A Review of Software Engineering Frameworks and Libraries for Green AI

By Noura Abdul Majid

Declaration:

I declare that I have personally prepared this assignment. The work is my own, carried out personally by me unless otherwise stated and has not been generated using Artificial Intelligence tools unless specified as a clearly stated approved component of the assessment brief. All sources of information, including quotations, are acknowledged by means of the appropriate citations and references. I declare that this work has not gained credit previously for another module at this or another University.

I understand that plagiarism, collusion, copying another student and commissioning (which for the avoidance of doubt includes the use of essay mills and other paid for assessment writing services, as well as unattributed use of work generated by Artificial Intelligence tools) are regarded as offences against the University's Assessment Regulations and may result in formal disciplinary proceedings.

I understand that by submitting this assessment, I declare myself fit to be able to undertake the assessment and accept the outcome of the assessment as valid.

1. Background

Artificial Intelligence (AI), defined as computational modelling that can solve sophisticated tasks [1], has achieved numerous technological breakthroughs. However, AI requires a large amount of computing power, usually resulting in large amounts of energy used. Deep Learning (DL) is a subset of AI comprised of layers of neural networks, structures inspired by the human brain, used to learn from data. [2] A single DL algorithm has been found to produce as much CO_2 as that produced by five cars across their lifetime [3]. More generally, the computing resources required to train the largest AI models have doubled every 3.4 years, showing an exponential increase [4]. This raises questions surrounding the environmental impact of software, especially AI tools. Moreover, until recently, papers have mostly focussed on the performance of AI models in terms of metrics such as accuracy [5].

1.1 Green AI Terminology

The phrase 'Green AI' in the literature has been found to refer both to eco-friendly AI tools (sometimes termed 'Green-in-AI' [6]), as well as the application of AI for eco-friendly outcomes in other areas, such as to make farming more efficient (sometimes termed 'Green-by-AI'). According to a highly-cited paper [7], the former had received little interest in the literature as of 2021, in comparison to the latter. Since then, however, several papers have been published with a focus on the former (see Related Work). For the rest of this review, Green AI will refer to Green-in-AI.

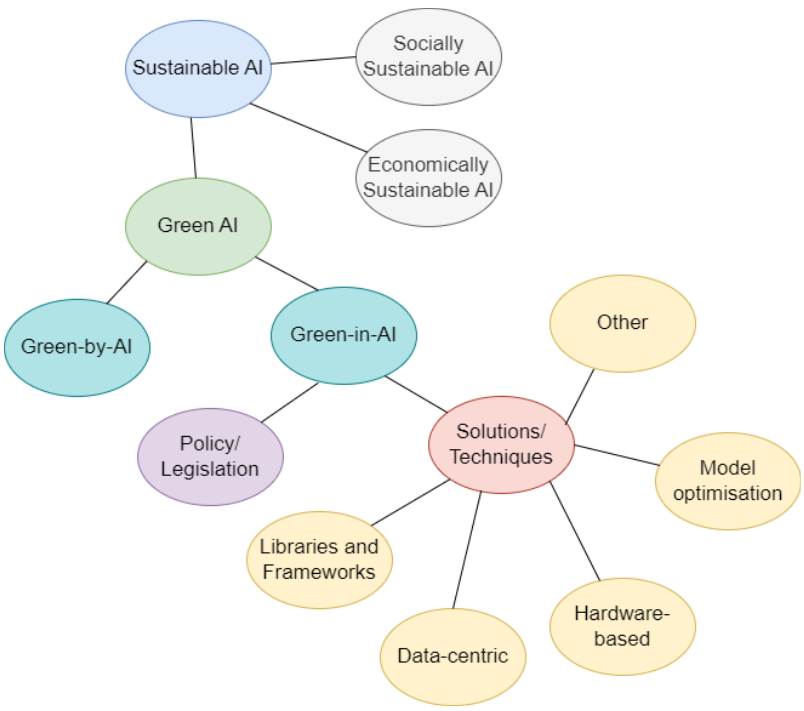

Figure 1: Taxonomy of Green AI Literature

1.2 Research Questions and Rationale for Review

An ethically-aware software engineer would need to make decisions regarding not just accuracy, speed, cost, but also eco-friendliness of the software. To help with the above with regard to Green principles in relation to AI software, this literature review aims to answer the following research question:

RQ1: What tools and techniques are available to minimise the environmental impact of AI, which involve the use of frameworks and libraries?

