advance the field of research it would be beneficial to implement Green AI.

References

[1] Y. I. Alzoubi and A. Mishra, "Green artificial intelligence initiatives: Potentials and challenges", vol. 468, 2024, ISSN: 0959-6526, 1879-1786. DOI: 10.1016/j.jclepro.2024.143090.

[2] M. Anselmo and M. Vitali, "A data-centric approach for reducing carbon emissions in deep learning", in *International Conference on Advanced Information Systems Engineering*, Springer, 2023, pp. 123–138.

[3] N. C. Frey et al., "Energy-aware neural architecture selection and hyperparameter optimization", in *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, IEEE, 2022, pp. 732–741.

[4] A. Badar et al., "Highlighting the importance of reducing research bias and carbon emissions in cnns", in *International Conference of the Italian Association for Artificial Intelligence*, Springer, 2021, pp. 515–531.

[5] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future", *Neurocomputing*, vol. 599, p. 128 096, 2024.

[6] M. A. Shafique, A. Munir, and J. Kong, "Deep learning performance characterization on gpus for various quantization frameworks", *AI*, vol. 4, no. 4, pp. 926–948, 2023.

[7] C. Clemm, L. Stobbe, K. Wimalawarne, and J. Druschke, "Towards green ai: Current status and future research", in *2024 Electronics Goes Green 2024+(EGG)*, IEEE, 2024, pp. 1–11.

[8] T. Eilam, P. Bello-Maldonado, B. Bhattacharjee, C. Costa, E. K. Lee, and A. Tantawi, "Towards a methodology and framework for ai sustainability metrics", in *Proceedings of the 2nd workshop on sustainable computer systems*, 2023, pp. 1–7.

[9] C. P. Ezenkwu, B. U.-A. Stephen, I. Affiah, and B. Daniel, "A green ai model selection strategy for computer-aided mpox detection", in *2023 IEEE AFRICON*, IEEE, 2023, pp. 1–6.

[10] G. C. Marinó, A. Petrini, D. Malchiodi, and M. Frasca, "Deep neural networks compression: A comparative survey and choice recommendations", *Neurocomputing*, vol. 520, pp. 152–170, 2023.

[11] N. Lenherr, R. Pawlitzek, and B. Michel, "New universal sustainability metrics to assess edge intelligence", *Sustainable Computing: Informatics and Systems*, vol. 31, p. 100 580, 2021.

[12] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai", *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[13] A. Karamchandani, A. Mozo, S. Gómez-Canaval, and A. Pastor, "A methodological framework for optimizing the energy consumption of deep neural networks: A case study of a cyber threat detector", *Neural Computing and Applications*, vol. 36, no. 17, pp. 10 297–10 338, 2024.

[14] *Blue angel | the german ecolabel*, `https://www.blauer-engel.de/en`, Accessed: 2026-05-11.

[15] C.-J. Wu et al., "Sustainable ai: Environmental implications, challenges and opportunities", *Proceedings of machine learning and systems*, vol. 4, pp. 795–813, 2022.

[16] *Prisma statement*, `https://www.prisma-statement.org/`, Accessed: 2026-05-11.

[17] R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of green ai", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 4, e1507, 2023.

[18] E. Barbierato and A. Gatti, "Toward green ai: A methodological survey of the scientific literature", *Ieee Access*, vol. 12, pp. 23 989–24 013, 2024.

[19] M. M. Rizvee, M. H. Rahman, P. Chakraborty, and S. Shomaji, "Understanding the innovations required for a green & secure artificial intelligence paradigm", in *2023 IEEE 16th Dallas Circuits and Systems Conference (DCAS)*, IEEE, 2023, pp. 1–6.

[20] R. Verdecchia, L. Cruz, J. Sallou, M. Lin, J. Wickenden, and E. Hotellier, "Data-centric green ai an exploratory empirical study", in *2022 international conference on ICT for sustainability (ICT4S)*, IEEE, 2022, pp. 35–45.

[21] A. Boumendil, W. Bechkit, and K. Benatchba, "On data selection for the energy efficiency of neural networks: Towards a new solution based on a dynamic selectivity ratio", in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2023, pp. 325–332.

[22] L. Balderas, M. Lastra, and J. M. Benítez, "An efficient green ai approach to time series forecasting based on deep learning", *Big Data and Cognitive Computing*, vol. 8, no. 9, p. 120, 2024.

