Array 29 (2026) 100652

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Green AI techniques for reducing energy consumption in AI systems

Sunawar Khan [a], Naila Sammar Naz [a], Tehseen Mazhar [a,i,*], Muhammad Usman Tariq [b], Tariq Shahzad [c], Sghaier Guizani [d,**], Habib Hamam [e,f,g,h]

[a] School of Computer Science, National College of Business Administration and Economics, Lahore, 54000, Pakistan
[b] Department of Marketing, Operations, and Information Systems, College of Business, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[c] Department of Computer Engineering, COMSATS University Islamabad, Sahiwal Campus, Sahiwal, 57000, Pakistan
[d] College of Engineering, Alfaisal University, Riyadh, 11533, Saudi Arabia
[e] Faculty of Engineering, Uni de Moncton, Moncton, NB, E1A3E9, Canada
[f] School of Electrical Eng., University of Johannesburg, Johannesburg, 2006, South Africa
[g] International Institute of Techno. and Management (IITG), Av. Grandes Ecoles, Libreville, Gabon
[h] Bridges for Academic Excellence - Spectrum, Tunis, Tunisia
[i] Department of Computer Science and Information Technology, School Education Department, Government of Punjab, Layyah 31200, Pakistan

A R T I C L E I N F O

Keywords:
Green AI
Energy efficiency
Model compression
Low-precision computing
Sustainable AI

A B S T R A C T

This systematic review synthesizes current evidence on energy-reduction techniques across algorithmic, hardware, and infrastructure layers of AI systems. Model compression and knowledge distillation (e.g., DistilBERT) deliver ~60 % faster inference with ~40 % fewer parameters while retaining ~97 % of baseline performance. Low-precision computation (quantization) yields up to ~50 % energy reductions, and architecture-level strategies—such as neural architecture search and depthwise-separable convolutions in MobileNetV2—significantly lower compute and memory demand. Specialized accelerators (TPUs) and neuromorphic hardware further improve efficiency, while data-center measures (advanced cooling, virtualization, renewable integration) reduce system-level consumption. For generative-AI workloads, distillation, quantization, efficient architectures, and accelerator-optimized inference remain the primary pathways to lowering both training and inference energy. Across studies, recurring gaps include inconsistent energy-metric reporting, limited standardized benchmarks, and a dominant focus on accuracy over efficiency. Regulatory progress is uneven: the EU has introduced stronger transparency requirements, whereas comparable obligations are not yet global. Review limitations include heterogeneous methodologies and incomplete transparency artifacts, which restrict cross-study comparability. Future research directions include algorithm–hardware co-design, neuromorphic methods, energy-harvesting AI devices, improved data-center operations, and explainable-AI tools to support reliable, energy-aware deployment at scale.

1. Introduction

1.1. Classical AI

The growing size and complexity of modern AI models have raised concerns across scientific and industrial communities about energy consumption. Scaling models for training and deployment demands substantial compute, increasing both operational costs and environmental footprint. As a result, AI energy efficiency—reducing power use while maintaining or improving performance—has become a central design objective. Recent work highlights Green AI as a key pillar of sustainable AI, emphasizing techniques such as pruning, quantization, and energy-efficient hardware to lower energy use.

A classical AI life cycle comprises five stages: Planning and Data Collection, Model Development, Model Training, Model Deployment, and Monitoring and Maintenance. Planning/Data Collection and Monitoring/Maintenance rely largely on estimates, whereas Model Development, Training, and Deployment are better supported by empirical

This article is part of a special issue entitled: GenAI and Green Tech published in Array.
* Corresponding author.
** Corresponding author.
E-mail addresses: sunawarkhan@gmail.com (S. Khan), nailanaz@ieee.org (N.S. Naz), tehseenmazhar719@gmail.com (T. Mazhar), Muhammad.kazi@adu.ac.ae (M.U. Tariq), tariq@cuisahiwal.edu.pk (T. Shahzad), sguizani@alfaisal.edu (S. Guizani), habib.hamam@umoncton.ca (H. Hamam).

https://doi.org/10.1016/j.array.2025.100652
Received 31 July 2025; Received in revised form 6 December 2025; Accepted 15 December 2025
Available online 23 December 2025

evidence and typically dominate financial and compute costs. Stage-level energy use is heterogeneous. For 2019–2021, Google directly measured that approximately ~60 % of ML energy is attributable to inference (deployment) and ~40 % to training [1]; no confidence intervals are reported. Because comparable standardized measurements for Planning and Development are unavailable, those stages are presented schematically only. Actual shares vary with workload characteristics, system scale, siting/grid carbon intensity, cooling Power usage efficiency (PUE), and utilization [1,2]. The distribution of power consumption at each stage is summarized in Fig. 1.

Planning and Development are illustrative only; no quantitative estimate is implied or should be inferred for these stages.

AI has transformed numerous sectors and has delivered significant economic and societal benefits, reshaping communications in healthcare, banking, transportation, and entertainment [3]. Research, technological innovation, human-computer interaction, and autonomous systems rely on ML, NLP, computer vision, and other technologies [4,5]. This growth demands vast infrastructure whose energy consumption is a major challenge. Training deep learning architectures can require substantial energy—such that GPT-3 or AlphaGo require large-scale compute resources (the GPT-3 model alone is reported to have used approximately 1287 MWh for training and ~564 MWh per day for sustained inference, as cited in published technical documentation) [6–8]. As AI-focused data centers increase electricity consumption, related greenhouse-gas emissions rise accordingly, a trend likely to intensify with growing computational demand. Thus, sustainable or 'Green AI'—combining energy-aware algorithms, model/architecture optimization, and renewable-powered data centers—has become a priority.

Energy efficiency reduces greenhouse-gas emissions and pollution while lowering utility costs and supporting a more sustainable future as documented in Ref. [9]. In AI, energy-efficient models can reduce training and inference energy requirements and support deployment on resource-constrained devices such as smartphones and IoT sensors, as demonstrated in Ref. [10]. Table 1 summarizes only the empirically supported benefits documented in these referenced studies [9,10].

Despite substantial accuracy gains, recent AI progress has increased energy and water use (e.g., data-center cooling). While sustainability standards are emerging, industry estimates suggest such requirements could exceed 30 % of worldwide energy expenditures by 2030 [11]. Fig. 2 depicts CO_2-equivalent (CO_2e) emissions for selected ML model training runs. The largest emissions appear for GPT-3 (175B) and Gopher (280B), at 502 and 352 tCO_2e respectively; other models (e.g., OPT 175B, BLOOM 176B) are lower than "car use" and "air travel" comparators as illustrated.

Table 1
Summary of empirically supported benefits.

Aspect	Details
Environmental Impact	Reduces greenhouse-gas emissions and pollution linked to energy use, supporting measurable environmental sustainability outcomes [9].
AI Energy Efficiency	Lowers energy requirements for model training and inference and enables deployment on resource-constrained devices (e.g., smartphones, IoT sensors) as reported in Ref. [10].

1.2. Generative AI

Generative AI systems such as ChatGPT and DALL-E underscore sustainability concerns across both training and inference. A peer-reviewed synthesis by Verdecchia et al. [12] shows that, among 98 papers published since 2015, 49 focus on training while only 17 address inference. In contrast, external industry estimates for ChatGPT operations cite thousands of NVIDIA HGX A100 servers and hundreds of MWh per day [13], potentially exceeding GPT-3's reported 1287 MWh training energy in the underlying technical report [14]. Google's internal measurements for 2019–2021 further indicate that inference accounts for ~60 % of ML energy use, compared with ~40 % for training [1]. At the same time, BLOOM is reported to have lower inference energy than training, indicating that retraining cadence, workload mix, and deployment scale strongly influence stage shares. Fig. 3 summarizes the reported training energies for several representative models [14].

1.3. Green AI

Green AI refers to the intentional development and use of AI to reduce environmental burden—lowering carbon emissions and energy use—by optimizing algorithms and architectures, adopting efficient hardware, and utilizing renewable energy [15]. These efforts align with the United Nations Sustainable Development Goals (SDGs) and broader sustainability standards, reinforcing ethical responsibilities in AI innovation [16,17]. Approaches such as federated learning can also reduce

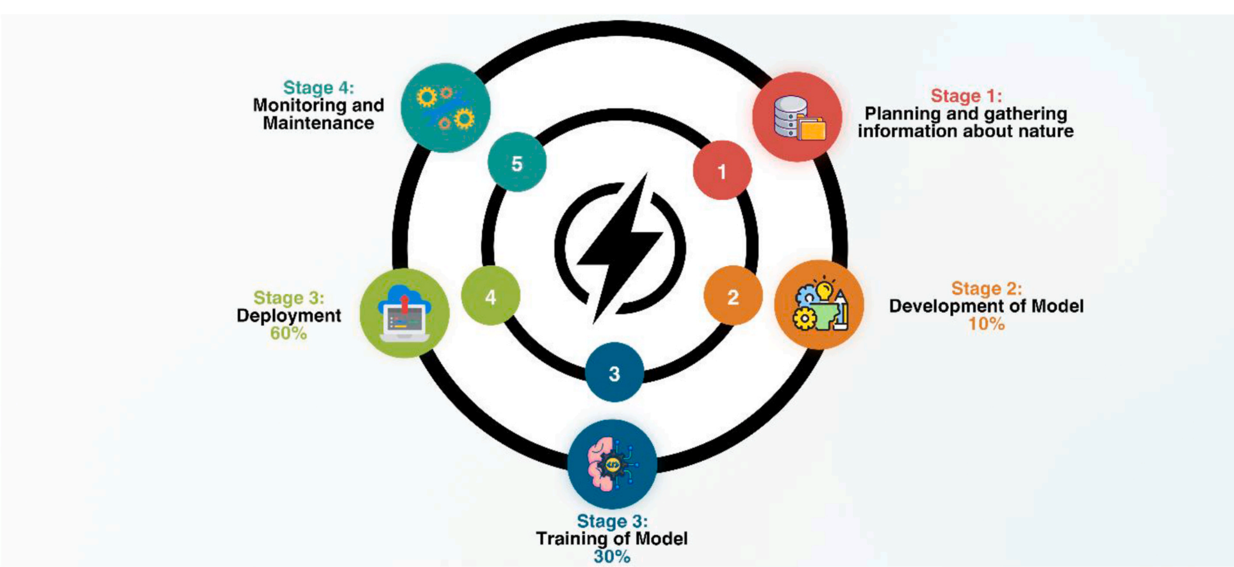

Fig. 1. Stage-wise energy in AI workflows. Deployment (inference) and training shares reflect measured values for 2019–2021 (~60 % inference, ~40 % training) [1]. "Planning" and "Development" are schematic (illustrative), as standardized measurements are not available. Actual shares vary with workload, system scale, siting/grid mix, cooling/PUE, and utilization [1,2].

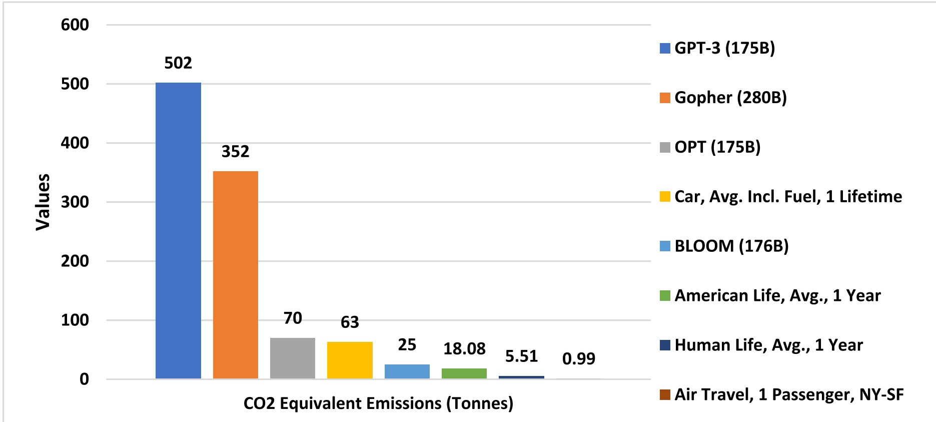

Fig. 2. CO_2 emissions for ML model training. CO_2e for each model is computed as $CO_2e\_model = E\_IT \times PUE \times CI\_region$, where E_IT is measured/reported training energy (or average power × training time), PUE accounts for facility overhead when available, and CI_region is the regional grid carbon intensity (g CO_2/kWh) for the model's reported site and time window. "Car use" and "Air travel" comparators convert distance to emissions via per-km or per-passenger-km factors and the distances shown in the labels; we indicate whether factors are lifecycle (LCA) or operational (tailpipe/TTW) and whether high-altitude forcing is included. Car lifecycle emissions use the European Environment Agency (EEA 2023) mean factor of ≈0.18 kg CO_2e/km (full LCA basis). Aviation emissions are based on ICAO (2023) operational CO_2 factors; no radiative forcing multiplier was disclosed in that source and is therefore marked "NR." Actual values vary by workload, site, and source; these comparisons provide order-of-magnitude context rather than a strict ranking and Table S1.