To our knowledge, this is the first literature review to focus on libraries and frameworks for Green AI. Numerous prior reviews, such as several papers in Related Work, have called for the increased adoption of Green AI methods, and highlighting such solutions should therefore encourage a more widespread adoption of Green AI techniques by software developers.

Literature Review

2. Methodology

A search was performed on the ACM Digital Library, using the following keywords and related words:

Keyword/ phrase	Related words/ phrases
Green	green in AI green-in-AI Sustainab* Efficien* Eco-friendl* Eco friendl* Environment* Ecolog* Carbon Energy
Artificial intelligence	AI Machine learning ML Deep learning DL
Librar*	Framework* Package* API*

A search string was generated using the ACM digital library advanced search features (see Appendix I). The search was performed on 29/12/2024.

Inclusion criteria:

- Published from 2014 onwards.
- Primary studies.
- Empirical studies.
- About Green AI.
- Technique used to increase energy eco-friendliness should involve use of libraries and/ or frameworks.

Exclusion criteria:

- Incomplete papers or stand-alone abstracts.
- Reporting efficiency as runtime only. This has been found to be heavily dependent on hardware and other factors. [5]
- Focusing solely on social or economic sustainability instead of environmental.
- Solely focusing on Green-by-AI.
- Not about software techniques.
- Focusing on the following: mobile devices, fog, edge devices and Internet of Things.
- Specifically having the word 'efficiency'/ 'efficient' in the title, but these and related words/ phrases in this category are not mentioned in the abstract.

354 search results were retrieved. 5 retracted articles and 2 duplicates were excluded, leaving 347 unique results. For 174 of these, abstracts and titles were read, with 49 articles included; for the remainder, titles and/ or abstracts were scanned due to time constraints, resulting in a further 26 articles included. The total included 75 articles went through another selection round, during which titles, abstracts and/or parts of full text were read, identifying 3 articles for inclusion. A further 3 articles to be included were identified from other sources, and a further 4 articles were identified from the dataset of a systematic review [8], resulting in 10 articles included for full text reading, including 1 grey literature document. Due to time constraints, only 4 of these articles had their full text read, and the remainder had their full text scanned to identify useful information.

3. Related Work

Eight related reviews about Green AI techniques have been outlined in Appendix II. Six of the reviews did not focus on any one type of technique. However, one review [6] focussed only on techniques to alter the AI model itself, and another review [9] focussed only on architectural decisions involved in the inference phase for Green AI. The 'libraries' technique identified by [8] provided the inspiration for the topic of this review; however, the write-up for their systematic review does not go into much detail on this technique and only cites one paper, though they stated that they had identified eight papers as belonging to this category in total.

Besides [8], no other reviews about Green AI, to our knowledge, identified libraries or frameworks generally as a Green AI technique, pointing toward a gap in the literature.

4. Results

The parts of the papers identified which are relevant to this review have been summarised in the two tables below.