[23] S. Rajput and T. Sharma, "Benchmarking emerging deep learning quantization methods for energy efficiency", in *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, IEEE, 2024, pp. 238–242.

[24] S. Wiedemann, K. Müller, and W. Samek, "Compact and computationally efficient representation of deep neural networks", *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, vol. 31, no. 3, pp. 772–785, 2020, ISSN: 2162-237X. DOI: 10.1109/TNNLS.2019.2910073.

[25] M. B. Noach and Y. Goldberg, "Compressing pre-trained language models by matrix decomposition", in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 884–889.

[26] Z. Wang, T. Luo, M. Li, J. T. Zhou, R. S. M. Goh, and L. Zhen, "Evolutionary multi-objective model compression for deep neural networks", *IEEE Computational Intelligence Magazine*, vol. 16, no. 3, pp. 10–21, 2021.

[27] S. S. Acmali, Y. Ortakci, and H. Seker, "Green ai-driven concept for the development of cost-effective and energy-efficient deep learning method: Application in the detection of eimeria parasites as a case study", *Advanced Intelligent Systems*, vol. 6, no. 7, p. 2 300 644, 2024.

[28] M. Dörrich, M. Fan, and A. M. Kist, "Impact of mixed precision techniques on training and inference efficiency of deep neural networks", *IEEE Access*, vol. 11, pp. 57 627–57 634, 2023. DOI: 10.1109/ACCESS.2023.3284388.

[29] X. Ma et al., "Non-structured dnn weight pruning—is it beneficial in any platform?", *IEEE transactions on neural networks and learning systems*, vol. 33, no. 9, pp. 4930–4944, 2021.

[30] L. Balderas, M. Lastra, and J. M. Benítez, "Optimizing convolutional neural network architectures", *Mathematics*, vol. 12, no. 19, p. 3032, 2024.

[31] L. B. Letaifa and J.-L. Rouas, "Towards green ai: Assessing the robustness of conformer and transformer models under compression", in *2024 32nd European Signal Processing Conference (EUSIPCO)*, IEEE, 2024, pp. 336–340.

[32] Y. Yang, L. Deng, S. Wu, T. Yan, Y. Xie, and G. Li, "Training high-performance and large-scale deep neural networks with full 8-bit integers", *Neural Networks*, vol. 125, pp. 70–82, 2020.

[33] L. Balderas, M. Lastra, and J. M. Benítez, "A green ai methodology based on persistent homology for compressing bert", *Applied Sciences*, vol. 15, no. 1, p. 390, 2025.

[34] B. Xu, S. Yan, L. Liu, and F.-M. Schleif, "Optimizing yolov5 for green ai: A study on model pruning and lightweight networks", in *International Workshop on Self-Organizing Maps, Learning Vector Quantization & Beyond*, Springer, 2024, pp. 196–205.

[35] A. M. Yokoyama, M. Ferro, and B. Schulze, "A multi-objective hyperparameter optimization for machine learning using genetic algorithms: A green ai centric approach", in *Ibero-American Conference on Artificial Intelligence*, Springer, 2022, pp. 133–144.

[36] Á. D. Reguero, S. Martínez-Fernández, and R. Verdecchia, "Energy-efficient neural network training through runtime layer freezing, model quantization, and early stopping", *Computer Standards & Interfaces*, vol. 92, p. 103 906, 2025.

[37] K. R. Foglia, V. Colao, and E. Ritacco, "Halpernsgd: A halpern-inspired optimizer for accelerated neural network convergence and reduced carbon footprint", in *International symposium on methodologies for intelligent systems*, Springer, 2024, pp. 296–305.

[38] R. Fischer, M. Wever, S. Buschjaeger, and T. Liebig, "Metaqure: Meta-learning from model quality and resource consumption", English, in *MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES: RESEARCH TRACK, PT VII, ECML PKDD 2024*, A. Bifet, J. Davis, T. Krilavicius, M. Kull, E. Ntoutsi, and I. Zliobaite, Eds., ser. Lecture Notes in Artificial Intelligence, vol. 14947, Springer International Publishing Ag, 2024, pp. 209–226, ISBN: 978-3-031-70367-6. DOI: 10.1007/978-3-031-70368-3_13.