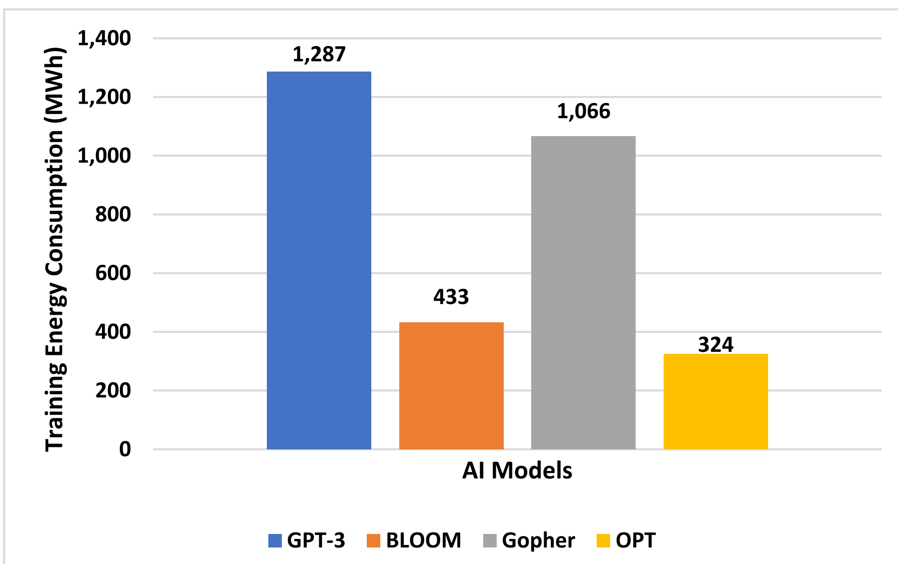

Fig. 3. Training energy consumption for various AI models (MWh) [14].

centralized training and related energy costs at scale. In this review, we organize Green AI along three interacting levels: (i) algorithmic techniques that reduce computational load (e.g., pruning, quantization, architecture search, and training-stage methods); (ii) hardware-level choices that improve performance per watt (e.g., GPUs, TPUs, FPGAs, neuromorphic devices); and (iii) infrastructure and data-centre measures that lower system-level energy and carbon intensity (e.g., efficient cooling, low-PUE design, siting, and renewable integration).

This systematic review identifies and evaluates Green AI methods that reduce energy use in AI systems. By surveying current approaches and developments in AI energy efficiency, we map how Green AI can shrink the carbon footprint of AI technologies and clarify pathways toward environmentally responsible AI. Within the algorithmic layer, we consider both inference-focused methods (such as compression and low-precision computing) and training-focused strategies (e.g., runtime layer freezing, model quantization, and early stopping), which can jointly reduce training energy while maintaining task performance. Our contributions are fourfold: (i) a systematic review of Green AI methods that reduce power consumption without compromising effectiveness; (ii) an assessment of integrating renewable energy sources into AI architectures, including case studies and real-world examples; (iii) an analysis of key barriers and trade-offs between performance and energy cost, with directions for future work; and (iv) a discussion of regulatory and ethical aspects to align environmental objectives with international sustainability standards.

Rising concerns about the environmental impact of AI systems and the need for sustainable growth underscore the importance of evaluating Green AI as a means to improve energy efficiency. This review therefore examines current energy-efficiency methods, integration of renewable energy, and broader sustainability practices in AI. To guide the investigation, the following research questions are central:

1. What are the most effective Green AI techniques currently being developed to reduce energy consumption in AI systems?
2. How can renewable energy sources be integrated into AI systems to make them more sustainable?

3. What are the real-world impacts of Green AI techniques on the energy consumption of AI systems in various industries?
4. How do the challenges of balancing performance and efficiency affect the adoption of Green AI in large-scale AI applications?
5. What are the regulatory and ethical considerations in the development of energy-efficient AI models?

The aim of this systematic review is to assess the range of Green AI approaches that reduce energy demand without compromising performance. Accordingly, the objectives are to:

1. To examine the current landscape of Green AI techniques and their effectiveness in reducing energy consumption in AI models.
2. To assess how renewable energy sources are being incorporated into AI research and development to mitigate the environmental impact.
3. To explore the balance between performance and energy efficiency, highlighting challenges faced by researchers and developers.
4. To review real-world case studies demonstrating the integration of Green AI and renewable energy solutions in AI systems across industries such as healthcare, autonomous systems, and data centers.
5. To provide recommendations for further research into Green AI and energy-efficient practices, including ethical considerations and regulatory compliance.

As AI becomes increasingly embedded in operational environments, energy-efficient design is essential. Techniques such as model trimming, quantization, hardware optimization, and the use of renewable energy offer promising pathways toward sustainable computation. This review highlights emerging innovations, key challenges, and future research directions.

2. Background

Traditional AI systems have significantly increased energy consumption, raising concerns about their environmental impact. Recent research emphasizes approaches that improve energy efficiency while maintaining or enhancing model performance. Reducing power demand is central to broader sustainability efforts, and AI can support this goal through smarter coordination with modern power systems.

Modern smart grids use bidirectional power flows and information exchange—supported by standards such as IEC 61850, IEEE 2030.5/SEP 2.0, and OpenADR—to balance supply and demand in real time. These mechanisms allow flexible loads and distributed resources to adjust operation based on price, reliability, or dispatch needs, enhancing grid efficiency and resilience.

Forecasting renewable energy generation is another critical component, and machine learning methods are widely used to manage variability in wind and solar supply. Representative studies report improvements such as RMSE reductions of ~30 % and MAE reductions of ~34 % for wind [18], MAE of 7.53 for Sotavento wind and 2,090, 959.30 for a Kaggle solar-radiation task [19], and lower MAE/RMSE for LSTM-based solar PV forecasting across year-long datasets [20]. These models leverage historical, operational, and physical-model data to support grid stability.

Green AI also enhances energy efficiency in built environments. Machine learning applied to sensor data in smart buildings [21,22] and cities [23,24] optimizes HVAC, lighting, and overall energy use, delivering sizeable reductions in consumption and carbon emissions [25,26]. The error metrics have been made as comparable as possible with some values of no reported (NR) being included, but much of the variation in the timing horizons of the data, their cadences of forecasting, the number of sites and the protocols used within a study to evaluate the performances prevent straightforward quantitative cross-comparisons of performance between studies.

Optimizing algorithmic performance provides clear energy benefits, as Green AI prioritizes methods that reduce computational and memory demands. Current approaches include sparse training [27], quantization [28,29], energy-aware pruning [30–33], and lower-precision arithmetic [34–36], all of which reduce training cost and inference load. Additional contributions include efficient feature-selection for domain adaptation without evolutionary algorithms (Garcia Castillo et al. [37]) and scalable, long-horizon time-series assessment using locality-sensitive hashing (Lourenco et al. [38]).

Energy efficiency is further improved through more capable hardware. GPUs vary significantly in FLOPS-per-watt: for example, the NVIDIA A100 (~400 W) offers ~2.5 × higher mixed-precision throughput than the V100 (~300 W), with MLPerf results showing ~1.8–2.2 × faster time-to-train on representative vision workloads [39]. Purpose-built accelerators such as TPUs [40] provide additional gains by aligning hardware with ML computational patterns [41,42]. Parallelization also reduces execution time; Anthony et al. [43] showed that increasing cores to 15 lowers runtime and emissions, though emissions rise sharply when runtime reductions fail to offset added power.

Edge computing provides another strategy by performing computation where data is generated, minimizing transmission to cloud or data centers and supporting the limited power budgets of IoT devices [44–46].

Data-center efficiency remains a critical determinant of AI's carbon footprint. Published CI_region values illustrate substantial regional differences, with approximately ~20 gCO_2e/kWh reported for low-carbon grids such as Switzerland and Norway and values exceeding 800 gCO_2e/kWh for high-carbon grids in Australia, South Africa, and certain U.S. regions, as reported in the cited dataset [47]. These CI values correspond to the dataset year and geographic breakdown provided in the referenced source [47], and the comparative framing in this section reflects those regional factors. To optimize performance, researchers have developed algorithms that balance server loads [48,49], regulate cooling, and manage resource allocation [50–52].

Efficient data-center design is therefore essential. Key practices include integrating renewable energy sources, adopting advanced cooling methods (liquid, immersion, free cooling), and using virtualization to reduce hardware footprint. Fig. 4 summarizes these principles, highlighting renewable power, thermal-efficient cooling, and virtualization as core enablers of sustainable, energy-aware data-center operation.

Data centers are a core element of AI infrastructure, providing the compute and storage needed for large-scale machine-learning workloads. Their electricity use is typically divided into three main components: IT equipment, cooling, and supporting infrastructure. Servers, storage, and networking equipment account for ~40–50 % of total energy consumption; cooling systems that maintain safe operating temperatures use ~30–40 %; and auxiliary systems (power supplies, security, lighting, etc.) consume ~10–30 % [47]. These shares are likely to evolve as AI workloads grow, driving further increases in overall data-center electricity demand. Fig. 5 illustrates a representative layout, highlighting key components such as uninterruptible power supplies (UPS), racks, cooling systems, fire protection, environmental controls, generators, security and access-control zones, and network operation centers (NOC), all of which must be coordinated to ensure safe and reliable operation [53].

Model optimization is equally important for reducing the energy needs of AI systems [54]. Two primary techniques dominate: pruning and quantization [55,56], often complemented by data augmentation. Model-size reduction can increase architectural flexibility—especially for convolutional networks [57]—while augmentation reduces the number of training cycles required, thereby saving compute resources [58]. Pruning removes parameters or neurons with minimal impact on outputs, shrinking model size and lowering training, inference, and storage costs [59]. Quantization reduces numerical precision, replacing floating-point representations with 16-bit or 8-bit integer formats [56,60]. This lowers memory requirements, accelerates computation, and shortens training and inference times, ultimately cutting energy use

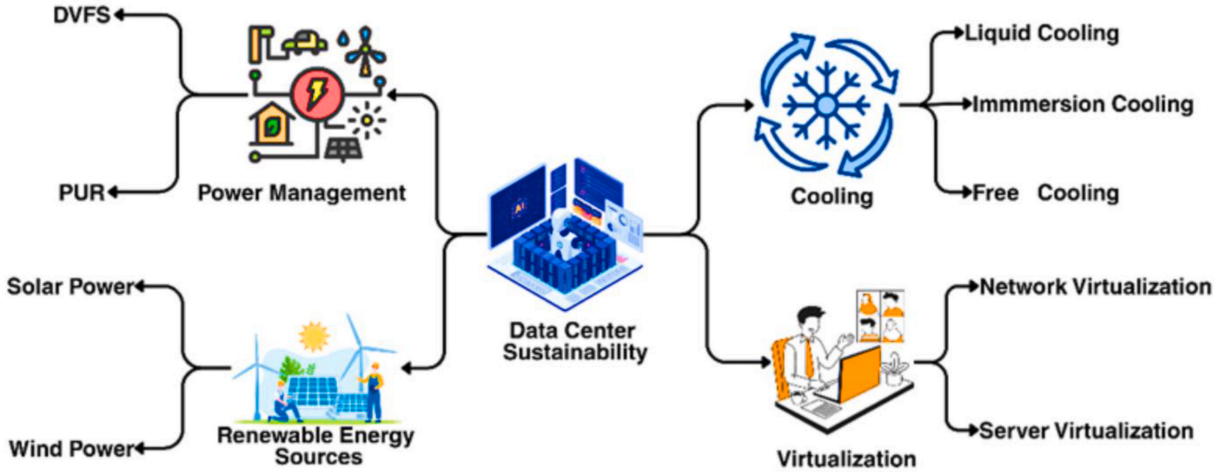

Fig. 4. The main measures to enhance the sustainability of data-centers. There are major categories, which are renewable energy integration (solar, wind), power management techniques (DVFS, PUE optimization), advanced cooling methods (liquid cooling, two-phase immersion cooling, free cooling), and virtualization (server and network virtualization). Achieved values of PUE are 1.05–1.07 when using two-phase immersion cooling and 1.1 when using direct-to-chip liquid cooling (versus industry average 1.56 when using air cooling), liquid/immersion systems achieve higher inlet temperatures (up to about 40 C) and facility-power saving in the order of 27 % and overall site-energy saving of up to 15.5 % over conventional air cooling (NR = not reported where applicable).