Table 1: Outline of Key Experiments

Authors and Year	Library/ Framework	Key Experiments	Model	ML Phase
Georgiou et al. [10] (2022)	TensorFlow, PyTorch	Comparing energy consumption for combinations of frameworks, models and phases (training/ inference)	DL	Training + inference[1]
Alizadeh and Castor [11] (2024)	Frameworks: TensorFlow, PyTorch, MXNet, ONNX Execution providers (libraries): TensorRT, CUDA	Comparing energy efficiency of runtime environments of 3 frameworks, and execution providers.	DL:	Inference
Shanbhag et al. [12] (2022)	Default/ built-in libraries for Tensor operations	Analysed StackOverflow posts to produce energy-efficiency recommendations. Survey of 14 DL developers on recommendations	DL	Training + inference
Rajput et al. [13] (2024)	TensorFlow APIs	Used FeCoM (fine grained energy consumption measurement tool) to measure energy consumption of APIs.	DL	Training + inference
Rajput et al. [14] (2024)	TensorFlow APIs	Measured energy consumption of TensorFLow APIs under different configurations.	DL	Training + inference
Duran et al. [15] (2024)	Execution providers: CPU, CUDA.	Analysed seven combinations of runtime engines and execution providers for energy efficiency.	DL: small language model (SLM)	Inference
Li et al. [16] (2016)	TensorFlow, Caffe, Torch, MXNet, Nervana GPU libraries: cuDNN vs native implementations of GPU library.	Tested energy efficiency of 4 models, 5 training frameworks, hardware (CPU, GPU) and 2 hardware libraries.	DL	Training
Garc´ıa-Vico et al. [17] (2021)	7 libraries used to build Spiking Neural Network[2] (SNN).	Tested each of the libraries' ability to convert a traditional NN to a SNN. Accuracy of the SNN inference and time taken to train were measured.	DL: SNN	N/A
Yao et al. [18] (2021)	TensorFlow, TensorRT	Carried out inference using 3 CNN models on two frameworks, on one GPU.	DL: CNNs	Inference

1 Training refers to the development of an AI model on training data, whereas inference is when a developed model is utilised on new, unseen data to make predictions, classifications, etc.

2 Neural Networks (NNs) here refer to artificial neural networks, which are DL models that are modelled after the human brain, comprising of layers of interconnected nodes. They have many applications, including natural language processing (NLP); the intelligent processing of speech by AI.

Sun et al. [19] (2021)	GPU acceleration library: cuDNN. CPU acceleration libraries. Frameworks: TensorFlow, Torch, MXNet, Caffe, CXXNet.	Training of 5 different CNN models done with and without cuDNN acceleration library. Training of CNN models done with 3 CPU acceleration libraries. 5 DL frameworks used when training one CNN model; energy and time taken are measured.	DL: CNNs	Training

Table 2: Key Findings

Reference	**Key Experimental Findings**
[10]	Which framework is most efficient depends on the model and whether training/ inference phase. TensorFlow is more efficient for most scenarios, but PyTorch is more efficient for a significant number of scenarios. Both have similar accuracy. Found that some functions consume most energy within framework.
[11]	Struggled to find general rules as performance difficult to predict. PyTorch and MXNet more efficient than TensorFlow for small batch sizes, but all three comparable for large batch sizes. Variations and exceptions with certain models, however. In terms of runtime, TensorRT always outperforms CUDA
[12]	Theme to use built-in libraries to execute operations rather than writing custom operations to optimise energy efficiency. 85.8% of developers agreed that this helps to optimise energy efficiency.
[13]	Energy consumption of APIs strongly correlates with size of dataset and execution time.
[14]	Produced dataset by measuring energy consumption of 527 TensorFlow APIs. There is a high variability in energy consumed for different operations by a single API, and also a large difference in energy consumption between different APIs.
[15]	CUDA was the most efficient execution provider, and Torch + CUDA is the most efficient combination.
[16]	Torch was the most efficient framework on one GPU, and was tied in with Nervana and Caffe as most efficient on another GPU. TensorFlow and MXNet are the least efficient. CuDNN was more efficient than native libraries. Runtime was correlated to efficiency.
[17]	SNN built using Norse library was most accurate, and as accurate as original NN model. Newer libraries had better accuracy and less training time compared to mature libraries.
[18]	TensorRT is 1.53x more energy efficient than TensorFlow.
[19]	Training with cuDNN results in average reduction in energy by 3.91%. Optimising CPU providers results in 38.7% lower energy consumption.

	CXXNet (most efficient) 21.9% more energy efficient than TensorFlow (least efficient). Torch also inefficient.

5. Critical Evaluation

Frameworks and library-related software decisions have been found to result in great changes in energy consumption, uncovering this as a promising technique for Green AI. For example, different APIs consume very different amounts of energy [14], one framework is 1.5x more efficient than the other [18], and using CXXNet as a training framework saves 21.9% more energy compared to using TensorFlow [19].