[39] D. Castellanos-Nieves and L. García-Forte, "Strategies of automated machine learning for energy sustainability in green artificial intelligence", *Applied Sciences*, vol. 14, no. 14, p. 6196, 2024.

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

[41] F. S. Martínez, R. Parada, and J. Casas-Roma, "Co2 impact on convolutional network model training for autonomous driving through behavioral cloning", *Advanced Engineering Informatics*, vol. 56, p. 101 968, 2023.

[42] M. Barlaud and F. Guyard, "Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai", in *2020 25th international conference on pattern recognition (ICPR)*, IEEE, 2021, pp. 1566–1573.

[43] S. Wu, A. Hadachi, C. Lu, and D. Vivet, "Mott: A new model for multi-object tracking based on green learning paradigm", *AI Open*, vol. 4, pp. 145–153, 2023.

[44] T. Yarally, L. Cruz, D. Feitosa, J. Sallou, and A. van Deursen, "Batching for green ai-an exploratory study on inference", in *2023 49th Euromicro Con-*

ference on Software Engineering and Advanced Applications (SEAA), IEEE, 2023, pp. 112–119.

[45] N. Alizadeh and F. Castor, "Green ai: A preliminary empirical study on energy consumption in dl models across different runtime infrastructures", in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 134–139.

[46] F. Duran, S. Martinez-Fernandez, M. Martinez, and P. Lago, "Identifying architectural design decisions for achieving green ml serving", in *PROCEEDINGS 2024 IEEE/ACM 3RD INTERNATIONAL CONFERENCE ON AI ENGINEERING-SOFTWARE ENGINEERING FOR AI, CAIN 2024*, Assoc Computing Machinery, 2024, pp. 18–23, ISBN: 979-8-4007-0591-5.

[47] V. Chavan, P. Koch, M. Schlüter, and C. Briese, "Towards realistic evaluation of industrial continual learning scenarios with an emphasis on energy consumption and computational footprint", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 506–11 518.

[48] S. Rajput, T. Widmayer, Z. Shang, M. Kechagia, F. Sarro, and T. Sharma, "Enhancing energy-awareness in deep learning through fine-grained energy measurement", *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 8, pp. 1–34, 2024.

[49] P. Duran, J. Castaño, C. Gómez, and S. Martínez-Fernández, "Gaissalabel: A tool for energy labeling of ml models", in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 622–626.

[50] J. Kannan, S. Barnett, A. Simmons, T. Selvi, and L. Cruz, "Green runner: A tool for efficient deep learning component selection", in *Proceedings of the*

IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, 2024, pp. 112–117.

[51] S. Rajput, M. Kechagia, F. Sarro, and T. Sharma, "Greenlight: Highlighting tensorflow apis energy footprint", in *Proceedings of the 21st International Conference on Mining Software Repositories*, 2024, pp. 304–308.

[52] R. Fischer, M. Jakobs, S. Mücke, and K. Morik, "A unified framework for assessing energy efficiency of machine learning", in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2022, pp. 39–54.

[53] D. Patterson et al., "Carbon emissions and large neural network training", *arXiv preprint arXiv:2104.10350*, 2021.

[54] S. del Rey, S. Martínez-Fernández, L. Cruz, and X. Franch, "Do dl models and training environments have an impact on energy consumption?", in *2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, IEEE, 2023, pp. 150–158.

[55] S. A. Budennyy et al., "Eco2ai: Carbon emissions tracking of machine learning models as the first step towards sustainable ai", in *Doklady mathematics*, Springer, vol. 106, 2022, S118–S128.

[56] Y. Xu, S. Martínez-Fernández, M. Martinez, and X. Franch, "Energy efficiency of training neural network architectures: An empirical study", *arXiv preprint arXiv:2302.00967*, 2023.

[57] S. Ali, E. T. Fapi, B. Jaumard, and A. Planche, "Focus on carbon dioxide footprint of ai/ml model training", in *2024 Intelligent Methods, Systems, and Applications (IMSA)*, IEEE, 2024, pp. 524–529.

[58] A. Boumendil, W. Bechkit, and K. Benatchba, "On-device deep learning: Survey on techniques improving energy efficiency of dnns", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 7806–7821, 2024.

[59] L. Bouza, A. Bugeau, and L. Lannelongue, "How to estimate carbon footprint when training deep learning models? a guide and review", *Environmental Research Communications*, vol. 5, no. 11, p. 115 014, 2023.