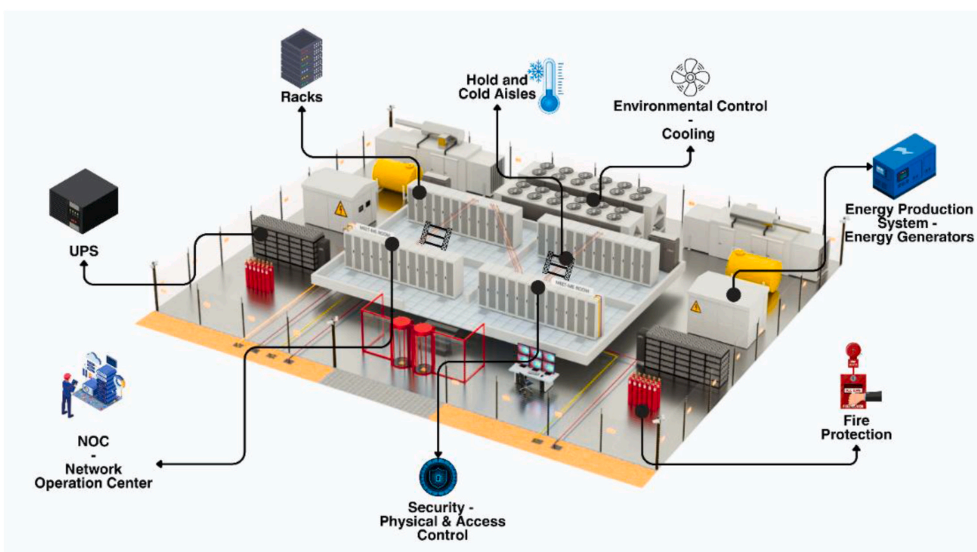

Fig. 5. Example of data center layout [53].

through reduced memory traffic and faster arithmetic operations. Quantization can be applied during training or only at deployment [60]; both post-training quantization and quantization-aware training aim to preserve accuracy while achieving substantial energy savings.

Foundational work prior to 2019 underpins these Green AI mechanisms. Early quantization and compression studies formalized precision–compute–accuracy trade-offs [32,35,56,60]; pruning and sparsity research culminated in the Lottery Ticket Hypothesis [59]; knowledge distillation provided a compression route exemplified by DistilBERT [61]; and efficiency-first architectures and search methods (e.g., MobileNetV2, NAS, ProxylessNAS) explicitly optimized operation count, memory access, and latency [62,63,64]. These antecedents form the methodological basis for the techniques evaluated in our recent evidence—lower-precision arithmetic, structured sparsity, student–teacher compression, and operation-aware design—without altering any quantitative results reported here.

The optimization of neural network training and inference to achieve energy efficiency has a history prior to current wave of Green AI publications. Even the initial analysis of quantization and sparsity-inducing regularization and pruning showed that extreme tradeoffs in memory, bandwidth, and arithmetic operations could be achieved with very few accuracy-losses-through-acquisition-like-lottery-ticket-like-pruning-techniques, with no sign of new 8-bit/4-bit quantization, structured and dynamic sparsity, and lottery-ticket-like-pruning patterns, directly inspired by the knowledge gained in the early days of growth hormone therapy.

3. Research methodology

The research dated back to a systematic review examining the Green AI methods leading to a decrease in energy in the AI systems. In the study, databases such as IEEE Xplore, Google Scholar, Springer, and ScienceDirect were used to find literature. Only studies published between 2017 and 2025 were targeted, with a possible date range of 5–7 years to capture current developments, in which only a limited number of papers group viewers aged above this. Regarding inclusion criteria, studies examining energy-efficient algorithms, hardware, and data centers, and optimization of AI models with quantitative information were emphasized.

3.1. Eligibility criteria

Included studies were peer-reviewed publications reporting quantitative energy or power measurements (or reproducible estimates) for at least one Green AI technique (model compression, quantization, pruning, efficient architectures, specialized hardware, data-center optimizations). Exclusions: non-English, non-peer-reviewed, qualitative only, no energy metric.

Inclusion Criteria:

1. Studies that focus on energy-efficient algorithms, such as pruning, quantization, and sparse training techniques.
2. Research on hardware solutions like Tensor Processing Units (TPUs), Field-Programmable Gate Arrays (FPGAs), and other specialized hardware that contributes to energy efficiency in AI systems.
3. Research related to energy-efficient data centers and infrastructure optimization strategies, such as cooling systems and power management.
4. Studies that address AI model optimization techniques aimed at reducing energy consumption without compromising performance.
5. Studies that provide quantitative data on energy or power measurements associated with AI workloads, regardless of the magnitude or direction of the measured effect.

Exclusion Criteria:

1. Studies that do not directly address energy consumption or fail to provide empirical or quantitative data on energy reduction.
2. Research that lacks a clear focus on energy-efficient techniques or does not offer measurable insights into energy efficiency in the context of AI systems.
3. Studies that are purely qualitative or theoretical, without empirical evidence related to energy use, theoretical frameworks for sustainability, or ethical considerations without energy efficiency data.
4. Research that does not specifically address AI's energy footprint, such as studies on unrelated aspects of AI applications in non-energy-focused domains.

3.2. Screening and data extraction

Two independent reviewers screened all records and extracted data and lastly a third senior reviewer was to resolve any disagreements. The entire screening (deduplication, title/abstract screening, full-text eligibility screening, data extraction) was done manually with the assistance of collaborative spreadsheets Microsoft Excel was used to manage reference management and export citation only. There was no systematic review software used.

Only peer-reviewed journal articles and referred conference papers were eligible for screening and inclusion in the evidence synthesis. Items from Google Scholar were included only if they were peer-reviewed. Industry white papers, vendor blogs/news posts, and other non-peer-reviewed materials were excluded from the screened study set and were not used for quantitative comparisons or study counts. Regulatory and standards could be cited for policy/definitions/background but were not part of the empirical synthesis. Where non-peer-reviewed sources are mentioned, they are cited solely for background. If any numerical values from such sources are shown for context, those values are moved out of the main text and placed in an appendix table, clearly flagged as "non-peer-reviewed," and excluded from all study counts, pooled summaries, and effect-size discussions. Accordingly, no quantitative comparisons in Section 4 (Results) or the narrative synthesis are derived from non-peer-reviewed sources; any contextual numbers appear only in Appendix Table S2 ("Contextual quantitative data from non-peer-reviewed sources") with explicit labeling.

3.3. Search strategy

The search strategy that will be specified in the paper is focused and relevant to these words: Green AI, energy-efficient AI algorithms, AI model optimization, and sustainable AI systems. Such keywords are very topical and correspond to the goals of creating the document, discussing the most important issues of the Green AI methods, optimization approaches, and sustainability. Grouping with such terms was probably a catch-all term to cover a broad range of relevant research, as in the references to techniques such as pruning [55], quantization [56], and data center efficiency [51].

The range of 5–10 years (2017–2025) can be explained by the explosive growth of the study of AI and Green AI. The sources of this era comprise 67.2 % of the overall amount, and 84 of them are located within the 2017–2025-time frame. There is an interesting bump in 2023, where 19 references were given to reflect the continued innovation of state-of-the-art in things like energy-efficient NLP models, DistilBERT [61], or hardware neuromorphic computing. Such emphasis on recent trends will make the research relevant to the current trends. The date range, however, may be a possibility to eliminate some of the lower studies, which formed the basis of these developments. Examples of all of these include Moore's law (1998) [65] and the theory of quantization, not as well represented with only 6 pre-2017 references, but with valuable historical context. Granting the date range for the past 10 years may give an insight into the development of Green AI in a wider scope without compromising on the recent nature work.

Fig. 6 outlines the search query process. It starts with defining a complex research string related to "Green AI" and energy efficiency. After executing the search, the results are assessed to determine if they meet the criteria. If yes, initial studies are reviewed; if not, the search string is updated, and the process is repeated until relevant results are obtained.

3.4. Selection process

The studies should be selected to make the process reliable and reduce bias. Two independent reviewers screened the studies based on their relevance, and in case of disagreement, their positions were resolved by consensus. Using two reviewers will increase the reliability of the study selection process since it eliminates individual biases, and the studies will meet the established eligibility criteria.

The choice of databases for the proposed Green AI research is relevant and suits the purpose of finding the appropriate research. The top source, IEEE Xplore, provided 29 references (23.2 %), primarily focusing on technical studies in the field of AI hardware, e.g., Tensor Processing Units (TPUs) [40], and algorithms, quantization techniques [56]. ScienceDirect (Elsevier) has 25 sources (20 %), and it provides access to energy and sustainability journals, so it is rather helpful in forecasting renewable energy [66]. The resource of Springer with 15 references (12 %) is another source of conference proceedings and machine learning research, and research on NAS [62]. Google Scholar has 16 sources (12.8 %) and guarantees that one will find coverage and that the newest work is easily available, such as the DistilBERT model [61], especially via arXiv preprints. The reference data shows that 69 references (55.2 %) are found on these leading platforms, indicating that high-quality peer-reviewed studies are emphasized.

Yet, the databases are inclusive, but since Google Scholar has been included, this may pose a possible variation in the quality of sources since it indexes peer-reviewed and non-peer-reviewed materials. A wider but not so uniform range of research could be the result. Also, the other category, referring to 32 sources (25.6 %), indicates the possibility of using other sources like industry reports [67]. However, their undocumented character restricts the level of transparency and makes this

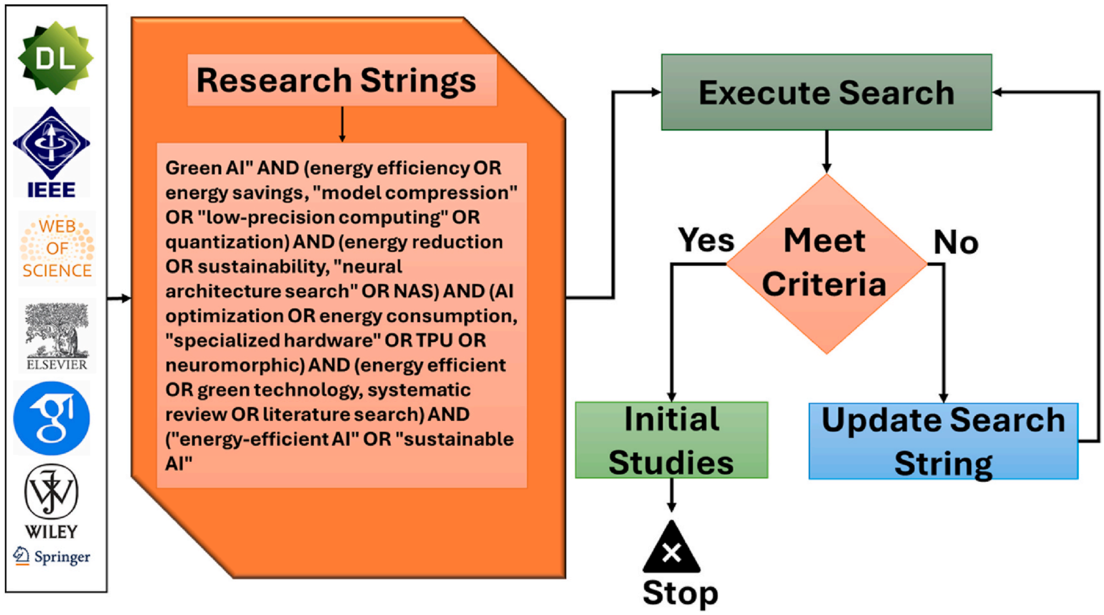

Fig. 6. PRISMA-style study-selection flow. Identification: records identified via databases (n = 683) and other sources (n = 0). Screening: records after deduplication (n = 494) screened; records excluded on title/abstract (n = 202). Eligibility: full-text articles assessed (n = 292); full-text articles excluded, with reasons (n = 165; see Appendix Table S3). Included: studies in qualitative synthesis (n = 131) and, where applicable, quantitative synthesis (n = 0). Counts reflect our screening log for the timeframe and sources defined in Section 3.

source hard to evaluate in terms of reliability and usefulness.

Fig. 7 illustrates the distributions of references by publication year in the field of Green AI. It shows a sudden influx, the peaks of which occur around 2022, and afterwards it reduces to a steady downward trend around 2025. The graph shows that there is an increase in the number of people who are interested in studying Green AI over the years.

Fig. 8 illustrates the distribution of publications in %age across different publishers. The chart demonstrates the %age of publications from various sources, including IEEE, Elsevier, arXiv, Wiley, ACM, MDPI, Springer, and others, such as blogs, reports, and web pages.