All studies are from 2020 onwards, with one exception [16], highlighting that this is an area with very recent interest.

All of the papers identified have specifically focussed on the use of frameworks and libraries with DL models, even though all forms of AI models were included in the search; therefore, there has not been enough literature to generate any information for non-DL AI models. Two unique examples of DL models are as follows. Small language models (SLMs) were used in one study [15] , which highlighted their relevance to smaller companies with limited budgets, who are more likely to take an interest in energy conservation. However, it can be argued that energy efficiency is even more important for larger, energy intensive models. Spiking Neural Networks (SNNs) were another notable model [17]; they are a type of neural network modelled closely after the human brain, which has shown great promise in energy reduction. This study [17] stands out amongst those included in that it does not measure the energy efficiency of the models, but rather only measures their accuracy; this is nonetheless relevant to our review, as we assume the models *will* result in significant energy savings regardless; we are only concerned about their lack of accuracy hindering their use. However, quantifying the energy saving with SNNs would still have been useful here.

Another study which did not include a measure of energy is [12], however it did include a survey which helped to validate the results. Due to the low sample size of the survey, this study may perhaps be backed by the least evidence, however it is nonetheless useful as it highlighted the utility of StackOverflow posts in energy efficiency research, uncovering that this is a common concern among developers. Its premise that native implementations were always more efficient was in contradiction to evidence by [16].

However, it is difficult to predict the effect of libraries/ frameworks on energy use in specific situations, due to their effect being dependent on a number of other factors, such as dataset size [13], GPU [16] , and AI model [10]. This lack of ability to predict which library/ framework is most energy-efficient is exacerbated by the present shortage of documentation regarding their effect on energy efficiency, as highlighted by several papers including [10]. Attempts to produce such documentation appear to be underway [14]. As a practical solution to this for developers in the meantime, two papers [10], [11] recommended that smaller implementations of AI models are run before larger implementations, to test the efficiency of the proposed approach, and possibly adjust frameworks/ libraries used. Given the findings across all the papers within this review, this is sound advice.

Moreover, the 2016 study [16] considers GPUs such as Titan X, which are less commonly used today and for which support is phasing out, highlighting the importance of not just comprehensive, but timely energy documentation.

The majority of the studies only tested a small number of frameworks and configurations, and it appears to be an intensive process. This suggests that potentially the large companies developing AI software should play a greater role in producing the documentation, particularly as native libraries may be more efficient [12]. Moreover

6. Conclusion

We reviewed the literature on the use of frameworks and libraries for greener AI, finding that framework and library selection tended to have a large effect on the energy consumption of AI models, showing great promise in reducing the environmental footprint of these models. We presented what has been documented so far, and made recommendations for practice and research.

References

[1] Google, ‘Machine Learning Glossary’, Google for Developers. Accessed: Dec. 31, 2024. [Online]. Available: https://developers.google.com/machine-learning/glossary

[2] ‘What is Deep Learning? Applications & Examples’, Google Cloud. Accessed: Jan. 04, 2025. [Online]. Available: https://cloud.google.com/discover/what-is-deep-learning

[3] E. Strubell, A. Ganesh, and A. McCallum, ‘Energy and Policy Considerations for Modern Deep Learning Research’, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 09, Art. no. 09, Apr. 2020, doi: 10.1609/aaai.v34i09.7123.

[4] D. Amodei and D. Hernandez, ‘AI and compute’. Accessed: Dec. 31, 2024. [Online]. Available: https://openai.com/index/ai-and-compute/

[5] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, ‘Green AI’, *Commun ACM*, vol. 63, no. 12, pp. 54–63, Nov. 2020, doi: 10.1145/3381831.

[6] M. Gutiérrez, M. Á. Moraga, F. García, and C. Calero, ‘Green IN Artificial Intelligence from a Software Perspective: State-of-the-Art and Green Decalogue’, *ACM Comput Surv*, vol. 57, no. 3, p. 64:1-64:30, Nov. 2024, doi: 10.1145/3698111.