[60] T. Tornede, A. Tornede, J. Hanselle, F. Mohr, M. Wever, and E. Hüllermeier, "Towards green automated machine learning: Status quo and future directions", *Journal of Artificial Intelligence Research*, vol. 77, pp. 427–457, 2023.

[61] Y. Wang et al., "Benchmarking the performance and energy efficiency of ai accelerators for ai training", in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, IEEE, 2020, pp. 744–751.

[62] S. Samsi et al., "From words to watts: Benchmarking the energy costs of large language model inference", in *2023 IEEE high performance extreme computing conference (HPEC)*, IEEE, 2023, pp. 1–9.

[63] S. Georgiou, M. Kechagia, T. Sharma, F. Sarro, and Y. Zou, "Green ai: Do deep learning frameworks have different costs?", in *Proceedings of the 44th international conference on software engineering*, 2022, pp. 1082–1094.

[64] R. Al-Qurran, M. Al-Ayyoub, and A. Shatnawi, "Plant classification in the wild: Energy evaluation for deep learning models", *Multimedia Tools and Applications*, vol. 81, no. 21, pp. 30 143–30 167, 2022.

[65] R. Caspart et al., "Precise energy consumption measurements of heterogeneous artificial intelligence workloads", in *International Conference on High Performance Computing*, Springer, 2022, pp. 108–121.

[66] S. Lahmer, A. Khoshsirat, M. Rossi, and A. Zanella, "Energy consumption of neural networks on nvidia edge boards: An empirical model", in *2022 20th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WiOpt)*, IEEE, 2022, pp. 365–371.

[67] X. Qiu et al., "A first look into the carbon footprint of federated learning", *Journal of Machine Learning Research*, vol. 24, no. 129, pp. 1–23, 2023.

[68] N. Nijkamp, J. Sallou, N. Van Der Heijden, and L. Cruz, "Green ai in action: Strategic model selection for ensembles in production", in *Proceedings of the 1st ACM International Conference on AI-Powered Software*, 2024, pp. 50–58.

[69] Y. Mao, X. Yu, K. Huang, Y.-J. A. Zhang, and J. Zhang, "Green edge ai: A contemporary survey", *Proceedings of the IEEE*, vol. 112, no. 7, pp. 880–911, 2024.

[70] A. Asperti, D. Evangelista, and E. Loli Piccolomini, "A survey on variational autoencoders from a green ai perspective", *SN Computer Science*, vol. 2, no. 4, p. 301, 2021.

Glossary

ANN artificial neural network. 5

CNN convolutional neural network. 22

CO2e Carbon dioxide equivalent, a standardised unit used to quantify carbon footprint. 12

CPU central processing unit. 15

dataset A collection of data, typically consisting of samples related to a specific problem domain. These are needed to train models. 6, 19

DL deep learning. 1, 5

DNN deep neural network. 6

emission intensity How much pollutant is released from a specific activity, such as energy production, per the amount of work done. 34

epoch During model training, a complete pass through the training dataset where each sample is fed to the model and parameter adjustments are made based on the results. 17

foundation model Large, high performance models with wide knowledge across many tasks. This means they can be used for knowledge distillation or transfer learning, to train smaller models with less work. 20

GPU graphics processing unit. 15

hyperparameter Parameters governing the learning process of a neural network during training, such as by how much the weights are updated. 7

inference A trained neural network making predictions on unseen data, in other words, being used. 11

knowledge distillation Building a smaller model by transferring knowledge from a larger one to attain similar performance with lower costs. 20

ML machine learning. 5

NAS neural architecture search. 11, 31, 58

NLP natural language processing. 18

optimizer The algorithm that adjusts model parameters during training, aiming to minimize a specific objective function. Often this function is loss, or the errors made by the network. 19

Pruning Removing parameters, such as neurons or their connections, from a neural network. 19

PUE power usage effectiveness. 14

quantization Mapping numerical values from a larger set to a smaller set, usually with some information loss. Rounding to the nearest integer is a form of quantization. 20

SGD stochastic gradient descent. 21

TPU tensor processing unit. 15

transfer learning Training a model with pre-existing similar knowledge to handle a new problem domain, usually with limited training data available. This boosts performance in comparison to being trained on only the new domain. 60

transformer A large language model built on the attention mechanism. Attention allows the models to pay attention to a large context window, which led to the current boom of generative AI models. 7