3.5. Data extraction

This document can quantify energy savings and test various techniques, AI models, and their results since the information collected in the studies and used as data points is very close to the research

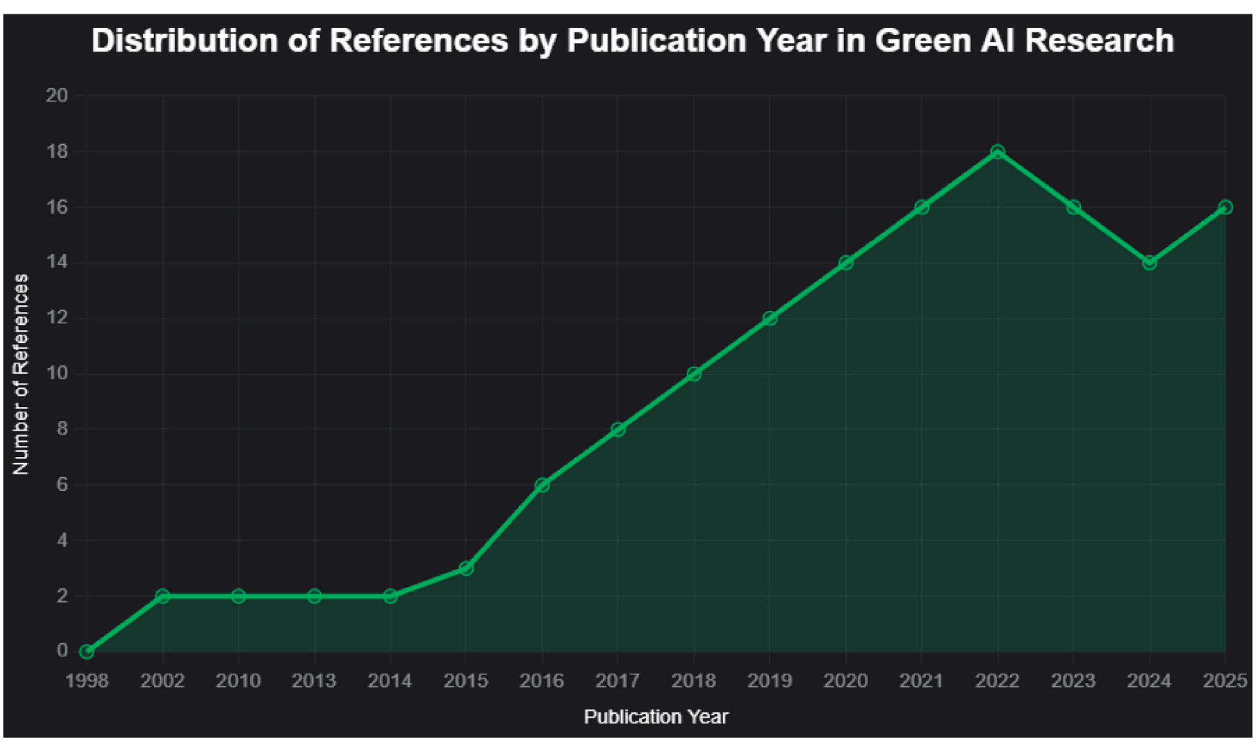

Fig. 7. Distribution of included studies by publication year. X-axis: Publication year (1998–2025). Y-axis: Number of included studies (n). Bar labels show each year's share as a %age of the total included studies (N = 131).

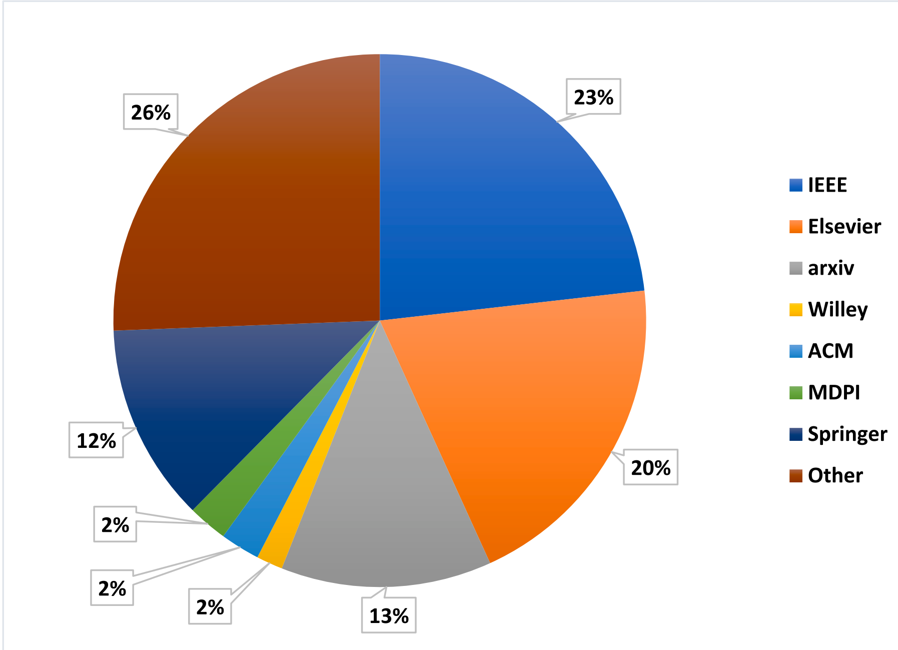

Fig. 8. Distribution of included studies by topical category. X-axis: Category. Y-axis: Number of studies included (n). Bar labels show each category's share as a %age of the total included studies (N = 131). Note: Categories may be non-exclusive for multi-topic papers; ages can therefore sum to more than 100 %.

questions. As an example, the document presents a 50 % reduction in energy consumption using quantization. It even proposes the benefits of such techniques, as pruning and Neural Architecture Search (NAS) [62, 63]. It also references such AI models as DistilBERT and MobileNetV2, and the document gives evidence of how the Green AI approach might allow for considerable energy savings without affecting application performance. The examples also effectively support the statements in the document, because there is a high probability of Green AI, and its performance in energy savings would have been better than that of other artificial intelligence systems that have traditionally been in existence.

We report per-study characteristics in Appendix Table S4 (domain/task, dataset/benchmark, and horizon, sample size, hardware/platform, energy metric and unit, measurement method, treatment of PUE and regional carbon intensity, baseline comparator, headline result as reported, and risk-of-bias rating).

3.6. Risk of bias

To determine the quality of the study, standard quality assessment tools were used, which is very important to ensure that the studies included in the research are of quality and the arguments concerning the energy savings and Green AI methods are supported with evidence. The risk of bias assessment will help reduce low-quality or biased research, which is paramount to the credibility of the findings. It does not mention specific tools, but the document is likely to rely on conventional instruments like the Cochrane Risk of Bias or AMSTAR, which are frequently applied when conducting systematic reviews.

Nevertheless, the absence of information on which exact tools or standards were applied to assess the bias lowers the transparency of the assessment. As an illustration, the document fails to explain the manner in which specific biases, e.g., publication bias or industry influence by companies like Google [1,67] are dealt with. Such possible biases may pose a threat to the outcome and the conclusion, especially when studying research funded by or directly associated with industry representatives. This level of transparency would also improve the quality of research by providing more detailed information about the tools and how this bias was addressed.

Assessment tool and procedure. We did not apply AMSTAR, ROBIS, or Cochrane risk-of-bias tools verbatim. Given the engineering/CS scope of the included studies, we used pre-specified rubric aligned with their core domains (eligibility; identification/selection; data collection/measurement; synthesis/findings) and tailored to Green-AI contexts. Each included study was rated on: (i) dataset representativeness and leakage controls; (ii) energy-metric validity (clear definition of power/energy, measurement method, treatment of PUE and regional carbon intensity); (iii) hardware/platform transparency; (iv) reproducibility (code/data availability, hyperparameter disclosure); and (v) baseline comparability. Ratings used a three-level scale—Low risk/Some concerns/High risk—and assessments were performed independently and reconciled by consensus among the authors; disagreements were resolved through discussion. Risk-of-bias ratings informed narrative synthesis and sensitivity commentary; they were not used to compute pooled effect sizes.

Inter-rater reliability and adjudication: Two authors independently rated all included studies (N = 131). We recorded initial disagreements at the domain level and resolved them through a structured consensus meeting: both reviewers re-checked the evidence against the rubric; when needed, pre-specified rules were applied (e.g., absence of PUE and/or regional carbon-intensity treatment ⇒ at least "Some concerns" in the energy-metric domain; non-comparable baseline settings—batch size/precision—⇒ elevate to "Some concerns" in baseline comparability). No third reviewer was required.

During the review process, disagreements were tracked qualitatively at the domain level and resolved using these pre-specified rules, but stable numerical tallies of initial disagreements per domain were not retained in the extraction sheet. As a result, we cannot report exact disagreement counts retrospectively without risking inaccuracy. We therefore report only the final reconciled ratings and explicitly acknowledge the absence of per-domain disagreement counts as a limitation of our risk-of-bias reporting.

Initial disagreements by domain were Energy-metric validity — NR; Hardware/platform transparency — NR; Reproducibility — NR; Baseline comparability — NR; Dataset representativeness — NR. We refrain from computing a single global κ due to heterogeneity and small per-domain cell sizes; instead, we disclose these per-domain counts and the final reconciled ratings in the text.

Illustrative adjudication examples:

a) Energy-metric validity: One reviewer marked Low risk based on average power reporting; the other marked Some concerns because PUE and regional carbon intensity were missing. Rule applied: missing PUE/CI ⇒ Some concerns.

b) Hardware/platform transparency: One reviewer marked Some concerns using vendor-only performance figures; the other marked High risk due to no measurement protocol/instrumentation details. Rule

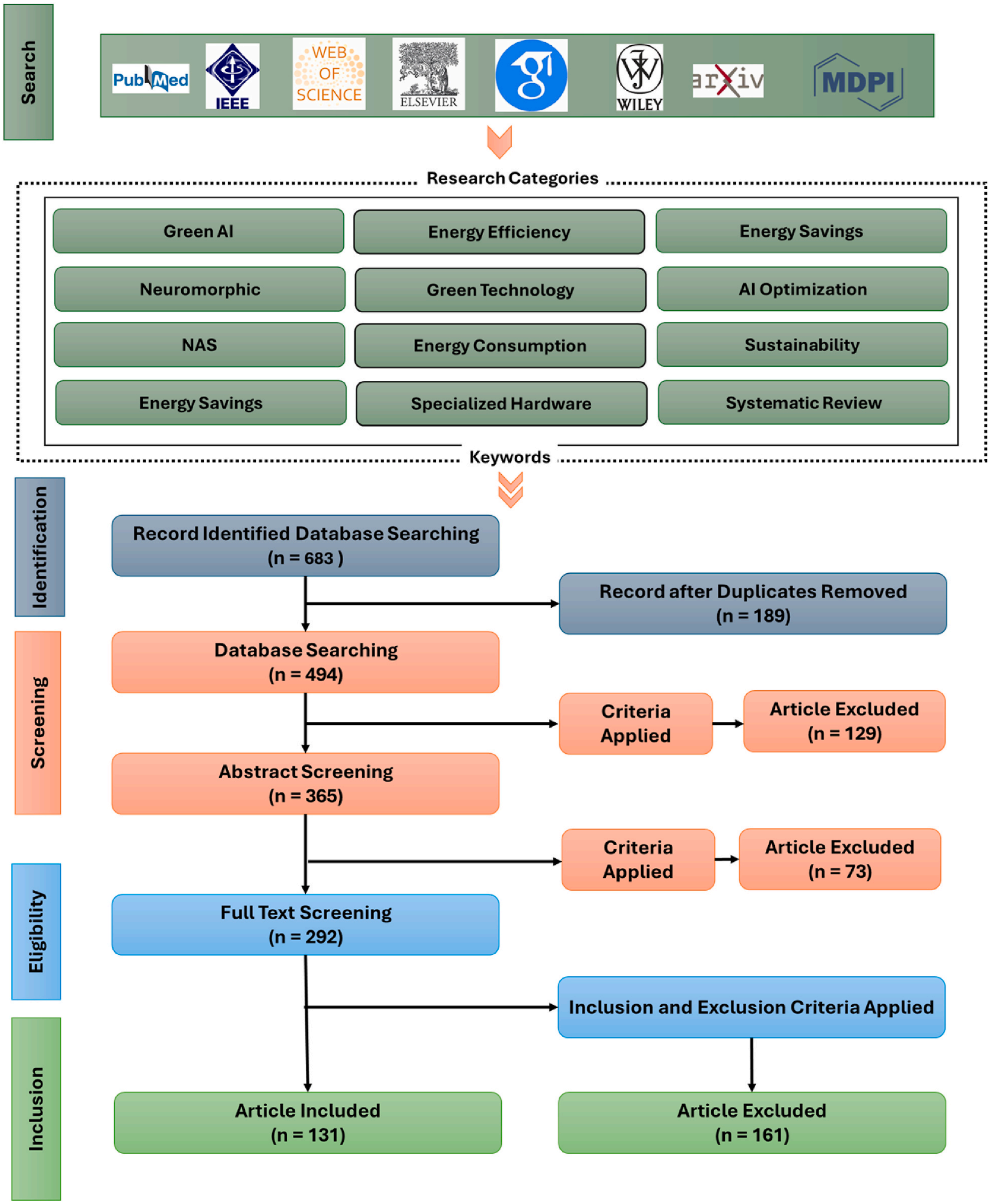

Fig. 9. PRISMA 2020 flow diagram for systematic review. A total of 683 records were identified through database searching. After removing 189 duplicates, 494 unique records remained and underwent title and abstract screening. Of these, 129 records were excluded, leaving 365 full-text articles to be assessed for eligibility. Following full-text review, a further 234 articles were excluded (with reasons provided in Appendix Table S3), resulting in 131 studies included in the final qualitative and quantitative synthesis. Detailed exclusion reasons and counts are provided in Appendix Table S3.

applied: vendor figures without methods ⇒ ≥ Some concerns; absent method/reproducibility ⇒ High risk.

c) Baseline comparability: One reviewer marked Low risk (baseline present); the other marked Some concerns because batch size and precision differed between baseline and treatment. Rule applied: non-comparable baselines elevate risk ⇒ Some concerns.