[7] A. Van Wynsberghe, ‘Sustainable AI: AI for sustainability and the sustainability of AI’, *AI Ethics*, vol. 1, no. 3, pp. 213–218, Aug. 2021, doi: 10.1007/s43681-021-00043-6.

[8] R. Verdecchia, J. Sallou, and L. Cruz, ‘A systematic review of Green AI’, *WIREs Data Min. Knowl. Discov.*, vol. 13, no. 4, p. e1507, 2023, doi: 10.1002/widm.1507.

[9] F. Durán, S. Martinez-Fernandez, M. Martinez, and P. Lago, ‘Identifying Architectural Design Decisions for Achieving Green ML Serving’, in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, in CAIN ’24. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 18–23. doi: 10.1145/3644815.3644962.

[10] 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, in ICSE '22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 1082–1094. doi: 10.1145/3510003.3510221.

[11] N. Alizadeh and F. Castor, 'Green AI: A Preliminary Empirical Study on Energy Consumption in DL Models Across Different Runtime Infrastructures', in *2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, Apr. 2024, pp. 134–139. Accessed: Dec. 31, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10556302

[12] S. Shanbhag, S. Chimalakonda, V. S. Sharma, and V. Kaulgud, 'Towards a Catalog of Energy Patterns in Deep Learning Development', in *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, in EASE '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 150–159. doi: 10.1145/3530019.3530035.

[13] 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 Trans Softw Eng Methodol*, vol. 33, no. 8, Dec. 2024, doi: 10.1145/3680470.

[14] S. Rajput, M. Kechagia, F. Sarro, and T. Sharma, 'Greenlight: Highlighting TensorFlow APIs Energy Footprint', in *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*, Apr. 2024, pp. 304–308. Accessed: Jan. 05, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10555878/?arnumber=10555878

[15] F. Durán, M. Martinez, P. Lago, and S. Martínez-Fernández, 'Energy consumption of code small language models serving with runtime engines and execution providers', Dec. 19, 2024, *arXiv*: arXiv:2412.15441. doi: 10.48550/arXiv.2412.15441.

[16] D. Li, X. Chen, M. Becchi, and Z. Zong, 'Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on CPUs and GPUs', in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, Oct. 2016, pp. 477–484. doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.76.

[17] Á. M. García-Vico and F. Herrera, 'A Preliminary Analysis on Software Frameworks for the Development of Spiking Neural Networks', in *Hybrid Artificial Intelligent Systems: 16th International Conference, HAIS 2021, Bilbao, Spain, September 22–24, 2021, Proceedings*, Berlin, Heidelberg: Springer-Verlag, Sep. 2021, pp. 564–575. doi: 10.1007/978-3-030-86271-8_47.

[18] C. Yao *et al.*, 'Evaluating and analyzing the energy efficiency of CNN inference on high-performance GPU', *Concurr. Comput. Pract. Exp.*, vol. 33, no. 6, p. e6064, 2021, doi: 10.1002/cpe.6064.

[19] Y. Sun *et al.*, 'Evaluating Performance, Power and Energy of Deep Neural Networks on CPUs and GPUs', in *Theoretical Computer Science*, Z. Cai, J. Li, and J. Zhang, Eds., Singapore: Springer, 2021, pp. 196–221. doi: 10.1007/978-981-16-7443-3_12.

[20] Z. Chen, M. Wu, A. Chan, X. Li, and Y.-S. Ong, 'Survey on AI Sustainability: Emerging Trends on Learning Algorithms and Research Challenges [Review Article]', *IEEE Comput. Intell. Mag.*, vol. 18, no. 2, pp. 60–77, May 2023, doi: 10.1109/MCI.2023.3245733.

[21] E. Barbierato and A. Gatti, 'Toward Green AI: A Methodological Survey of the Scientific Literature', *IEEE Access*, vol. 12, pp. 23989–24013, 2024, doi: 10.1109/ACCESS.2024.3360705.

[22] 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. 128096, Sep. 2024, doi: 10.1016/j.neucom.2024.128096.