We present aggregated ratings (Low/Some concerns/High) per domain in Appendix Table S5 and use these to qualify the narrative synthesis. Two authors rated each study independently using the rubric in §3.5 and reconciled disagreements by consensus; due to heterogeneity and small per-domain cell sizes typical of engineering evidence, we report category counts rather than a single global κ statistic. Where specific fields were not provided by the original studies (e.g., missing PUE/CI), we mark them as "not reported" (NR) at the study level and apply the predetermined elevation rules during adjudication; this use of NR is distinct from inter-rater disagreement reporting, for which numeric counts are not available.

4. Results

4.1. Study selection

The inclusion of the PRISMA Flow Diagram, presented as Fig. 9, is a standard tool for presenting systematic reviews to demonstrate the process of study selection, including the number of studies identified, screened, assessed for eligibility, and included. The tool facilitates transparency, which corresponds to the methodology that utilizes two independent reviewers and the resolution through a consensus. Still, the PRISMA flow chart is absent, so there is less transparency, and it is not easy to ascertain the extent of the review. In the absence of this flowchart, the number of studies that were excluded, included, or screened cannot be exactly evaluated, and this is a restriction that does not help to determine the comprehensiveness and reliability of the review.

The reference data involved in the study includes a high proportion of recent studies (67.2 % of the publications conducted in 2017–2025). Still, due to the absence of a table, it is not possible to understand which studies were included in the review. Lack of such a table means that it cannot evaluate the number of works or compare the results of the work systematically, and this is a huge gap, especially because, as noted in these studies in the Discussion section, Green AI techniques indeed have considerable promise.

The reference data feature a heavy body of study in the last few years, with 84 studies in 2019–2025, including some of the current state-of-the-art, such as knowledge distillation of models such as DistilBERT [61] and energy-efficient designs such as MobileNetV2 [64]. This makes the studies more relevant, though it may overlook more initial concepts, such as previous work on the theory of quantization [32,35,56,60], which might be used to put it in context. The restriction to roll out on the most recent studies is quite fitting to mention the current trends; however, the omission of the studies that are not recent might reduce the amount of the overview and prevent understanding of the historical progression of Green AI methods. All in all, even though the use of recent studies is an advantage, the lack of the PRISMA flowchart and the Study Characteristics table is a substantial setback to the transparency of the reviewed analysis and the possibility of substantiating the empirical statements provided in the Discussion.

Across the cited studies [28,29,56,60], the reported 21–54 % energy savings figure refers to inference-stage energy reductions when applying INT8 post-training quantization (PTQ) relative to FP32 baselines. Representative evaluations in these sources use standard benchmarks and models—e.g., ResNet-50/MobileNetV2 on ImageNet for vision and BERT/BERT-based on GLUE or SQuAD for NLP—with energy measured as power × time on the hardware platforms reported in each study (with per-study GPU/accelerator/CPU details summarized in the "Hardware context" column of Table 2 and in Appendix Table S4). (Stage: inference | Models: ResNet-50, MobileNetV2 | Dataset: ImageNet | Study IDs: [28, 29]); (Stage: inference | Models: BERT/BERT-base | Dataset: GLUE or SQuAD | Study IDs: [56,60]). Where quantization-aware training (QAT) is employed in the same references, comparable or greater inference-stage reductions are reported at tighter accuracy deltas; we label PTQ vs. QAT explicitly next to each corresponding citation in the Results text. (Stage: inference | Technique: QAT vs. PTQ | Study IDs: [56, 60]).

4.2. Efficient algorithm

Energy-efficient AI systems cannot be made without the design of efficient algorithms that limit computational complexity and efficiently utilize available resources to reduce energy costs. Optimizing the structure of the data, simplification of mathematical operations, and use of sparse representation to reduce the number of communicated sets are among the key strategies to gain an algorithmic efficiency [68,69]. As an example, the effective neural architecture search (NAS) [62] may automate the process of designing the neural-network architectures by simultaneously optimizing the two values: the performance of the model and the expendable resources, resulting in a substantial decrease in energy consumption. In addition, approximate computing [70] sees the intentional trading-off of accuracy in pursuit of efficiency, a technique that can be very effective when 100 % correctness is not critical. Other potentially considerable energy savings can be achieved using techniques like probabilistic computing and adaptive precision to decrease the number and precision of operations needed on a task-by-task basis.

Energy efficiency is also increased using the transfer learning technique [71] to transfer models to other similar tasks, so a model does not need to be trained as extensively. This is particularly useful in areas where natural-language processing/computer vision takes place, where it is usually expected to train end-to-end models that need vast quantities of data and high computing power. In short, all these sustainable AI methods, such as model optimization and the design of efficient algorithms, should at least be considered vital ways of minimizing the environmental cost of AI technologies. Techniques such as pruning and quantization can simplify models, and the efficient design of algorithms enables the use of minimal resources, which contribute to the possible creation of energy-efficient AI systems and the achievement of overarching sustainability objectives.

4.3. Green in AI

The technological applications of AIs hold great potential when

Table 2
Case studies of Green AI implementations.

Hardware/ Technique	Model/Workload	Energy reduction vs. baseline	Throughput/Latency details	Batching used?	Latency constraints considered?
FPGA (Xilinx VU9P/ Alveo U280)	Llama-2-7B inference (per token)	8.25–12.75 × lower energy per token vs. RTX 3090 GPU/CPU	0.53 × throughput of RTX 3090; per-token latency 1.8–2.4 × higher than GPU	No (batch size = 1)	No (optimized for throughput; latency not constrained)
FPGA (High-Level Synthesis flow)	Llama-2-7B/13B	Up to 10 × energy efficiency vs. GPU	Throughput-focused design; measured at batch size 1; end-to-end latency NR	No	Partially (real-time token generation targeted but no hard constraint)

implemented in sustainability operations in various fields. However, there must also be careful consideration that the computational energy involvement of these systems themselves should not detract from the goals they are supposed to promote. Realizing positive environmental impact with the help of AI requires a strategy of developing so-called green AI, which constrains energy consumption without imposing an actual penalty on performance. The most prominent question is the one usually posed in terms of the conflict between the red AI and green AI, i.e., the trade-off between the computational capabilities and energy costs [72]. Studies conducted in the past show that most of the noticeable AI articles prefer precision to efficiency, where up to 90 % of the studies revolve around improving the model instead of streamlining the resources [15]. Moreover, training larger-scale ML models has increased over time by a factor of more than 2 at a rate of more than 3.4 months since 2012, and that pace of growth surpasses that of Moore's Law [65]. Although the resulting footprint of the energy of ML algorithms cannot be cast as unavoidable, the problem is one to take seriously and never at the expense of scientific advancements.

In turn, green AI is based on the need to achieve the desired performance results under conditions where either the costs of computation are reduced or maintained at the same level. The body of literature on this subject reveals a variety of design solutions, including software optimizations and hardware interventions capable of reducing the environmental burden of AI systems. All these suggestions come together as the most logical way of achieving the goal of making AI positively enable sustainability without excessive energy use. Green AI in addressing the energy consumption and environmental impact of AI. It proposes a framework for understanding, implementing, and advancing sustainable AI practices, including model optimization, efficient algorithms, energy-efficient hardware alternatives, and data center design strategies. The goal is to balance AI's performance and resource efficiency with global sustainability goals [73].

4.4. Green hardware alternatives

Within the modern endeavor to enhance artificial intelligence sustainability, the focus on energy-efficient hardware is already gaining new scholarly interest. Two major alternatives to regular graphics processing units have sprung forward, namely Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs) [74]. The current discussion explains the benefits and situations of these two green technologies with a case study that shows the successful use of the technologies. We now discuss neuromorphic and quantum computing in separate subsections—see §4.5 and §4.6—to avoid conceptual overlap with conventional accelerators. The first analysis will deal with compromises between the increase in model parameters and the related jumps in computational and environmental costs. Then, the new research on compression, pruning, and quantization through techniques is discussed as one of the responses to such issues.

GPU history has been a significant drive toward the development and expansion of more elaborate Natural Language Processing (NLP) and Computer Vision (CV) models. These constantly growing models can be described as having exponentially growing parameters but again have proven significant in identifying finer patterns in data and carrying out an extensive range of tasks. In turn, such multiplication of parameters leads to an increase in storage needs, prolongation of training, increased capital investments, and carbon footprint. However, empirically, there have been trends in both NLP and CVs where the number of parameters increases, not by powers of ten, but by powers of a billion, and the corresponding computation requirements multiply in commensurate complexity. These dynamics make it likely that specialized high-performance hardware is needed to train these large models, which in turn increases energy use. To address this, the methods of pruning, compression, and quantization have been heavily researched, all to reduce the computational requirements needed to run a model and reduce the model size and environmental cost with minor workload performance tradeoffs.

Deep learning has continued to grow, with larger and more intricate models, a scenario that puts the need for methodologies that mitigate demands on the computing system, as well as the model weight, to the forefront. The necessity goes beyond decreasing the dependence on expensive high-performance graphics processors; it is a purpose to simplify training, reduce memory usage, and allow deployment of the determined model in devices within the limit of resources at the edges. Here, several methods have been identified to be particularly useful: model compression, pruning, and quantization. Compression attempts to reduce the physical size of a model with a significant portion of retained performance, and knowledge distillation is a notable example. An example of such transfer learning models is DistilBERT, which reveals that a student model can keep 97 % of the accuracy of its larger, more capacious counterpart BERT, at a compression ratio of 40 %. Pruning discards unnecessary parameters or weights and thus reduces size and computational requirements; methods like magnitude-based pruning, where weights of low magnitude are removed, and structural pruning, where full neurons or layers are removed, have been up-and-coming, reinforced by the Lottery Ticket Hypothesis [59]. Quantization has also been popularized by its capability to reduce model size and computational overhead by limiting model weights to lower precisions, which is usually done by down-casting 32-bit floating-point real numbers to 16-bit integer numbers, and sometimes further, resulting in relatively small performance penalties [56,75].

Briefly, a jointly considered use of compression, pruning, and quantization has vastly reduced the reliance on high-end GPUs and allowed researchers with limited computational devices to access state-of-the-art systems and, accordingly, reduce energy costs and environmental consequences. The subsequent model size decreases have also extended the scope of robust architectures to edge devices, which include telephones and components of the Internet of Things, thus shifting deep learning capabilities into more realistic situations [76].

Recent large-scale accelerator evaluations now report explicit performance/Watt gains: Google's TPU v4 improves performance/Watt by $\sim$2.7 $\times$ over TPU v3 and, at comparable system scales, is reported as 1.2–1.7 $\times$ faster while using 1.3–1.9 $\times$ less power than Nvidia A100 on representative MLPerf workloads [40]. These results illustrate an emerging algorithm–hardware co-design pattern in Green AI: model architectures, numerical precision (e.g., BF16 and INT8), and accelerator characteristics are tuned jointly to maximize performance per watt. Energy-efficient transformer variants and LLM inference pipelines increasingly combine pruning, quantization, and attention/sequence optimizations with accelerator-specific kernels, treating hardware and model design as a coupled optimization problem rather than independent choices.

Standardized cross-vendor power measurements are available: MLPerf Inference v4.0 included $\sim$900 "Power" results across datacenter and edge systems, enabling transparent, apples-to-apples efficiency comparisons for LLMs, vision, recommendation, and other tasks; MLPerf Power (2024) formalizes the methodology for energy-efficiency reporting from microwatts to megawatts [77,78].

To maintain contemporaneity (2024–2025), we include a representative, measured "Power" example from MLPerf Inference v4.1 (Llama-2-70B, Offline). A single NVIDIA H100-SXM achieved $\sim$24,525 tokens/s at $\sim$700 W, while an AMD Instinct MI300X achieved $\sim$23,515 tokens/s at $\sim$750 W; this corresponds to $\sim$35.0 vs. $\sim$31.4 tokens/s/W, respectively (computed from the reported throughput and measured accelerator power). As with all MLPerf "Power" submissions, methodology and data collection follow the MLPerf Power rules; results are representative rather than universal and depend on software stack, precision, and system configuration [77,78].

FPGAs can deliver substantial energy reductions for transformer/LLM inference: an HLS-based Llama-2 design on a Xilinx VU9P reported 8.25 $\times$ reduction in energy per token vs. an RTX 3090 GPU (and 12.75 $\times$ vs. CPU) while maintaining 0.53 $\times$ of the GPU's throughput [79].

4.5. Neuromorphic computing (event-driven efficiency)

Neuromorphic computing (spiking/event-driven). In suitable workloads, neuromorphic platforms report high performance-per-watt relative to conventional pipelines, consistent with their event-driven operation and reduced data movement. Neuromorphic/event-driven hardware shows very high efficiency in suitable workloads: Intel's Loihi-2 achieved ~103.94 GOP/s/W at ~1.55 W in sensor-fusion tasks, and event-camera–centric edge platforms have demonstrated end-to-end perception-to-control power under ~50 mW with millisecond-scale latency [80].

Efficiency Loihi 2 and SpiNNaker 2 neuromorphic systems can achieve especially large efficiency improvements (1001,000 times less energy per inference than with standard GPUs) on spiking neural network (SNN) workloads that can use the temporal sparsity. Reports mainly involve real sensor loads like DVS gesture recognition, keyword spotting on streaming audio, radar/LiDAR event streams) although also some benchmark performance uses synthetic data converted CIFAR-10/ImageNet, N-MNIST. In almost every mentioned study, the accuracy, or task-performance metrics are provided together with the energy values (in most cases, these metrics are equivalent or within 15 per cent of ANN baselines following SNN conversion/training.

4.6. Quantum computing: algorithmic speedups and energy implications

Quantum computing (algorithmic speedups; energy implications are context-dependent). Quantum devices aim for asymptotic speedups for specific problem classes rather than circuit-level energy reductions. Any net energy benefit depends on wall-plug power, runtime (including error-correction), and problem suitability; accordingly, we refrain from asserting intrinsic energy savings and treat quantum as an emerging computational paradigm rather than a drop-in "green hardware" alternative.

4.7. Case studies in green AI

The examples of successful experience in the implementation of the principles of green AI provided in the frameworks of natural language processing (NLP) and computer vision (CV) can serve as case studies that prove it is possible to become energy efficient while maintaining high performance in all AI-related applications.

Memory/Energy-Efficient NLP Models: NLP is an expensive process, and beyond training, it consumes a lot of energy in the deployment environment. Recent advancements have therefore looked at energy-efficient NLP to reduce the environmental impact. The most noticeable one is the DistilBERT model, which Hugging Face presented. DistilBERT is pre-trained using knowledge distillation, where a smaller "student" model is trained to mimic a larger "teacher" model's behavior. DistilBERT retains ~97 % of BERT's task performance, is ~60 % faster at inference, and has ~40 % fewer parameters [61]. DistilBERT is a smaller and faster implementation of BERT (Bidirectional Encoder Representations from Transformers) [81]. Other energy savings have been achieved through methods such as pruning unnecessary parameters and quantization, which reduces the numerical precision of model weights. These strategies can reduce model size and computational cost while maintaining competitive accuracy in many settings. According to a study [82], incorporating such techniques on transformers such as GPT-3 reports energy consumption reductions of up to half. According to a study [82], incorporating such techniques on transformers such as GPT-3 reports energy consumption reductions of up to half.

Computer Vision on Resource-Constrained Devices: Computer vision tasks like image recognition and object detection are inherently high-energy demanding because they involve large amounts of data and complex computations. To address this, new methods are being developed that enhance efficiency. MobileNetV2 architecture [64] is an illustrative example that incorporates depthwise-separable convolutions to minimize the number of operations relative to conventional networks. This architecture supports low-power inference on mobile and edge devices while maintaining high accuracy. Reliable Neural Architecture Search (NAS) algorithms are another avenue for improvement. The effectiveness of NAS algorithms, such as ProxylessNAS [63], involves optimizing neural architectures with explicit attention to hardware limitations, performance, and energy efficiency. The resulting models reduce operations and on-device compute cost while maintaining competitive accuracy.

Energy-Efficient Robotics: Energy-efficient robots are essential to increase operating time and reliability. Autonomous navigation stacks increasingly optimize path planning, motion execution, and duty-cycling to improve energy efficiency and robust operation [83]. These optimizations are significant in field robots operating in remote environments or resource-constrained deployments, especially for long-duration missions where minimizing energy consumption is critical.

Healthcare Applications: In the medical field, AI systems also benefit from environmentally conscious design. Medical imaging applications for disease detection and diagnosis can be designed to be more efficient. Lightweight convolutional neural networks (CNNs) for detecting abnormalities in X-ray and MRI scans [84] are amenable to deployment on resource-constrained medical equipment, reducing device-level power draw during inference while supporting clinical utility. Using such combined advancements, AI in healthcare not only improves patient outcomes but can also reduce environmental impact. Table 2 represents the cases studies.

4.8. Design principles

With traditional air-conditioning systems being very energy demanding, the minimization of data-center energy usage is increasingly being bolstered using advanced cooling technologies [85].

The alternative methods, namely, liquid cooling, immersion cooling, and free cooling, are much more effective thermally. Liquid cooling will circulate a chillant directly onto the components producing heat, and it is more efficient than the standard air-based cooling systems [86]. This principle is further pushed through by immersion cooling, where servers in an electrically non-conductive but thermally conductive liquid are submerged, allowing for an even distribution of heat to the liquid [87,88]. Free cooling allows free utilization of environmental sources of air and water to provide cooling to server housings, because mechanical systems are minimally used, which reduces energy requirements [89]. It has been demonstrated empirically that the minimal power techniques allow cutting off energy consumption in data-centers up to 50 %, which in turn enhances the total energy efficiency [90]. The ideal cooling solution also depends on geographical factors (ambient temperatures and access to water) in the relevant geographical areas; cooling immersion would help when geographic temperatures are higher, and liquid cooling would be better where water is limited.

It should also be noted that this is critical to the optimization of power management. Techniques like the Dynamic Voltage and Frequency Scaling (DVFS) manage voltage and frequency regarding the system workload, and as such, reduce power consumption when the system is under-utilized [91]. Power Usage Effectiveness (PUE) is designed to give a quantitative measurement of the ratio of facility-infrastructure energy to IT-equipment energy, and its primary aim is energy waste reduction [92]. Efficiency is also dependent on the location options and architectural design. Having data centers in cooler locations will mean less artificial cooling is needed, and being in an area with access to renewable energy and stable grid infrastructure will help in being sustainable. Design intent has been shown to deliver significant energy savings through purposeful design, such as optimizing server rack layouts to enhance airflow and integrate natural-cooling-related elements [93].

As illustrated in the present study, the concept of implementing virtualization is a critical move towards improving energy rates at an organizational data center [94]. Virtualization has the effect of concentrating a large number of virtual servers on a single physical server, reducing energy demand and cooling needs, and hence requiring fewer physical machines. Additionally, technology maximizes the usage of the network resources, thus minimizing the use of the physical networking devices and thereby cutting down the energy consumption even further. The combination of increased efficiency in resource utilization, as well as reduced physical presence of hardware, essentially promotes energy efficiency.

The findings provide an urgent insight into what AI practitioners and policymakers must know in a managerial sense. By pursuing their Green AI initiatives, organizations will help reduce their operational costs by the same amount of energy savings, while supporting their broader environmental sustainability goals. The planned framework will help AI project managers weigh model performance and its consequences on the environment, especially in resource-limited situations. In the case of government agencies, facilitating Green AI adoption may be approached through policy interventions, such as tax breaks on energy-saving hardware and the obligation of AI frameworks to declare their carbon footprints. Also, the investments of the government into research on AI possibilities that are persistent and not harmful to the environment would result in the significantly faster development of the sphere and in further innovations that aim at making AI systems less harmful to the environment.

4.9. Applications for energy-efficient AI

Sustainable and cost-effective AI solutions are transforming the industry and are applicable in different sectors. AI models are optimized to run on smartphones and IoT devices in edge computing, so that they can work in real-time on low-power systems with short battery life. Cloud computing and server optimization enable better use of energy in data centers and augment efficiencies in AI training and the use of AI inference. Real-time decision-making and long work durations are some of the applications of energy-efficient AI in autonomous systems, such as self-driving vehicles and drones. The applications of wearable devices in healthcare range from meeting the requirements of health metrics monitoring, with the help of low-power AI models, to improvements in medical diagnostics using medical imaging systems that can produce less energy output in diagnosis.

Energy-efficient AI is also applied in a more fundamental sphere of center administration and enhancement, where energy misappropriation prevention, power theft detection, and adjustable load adjustment are some of the use cases sustained by start-ups, for example, eFlex. Companies such as Resilient Entanglement are advancing the aspect of smart grid and demand response systems, which include. Still, they are not limited to automated demand response and real-time grid monitoring. The innovators include Quadrical AI and Ratio Energy, as they offer predictive maintenance, energy demand forecasting, and storage management, with transformer health monitoring and solar panel analysis. Moreover, startups in areas of carbon emission reduction, energy trading, renewable energy integration, grid security, and consumer energy management, among others, include Ecodoho, Suena, EPAC Energy, Psymetis, and Wenerate, indicating a blossoming ecosystem of energy-efficient AI solutions.

4.10. Regulation and green AI

The increasing ease in obtaining tools to estimate the energy use in an AI system makes it imperative to optimize the carbon footprint of such systems. Nevertheless, the second crucial step is not limited to measuring but is followed by the implementation of regulations and the encouragement of Green AI. This shift symbolizes a shift towards the need to evaluate the environmental impact of AI in the process of managing and minimizing it. By setting standards and principles of energy efficiency in the development of AI, it is possible to guarantee the development of AI innovations that help not only transform the technological capabilities but also contribute to their compatibility with sustainability objectives. In a documented case study, adopting four best practices (efficient ML hardware, model choices, low-PUE facilities, and siting/24 × 7 carbon-free energy) yielded ~83 × lower training energy and ~747 × lower gross CO_2e relative to a 2017 baseline; the oft-quoted '~100 × /~1000 × ' figures are order-of-magnitude summaries rather than additional measured values [1,67]. We follow [1] and compute CO_2e as energy × PUE × regional carbon intensity. This includes optimization of efficient machine learning (ML) model architectures oriented towards minimizing computational needs over multi-purpose models [73,95–97], adoption of machine learning (ML) training optimized processors and ML-specific systems like TPUs, neuromorphic chips and vision processing units (VPUs) where training offers a large energy savings over traditional GPUs [98], migration to data centers provided by cloud providers to leverage their energy efficiency [99] and optimization of map selections to cloud providers which can exploit renewable energy resources to multiply. In the EU, these disclosure and documentation duties are anchored by the AI Act's obligations for providers of general-purpose AI models (Article 53(1)(a) and (b)), which require technical documentation (Annex XI) and downstream information (Annex XII); Article 53(5) empowers the Commission to specify "measurement and calculation methodologies" to ensure comparable, verifiable documentation, and Article 112(6) mandates Commission reporting on standardisation deliverables for the "energy-efficient development of GPAI" [26].

4.10.1. Regulatory frameworks and measurement standards

Beyond AI-specific instruments, compliance also depends on sustainability and energy-sector frameworks that set the operational conditions under which AI is acceptable. In practice, this intersects with: (i) oranizational carbon-footprint reporting and transparency across the AI lifecycle [100,101], including prior calls for declaring AI systems' carbon impacts; (ii) facility-level efficiency metrics and controls (e.g., PUE, DVFS, virtualization, siting/cooling design) that govern data-center energy performance [91–94]; and (iii) procurement and location strategies (e.g., 24/7 clean-power matching) that align computation with low-carbon supply [102]. Taken together with disclosure duties in the EU AI Act [26], these instruments specify when and how AI operations meet sustainability expectations: report what the system consumes and emits, run it in efficient facilities, and source power to minimize carbon intensity. Where facility metrics are cited, we reference established measurement standards, notably ISO/IEC 30134-2 (PUE) and EN 50600-4-2 for data-centre KPIs; best-practice siting and operational targets are additionally reflected in the EU Code of Conduct on Data Centers – Best Practice Guidelines (2024) (Joint Research Centre). To help practitioners navigate this landscape, we summarize the main frameworks and their roles in energy/CO_2e reporting for AI workloads in Table 5.

Together, these instruments frame how energy use and CO_2e should be documented for AI systems: the EU AI Act defines AI-specific documentation and transparency duties, ISO/IEC 30134-2 and EN 50600-4-2 structure data-centre efficiency metrics, and ISO 14064-1/GHG Protocol anchor greenhouse-gas accounting for AI services.

On the one hand, the contribution of AI researchers and companies is vital. Still, it is necessary to present some regulations that will foster the development of AI responsibly and sustainably. In line with this, the Ethics Guidelines for Trustworthy AI published by the High-Level Expert Group on AI in the EU in 2019 stated that systems based on AI must meet the needs of society and the environment in terms of well-being [100]. Acting on these recommendations, the EU subsequently issued a second report that promoted research in AI solutions capable of addressing the various challenges facing the world, including the Sustainable Development Goals (SDGs) of the United Nations, which may be fulfilled with the help of AI [101]. Companies and researchers are advised to respond

Table 3
Reported efficiency/accuracy trade-offs (examples). (Metrics are exactly as reported in sources; units vary by study. NR = not reported.)