[23] A. Tabbakh, L. Al Amin, M. Islam, G. M. I. Mahmud, I. K. Chowdhury, and M. S. H. Mukta, ‘Towards sustainable AI: a comprehensive framework for Green AI’, *Discov. Sustain.*, vol. 5, no. 1, p. 408, Nov. 2024, doi: 10.1007/s43621-024-00641-4.
[24] J. Xu, W. Zhou, Z. Fu, H. Zhou, and L. Li, ‘A Survey on Green Deep Learning’, Nov. 10, 2021, *arXiv*: arXiv:2111.05193. doi: 10.48550/arXiv.2111.05193.

Appendices

Appendix I - Search String

[[**Title**: green] **OR** [**Title**: sustainab*] **OR** [**Title**: efficien*] **OR** [**Title**: eco-friendl*] **OR** [[**Title**: eco] **AND** [**Title**: friendl*]] **OR** [**Title**: environment*] **OR** [**Title**: ecolog*] **OR** [**Title**: carbon] **OR** [**Title**: energy*]] **AND** [[**Title**: "artificial intelligence"] **OR** [**Title**: ai] **OR** [**Title**: "machine learning"] **OR** [**Title**: ml] **OR** [**Title**: "deep learning"] **OR** [**Title**: dl]]
AND
[[**Abstract**: librar*] **OR** [**Abstract**: framework*] **OR** [**Abstract**: package*] **OR** [[**Abstract**: api*] **AND** [[**Abstract**: green] **OR** [**Abstract**: green-in-ai] **OR** [**Abstract**: "green in ai"] **OR** [**Abstract**: sustainab*] **OR** [**Abstract**: efficien*] **OR** [**Abstract**: eco-friendl*] **OR** [[**Abstract**: eco] **AND** [**Abstract**: friendl*]] **OR** [**Abstract**: environment*] **OR** [**Abstract**: ecolog*] **OR** [**Abstract**: carbon] **OR** [**Abstract**: energy*]] **AND** [[**Abstract**: "artificial intelligence"] **OR** [**Abstract**: ai] **OR** [**Abstract**: "machine learning"] **OR** [**Abstract**: ml] **OR** [**Abstract**: "deep learning"] **OR** [**Abstract**: dl]]]]
AND [**E-Publication Date**: (01/01/2014 **TO** 12/31/2024)]

Appendix II - Table of Related Work

Review	Summary	Categories of techniques	Subcategories
[8]	A systematic review of solutions, observations and policy papers about Green AI	Hyperparameter tuning, model benchmarking, deployment, precision-energy trade-off, algorithm design, libraries, data-centric approaches, etc.	N/A
[20]	A review encompassing Green AI and socially sustainable AI (reviewed separately). Claims to be the first review to focus on the technical aspects of sustainable AI (as opposed to policy).	Model compression techniques	Quantisation, pruning, knowledge distillation
		Data-centric training approaches	Transfer learning, active learning, etc.
[21]	Methodological survey	Hardware-based	N/A
		Training	N/A
		Learning	N/A
		Heuristics	Model compression, early stopping, data augmentation, sparsity models, etc.
[22]	Review	Hardware optimisation	GPU selection, etc.

		Algorithm optimisation	Model compression, training approaches, etc.
		Data centre optimisation	Dynamically managing server loads, etc.
		Pragmatic scaling factors	Limiting number of times algorithm is run, time spent on hyperparameter tuning, etc.
[23]	Review	Software	Model compression, efficient algorithms, approximate computing, transfer learning
		Hardware	Use of TPUs
		Designing and operating efficient data centres	N/A
[6]	Review focussing specifically on Green AI techniques which directly modify the AI algorithm, not including model optimisation techniques.	Not relevant to this review.	N/A
[9]	Review focussing specifically on Green AI architecture techniques relating to ML serving (ML models binding to the ML system during their inference phase, which is when they are run)	Choice of serving infrastructure (most significant factor), etc.	N/A
[24]	Systematic review	Compact neural networks	N/A
		Training strategies	Progressive training, hyperparameter optimisation, etc.
		Inference approaches	Model optimisation techniques e.g. pruning
		Data-efficient approaches	Active learning, few shot learning, etc.