Model/Variant (Task)	Benchmark	Params/FLOPs (as reported)	Hardware context (as reported: accelerator/node count/batch size/precision)	Energy metric (unit, as reported)	Accuracy/Δ vs. baseline	Note/ Source
INT8 post-training quantization (PTQ) vs. FP32 (Vision/ NLP)	ImageNet; GLUE/ SQuAD	NR (precision reduced FP32→INT8; architecture unchanged)	Varies by study; accelerator model/node count/batch size often NR; precision = INT8 (PTQ) vs. FP32	Energy per inference ↓ 21–54 % (median ~40 %) across reviewed studies; power × time, Wh or mJ per sample	Accuracy drop 0–2 % typical; up to 5 % in worst cases (PTQ); labeled PTQ vs. QAT in text	[28,29, 56,60]
DistilBERT vs. BERT-base (NLP)	GLUE/ SQuAD	~40 % fewer params (DistilBERT) — FLOPs NR	NR/NR/NR/NR	NR (no energy reported)	~97 % of BERT, ~60 % faster inference	[61]
Llama-2 inference on FPGA vs. GPU (LLM)	Llama-2	NR	Xilinx VU9P/1 node/NR/NR; NVIDIA RTX 3090/1 node/ NR/NR	Energy per token ↓8.25 × vs. RTX 3090 (GPU); ↓12.75 × vs. CPU	Throughput: 0.53 × of RTX 3090 (accuracy NR)	[79]
GPT-3 (175B) training & inference	–	NR	NR/NR/NR/NR	Training ≈ 1287 MWh; inference energy reported as higher in sustained operation	NR	[14]

NR = not reported in the cited source. Quantitative entries in Table 3 reflect peer-reviewed evidence only and are aligned with Section 4 annotations (stage/model/dataset/ID).
Scope note: Where studies report J/inference or Wh/token, we retain those units; when energy or FLOPs are not provided in the source, we label NR and avoid cross-paper normalization due to hardware, batch-size, and protocol differences.

to questions about the possible harm of AI systems on the environment, as well as the systems to gauge and minimize the damage, and the mitigation against harm to the environment throughout the lifespan of the AI system, to be able to analyze its use of resources and energy consumption critically. The 2020 White Paper on AI by the European Commission suggested that AI was both capable of contributing to the resolution of climate change (green-by AI) and that it could be created sustainably (green-in AI), which would ensure its impact on the environment in terms of protection to the whole value chain, including data collection and algorithm implementation [25]. In March 2024, the AI Act was adopted, which creates a reporting requirement on the energy efficiency of high-risk AI systems, especially the new generative AI models, and encourages voluntary codes of conduct on non-high-risk AI systems that apply sustainability principles [26]. Through this regulation, there is an effort to promote responsible innovation of AI, but environmental sustainability and the protection of fundamental rights are also safeguards. In the present manuscript, we therefore map each policy recommendation below to specific legal or standards anchors (EU AI Act articles/annexes; ISO/IEC and EN standards; and EU Code of Conduct guidance) to enable verification.

Nevertheless, the governance of AI sustainability is yet to emerge. In the recent Executive Order on AI in the USA (October 2023), green algorithms are not mandated for development, but it briefly refers to the creation of sustainable development goals [15]. Correspondingly, the Chinese policies are also not regulating AI development based on the notion of its sustainability. The ICT sector contributes 3.9 % of overall GHG emissions [103], with a large portion due to AI (particularly large-scale generative models). The resulting environmental issue poses a big challenge. With this heavy burden not being sustainable, integration of sustainability as an AI-algorithms regulatory aspect, as proposed in the AI Act, is essential. Scientists and engineers can also start by including energy consumption measures alongside conventional accuracy or precision measures to increase awareness that the development of AI should be conscious of energy consumption [104].

4.10.2. Policy actions and ethical guardrails

Policy Recommendations—Actionable Checklist:

a. Standardize disclosures for high-risk AI: report device-level energy for training/inference, data-center PUE where applicable, and regional carbon intensity (EU AI Act Article 53(1)(a) & Annex XI; Article 53(5) on measurement methodologies; complements Annex XII information duties to downstream providers [26]; for facility metrics: ISO/IEC 30134-2 (PUE) and EN 50600-4-2).
b. Prefer low-carbon siting and 24/7 clean-power procurement: document location choices and grid-mix assumptions (EU Code of Conduct for Data Centers – Best Practice Guidelines, JRC 2024; aligns with lifecycle reporting frameworks in ISO 14064-1 and GHG Protocol for disclosure boundaries).
c. Require lifecycle-oriented carbon declarations at deployment (beyond training): include utilization and cooling assumptions (ISO 14064–1:2018 for organizational GHG inventories; GHG Protocol ICT Sector Guidance for service-level accounting).
d. Use efficiency-by-design procurement: set minimum performance-per-watt baselines and reference recognized power/efficiency benchmarks where available (MLCommons MLPerf Power methodology [77,78] for comparable AI-workload efficiency; EU Ecodesign Regulation 2019/424 for servers and data-storage products on minimum energy performance).
e. Encourage public-sector "model cards" that include energy/CO_2e sections, and support data-centre best practices (e.g., liquid/immersion cooling with PUE targets): (complements EU AI Act Annex XII information to integrators and Article 53(1)(b) transparency to downstream providers [26]; facility targets per ISO/IEC 30134-2 and EN 50600-4-2; operational guidance in the EU Code of Conduct, JRC 2024). At a minimum, energy/CO_2e-aware model cards should report: (i) device-level energy for training and inference; (ii) runtime together with batch size and numerical precision; (iii) hardware and firmware versions; and, where applicable, (iv) data-centre PUE and regional carbon intensity (including siting assumptions and time window), accompanied by a short description of the measurement boundary (device-only vs. wall-plug, inclusion of cooling) and the measurement protocol (instrumentation and sampling cadence).
f. Fund open, comparable measurement infrastructure: reduce variability in energy metrics and improve cross-study reproducibility (supported by EU AI Act Article 53(5) on delegated acts for "measurement and calculation methodologies" and Article 112(6) on energy-efficient GPAI standardisation deliverables [26]; benchmarking comparability via MLPerf Power [77,78]).

Ethical guardrails (deployment-focused):

g. Transparency & auditability: Publish measurement protocols (instrumentation, sampling cadence, batch/sequence settings, pre/post-processing) sufficient for third-party replication (mapped to Annex XI/XII documentation duties) [26].
h. Equity & environmental justice in siting: Disclose local water constraints, grid carbon intensity, and community impacts; track facility

Table 4
Challenges in sustainable AI. Unless explicitly labeled "authors' synthesis," each challenge reflects patterns reported in the reviewed studies; representative sources are listed in the footnote. Items labeled "authors' synthesis" are integrative themes distilled from cross-paper evidence rather than single-study claims.

Category	Challenge	Details
Energy Infrastructure and Availability	Power reliability	Delayed utility upgrades to meet data center demand can disrupt operations when the required demand isn't delivered.
	Cooling needs	Increased temperatures due to climate change led to heightened cooling demands, straining power during peak summer months.
Data and Computational	Data mobility and optimization	Optimizing data workloads based on energy availability remains complex, despite data processing moving across locations.
	Data quality	Poor data quality can reduce AI's energy efficiency, such as necessitating frequent retraining due to inefficient model performance.
Regulatory and Policy	Lack of standards	A lack of uniform global standards, taxonomies, and definitions for AI energy use and digital systems creates challenges in scaling and assessing impact.
	Regulatory complexities	Regional differences in regulations complicate compliance, especially as AI-focused regulations emerge.
	Industry-specific regulations	Regulations for building efficiency and renewable adoption vary widely across industries and regions.
Industry Collaboration and Partnerships	Complex value chains	Difficulty in mapping supply chains and gathering supplier data impacts collaboration.
	Dependence on key players	The concentration of R&D within a few companies limits access to innovation.
	Lack of local capacity	Inadequate local infrastructure and insufficient volume of skilled partners for implementing global AI systems.
	Risk aversion	Telecommunication and energy sectors are hesitant to adopt disruptive technologies, focusing instead on incremental changes.
Mindset, Awareness, and Cultural Shifts	Awareness barrier	There is low awareness about the financial benefits of sustainable AI.
	Mindset shift	There is hesitation to adopt energy-efficient practices across supply chains due to fears of disrupting established profit margins.
	Geographic variations in attitudes	Different regions prioritize AI and sustainability goals based on their socioeconomic and environmental context.
Operational and Technical Challenges	Data privacy and security	Privacy and security concerns arise as AI models require vast data, especially in regions with strict regulations.
	Energy security concerns	Dependency on external energy sources and concerns about energy reliability due to geopolitical issues create hurdles in AI applications.
	Real-time forecasting and optimization	Accurate forecasting tools are needed to optimize resource allocation and identify energy hotspots.

KPIs using ISO/IEC 30134-series and EN 50600-4-2 to avoid shifting burdens to water-stressed or high-carbon regions.

i. Rebound effects: Pair efficiency targets with periodic deployment-stage disclosures (total energy/CO_2e at realized utilization) so scale-up does not erase per-operation gains.

Table 5
Selected regulatory and standards frameworks relevant to energy/CO_2e reporting in AI systems.

Framework	Primary scope	Key energy/CO_2e reporting elements	Relevance for AI/model cards
EU AI Act (2024)	Regulation of AI systems in the EU, including high-risk systems and general-purpose AI models	Requires technical documentation and downstream information that can include energy- and resource-use characteristics for high-risk and general-purpose AI models; empowers the Commission to define "measurement and calculation methodologies" for comparable, verifiable documentation [26].	Anchors AI-specific documentation duties, including energy/CO_2e-related fields in technical files and model cards for systems deployed in the EU, especially for high-risk and general-purpose models.
ISO/IEC 30134-2 and EN 50600-4-2	Data-centre performance metrics and KPIs (e.g., PUE)	Define how to measure and report facility-level indicators such as Power Usage Effectiveness (PUE) and related data-centre KPIs.	Provide the facility-side metrics that AI developers should reference when reporting data-centre energy efficiency and PUE assumptions in model cards and technical documentation.
ISO 14064-1 and GHG Protocol	Organizational and service-level greenhouse-gas accounting	Specify principles and requirements for quantifying and reporting GHG emissions, including boundaries, scopes, and disclosure formats for CO_2e.	Offer a basis for translating measured energy (device energy, PUE, regional carbon intensity) into CO_2e inventories and service-level declarations that can be linked to AI model cards and deployment reports.

Resource-constrained environments: When promoting edge/on-device deployments to cut connectivity energy, report accuracy/latency trade-offs, maintenance burden, and update frequency to ensure benefits are not offset by degraded service quality.

5. Discussion

5.1. Interpretation of results

The assertion that "identified Green AI techniques have a lot of potential to be able to cut energy use" is substantially justified by a variety of algorithm optimizations, as well as dedicated hardware systems. The paper refers to several energy-efficient methods, including pruning, quantization, sparse training, and NAS. For example, pruning also decreases model size by eliminating superfluous parameters. It makes the model less computationally and energy expensive in both training and inference, as argued by the Lottery Ticket Hypothesis [59]. Computation of lower-precision numerical formats converting 32-bit format to 8-bit format provides less memory usage and faster processing, with reports suggesting up to a 50 % reduction in energy consumption [10,83,102]. NAS converts the process of designing energy-efficient networks into an automated task where both resource costs and efficiency are optimized. Recent work on energy-efficient neural-network training further exemplifies this combined strategy: Reguero et al. [105] jointly leverage runtime layer freezing, model quantization, and early stopping as complementary levers for reducing training energy while maintaining task performance. Such algorithmic soft points out the eigen-buttes of Green AI algorithms in terms of allowing a sufficient balance between performance and energy efficiency, which, overall, leads to energy reductions.

Beyond compression and hardware acceleration, structurally efficient model families also contribute to Green AI by reducing redundant computation. Graph neural networks and hypergraph-based frameworks exploit relational structure to focus message passing on informative connections, thereby limiting unnecessary feature propagation. For example, Zhu et al. [106] propose an AI-driven hypergraph neural network for predicting gasoline price trends, where hyperedges encode higher-order relationships between market variables and reduce redundancy in feature propagation while maintaining predictive performance. Such graph- and hypergraph-based designs represent a complementary axis of energy efficiency alongside pruning, quantization, and NAS.

Along with the enhancement of algorithms, the use of specialized hardware is a vital part of minimizing power consumption in AI systems. One of the hardware solutions is the use of Tensor Processing Units (TPUs) and Field-Programmable Gate Arrays (FPGAs). Even TPUs, by comparison, have more FLOPS per watt than conventional GPUs, and neuromorphic chips also allow even more energy savings as they are modeled after the brain. Practical energy-saving can be achieved without much compromise in performance with models like DistilBERT, which matches 97 % of BERT with only 60 % of the computational energy, and MobileNetV2, which is energy-efficient on edge devices through depth-wise separable convolutions [81]. Nevertheless, although these examples support the evidence of such statements as the concept of substantial promise, the report does not provide detailed quantitative comparisons specifying the range of energy savings, 10–50 % or more, or the range of savings in different AI applications, NLP versus CV. The extent of applicability of such techniques to a variety of AI models and fields is not fully addressed.

5.2. Comparison with traditional AI systems

The claim that "Green AI techniques outperform traditional AI models in terms of energy efficiency" is well-supported, as Green AI focuses on optimizing energy usage without compromising performance. Traditional AI models often prioritize accuracy over energy efficiency, leading to models with exponentially growing parameters and escalating energy demands. As noted in the document, up to 90 % of AI research prioritizes precision, which results in models like GPT-3 requiring a staggering 1287 MWh for training. In contrast, Green AI techniques actively work to balance energy efficiency with performance. For example, DistilBERT achieves 97 % of BERT's accuracy while requiring significantly less energy, thanks to knowledge distillation. Similarly, MobileNetV2 uses depth-wise separable convolutions, reducing computational power requirements, making it ideal for resource-constrained devices compared to traditional convolutional neural networks that are computationally intensive. Additionally, hardware solutions like TPUs and FPGAs outperform GPUs in energy efficiency by tailoring computations to specific AI tasks.

While the claim is compelling, the document lacks direct quantitative benchmarks comparing energy consumption across Green AI and traditional AI models, such as watt-hours per inference for each model. The claim would be more convincing with a clear comparison of energy metrics, such as a table or chart, to visually highlight energy savings across specific models. Additionally, while the document acknowledges the trade-off between performance and efficiency, it does not quantify this relationship, which could obscure cases where Green AI sacrifices accuracy in favor of energy savings. By including these details, the analysis would provide a clearer picture of the real-world implications of adopting Green AI techniques. To ensure interpretability and cross-study comparability, all quantitative comparisons in this section and in Table 2 now require the following fields to be shown next to the reported metric(s): (a) hardware/accelerator model (e.g., GPU/TPU/FPGA), (b) node count (and per-node accelerator count if applicable), (c) batch size, and (d) numerical precision (e.g., FP32, BF16, INT8); entries are marked NR when not reported in the source. We maintain source-native units and avoid cross-paper normalization.

To synthesize quantitative evidence across techniques, models, and domains, Table 3 presents quantitative examples across NLP, vision, and LLM inference, reporting efficiency/accuracy trade-offs using the sources' native units. For each entry, we include the model or variant, benchmark, hardware context (accelerator type/node count/batch size/precision where reported), and the corresponding energy metric, so that energy savings can be interpreted in light of hardware and configuration differences. A more detailed per-application breakdown—separating training and inference energy—appears in Appendix Table S6 (metrics exactly as reported; NR where unavailable). Where harmonization is possible, the energy values are reported in the original peer-reviewed units of the source where they were taken (Wh per inference; kWh or MWh per training run). The comparison between data rows is not advisable on a direct numeric basis because there is a significant variation in methodology, hardware and batch size, utilization and approach to measurement. The "Hardware context" column and the notes on measurement method identify the main confounding factors and instrumentation used in each source (NR = not reported).

5.3. Challenges in AI energy efficiency

With AI usage growing steadily, companies constructing data centers are becoming increasingly challenged by the lack of access to sufficient electricity, especially in the form of emission-free resources. The new and massive data centers consume a lot of power, which would cost money and time to serve, depending on the jurisdiction in which they are being constructed. The nature of renewable energy sources and the complexities of storage systems make it challenging to integrate the sources into data centers that can provide reliable power delivery. There are also transmission problems experienced in some regions where there are already almost complete high-tension lines. For example, Dominion Energy in Northern Virginia temporarily restricted new data-center interconnections in parts of Loudoun County due to transmission constraints, placing some projects on a waiting list and delaying service [107].

Although this may not be feasible for all industry players, the hyperscalers are considering using renewable energy sources to power their data centers. Oil and gas firms are increasingly becoming interested in investing in renewable energy to keep up with the rising demand for clean energy to support AI. Nevertheless, some risks have been raised regarding the equitable distribution of costs among client groups to support data center growth, as well as the implications of upgrading infrastructure [108,109]. When designing the price systems on consumer rates, the utility companies have to figure out a balance between sustainability, affordability, and fairness requirements to support the data center expansion. Further, it is essential to examine the grid effects because better demand response and load matching may be necessary to accommodate the growing needs of data centers. This can affect timelines in the licensing stage and construction, including supply chain problems, environmental aspects, and regulatory roadblocks, which might include delays in key grid infrastructure.

One of the most important challenges that AI technologies will face as a part of their future growth is the environmental side of their impact. Any energy source, whether clean or not, has an environmental footprint [110]. With the continued integration of AI in various activities of life, the entrenched concept that there will be energy scarcity is essential in the design concept for the development of the infrastructure of AI. This will be the strategy of making the energy transition with the help of AI instead of exacerbating resource depletion. Such a trade-off between focusing more on 24/7 [102] and carbon-free energy targets or net-zero emissions on the one hand and market objectives on the other hand will be essential in facilitating sustainable AI. The increase in the demand for data centers and changes in the energy industry may lead to a mismatch between the predicted emissions by the end of 2050 and the net-zero objectives.

As the demand for data centers is growing, there is a requirement for dispatchable and firm generation to meet the growing energy needs. This creates a need to generate novelty and investment in carbon-free solutions to overcome the challenges of developing infrastructure and net-zero targets. Renewable sources such as solar and wind are sources of sustainable energy. Still, their intermittency and poor energy storage capacity may cause problems for data centers that require consistent, high-demand electricity that is supplied with high-quality power. To meet the necessary energy, alternative energy solutions can be adopted, which include natural gas, bioenergy, carbon capture and storage (CCS), nuclear, geothermal, and long-duration storage [111]. Several nuclear companies embrace nuclear technologies to assist in the growth of data centers with zero-carbon power in response to COP28 and the global goal to triple nuclear energy production by 2050. This new trend includes small modular reactors (SMRs) and flexible low-carbon energy that comes with limited land occupation and a quick construction schedule. However, nuclear solutions are also being examined in the U. S. by companies such as Microsoft, Google, and Amazon, each offering its own set of benefits and drawbacks [112].

The issues of buying renewable energy for data centers are beyond the issue of intermittency. Still, they would include challenges of physical distance between data center sites and renewable generation facilities, and the costs of transmission are quite high and lack efficiency. To achieve the potential of sustainable AI, we must not just concern ourselves with the energy aspects but also learn much more and research the management of natural resources.

Table 4 summarizes the most relevant challenges to sustainable AI applications in sustainable energy infrastructure, data and computational challenges, regulatory issues, business cooperation between the industry, and operational barriers. It highlights the necessity of better standards, data optimization, and cross-regional and cross-sectoral collaboration to ensure efficient implementation. Except where explicitly labeled "authors" synthesis," the listed challenges are distilled from the peer-reviewed studies included in our review; items marked "authors' synthesis" reflect cross-study integration rather than single-study claims.

5.4. Broader contextual analysis

The discussion part of the presented document is a source of information about the challenges, future directions, and practical implications of Green AI. It highlights the importance of creating a sustainable ecosystem of AI.

Challenges in AI Energy Efficiency: The next challenge in Green AI is the incorporation of renewable power sources, where the variable nature of solar power and limited energy storage capacity is an immense obstacle when it comes to powering data centers, which need a reliable and high-quality power supply. The environmental impact of a data center that is built in a carbon-intensive area like Australia (800 gCO^2e/kWh) is greater than that of a data center built in a low-carbon region like Switzerland (20 gCO^2e/kWh). Other limits, such as congested transmission lines and power delivery capacity issues in certain locations, have further limited the scalability of AI infrastructure, particularly in hyperscale data centers used to train large AI systems. The next hurdle is that of performance and efficiency, as 90 % of the AI research is focused more on performance than on energy efficiency. To perform well, large-scale models, such as GPT-3, require immense quantities of energy (1287 MWh during training and 564 MWh per day of inference periods), which limits the use of Green AI practices in the real world.

Lastly, there are regulatory gaps that pose a major snag. Though the EU AI Act (2024) requires high-risk AI systems to report on energy efficiency, other nations, such as the USA and China, do not include such an obligation in their legislation, thus preventing the regular implementation of Green AI principles on a global scale.

Despite progress, several gaps remain. First, the field lacks unified, reproducible efficiency metrics: studies report power and energy in incompatible ways (e.g., instantaneous vs. average power, device-only vs. system-level, wall-plug vs. proxy figures), undermining cross-comparison. These observations are consistent with recent work on carbon-emission quantification for machine-learning workloads, which documents heterogeneous reporting practices and methodological assumptions across studies [113]. Second, lifecycle-oriented carbon accounting at deployment is seldom reported; beyond training-time estimates, few works quantify end-to-end operational emissions shaped by siting, grid-mix variability, cooling, and utilization. Third, inference-stage efficiency is underexplored relative to training—even though inference often dominates cumulative energy in production—leaving methods, benchmarks, and reporting uneven. Finally, alignment between AI design choices and sustainability regulation remains limited; algorithm, hardware, and siting decisions are rarely mapped to disclosure, procurement, and facility-level requirements across jurisdictions.

There are a few positive pointers in the future where Green AI can be a keystone in the endeavor towards sustainability. An example of these directions is the area of Explainable AI (XAI), which promotes transparency in green AI applications, especially in environmental fields such as agriculture and industry. Neuromorphic computing, which replicates the efficiency of the brain, has the possibility of transforming energy consumption, particularly in applications that need to be energy-efficient in terms of latency. Energy-harvesting AI devices that can utilize additional energy sources such as light and vibrations to reduce reliance on external power are another fascinating field of research. Moreover, environmental AI applications intended to help achieve sustainability outcomes include pollinator boots and underwater robots.

Green AI may have several practical implications for industries and policymakers. Other areas of application of green AI in industry are its use in low-power devices such as smartphones or IoT devices, autonomous systems, healthcare, and smart grids [84]. When it comes to policymakers, providing tax incentives to use energy-efficient devices by creating a tax break to buy efficient hardware or requiring the reporting of carbon footprint to be used would adhere to the EU approach to the AI Act and lead to broader adoption of sustainable AI practices. Besides, the economic advantages of Green AI, including power savings and reduced operational costs, ensure that it is a promising alternative to satisfy sustainability requirements and upgrade the bottom line of organizations. These practice implications indicate that Green AI can not only help to develop sustainability in the environment but also generate economic benefits to companies and industries.

Energy-efficiency results vary with dataset/horizon, hardware (GPU/TPU/FPGA), precision/batch size, and measurement protocol (device-only vs. wall-plug, inclusion of PUE and regional carbon intensity). These factors limit direct cross-paper comparisons; accordingly, Table 3 reports energy in native units (J/inference, Wh/token, MWh) and Appendix Table S4 documents study-level context to aid interpretation.

Bridging gaps to deployment. In practice, the lack of unified efficiency metrics complicates procurement, auditing, and cross-vendor comparability required by emerging disclosure duties [1,26]; absent lifecycle carbon accounting at deployment obscures siting- and PUE-dependent CO_2e and slows approvals [91–94,102]; underreported inference-stage energy impedes capacity planning and cost-of-service estimates; and weak alignment between algorithm/hardware design choices and energy-sector requirements raises compliance risk and limits uptake. To support adoption, we point practitioners and policymakers to the §4.7 "Policy Recommendations—Actionable Checklist," and highlight four near-term steps: (i) report device-level energy for training and inference plus PUE and regional carbon intensity; (ii) set efficiency-by-design procurement baselines (performance/Watt) and reference recognized power/efficiency benchmarks; (iii) prefer low-carbon siting and 24/7 clean-power procurement with documented assumptions; and (iv) publish energy/CO_2e sections in model cards to improve traceability.

In healthcare, adoption hinges on preserving clinical performance and traceability while meeting safety/validation and data-protection constraints; our reviewed applications (e.g., medical imaging) underscore the value of lightweight models at the edge, and we recommend attaching device-level energy/CO_2e fields to model cards and evaluation packets to support procurement and audits [84], with disclosures aligned to §4.7 [1,26]. In smart grids, variability and reliability requirements make inference-stage efficiency and demand-response--friendly deployment essential; reporting PUE and location-aware carbon intensity together with standardized energy metrics improves comparability for grid operators (§2; [18–20,66,114–116,91–94,102]). In autonomous/field systems, tight onboard power budgets and safety constraints favor pruning/quantization and efficient hardware, with duty-cycle energy reported alongside task performance to enable lifecycle-aware decisions (§4.5; [82]).

The literature remains heterogeneous in energy metrics (device-only vs. wall-plug; missing PUE and regional carbon intensity), hardware/batch/precision settings, and reproducibility artifacts; vendor-only reports and small samples limit generalizability. Our review inherits these constraints: we rely on source-reported measurements (no re-instrumentation), exclude grey literature to preserve comparability, and do not perform meta-analysis because outcomes/units are not commensurate. These factors may bias coverage toward well-documented subfields and high-resource settings.

Concrete directions. (i) Adopt a minimum reporting schema for energy/CO_2e (device energy, batch size/precision, runtime, PUE, regional carbon intensity, hardware/firmware versions); (ii) publish open, cross-domain benchmarks pairing accuracy with J/inference, Wh/token, and MWh for training; (iii) include energy/CO_2e sections in model cards; and (iv) provide sector-specific integration templates that align energy metrics with operational and regulatory requirements, building on §4.7 and Table 3.

This review serves three audiences. For policymakers, it links AI-specific reporting duties with energy-sector practices—PUE/DVFS/virtualization, siting/cooling choices, and 24/7 clean-power procurement—to shape actionable compliance checklists. For practitioners, it consolidates algorithmic, hardware, and facility-level levers that can be operationalized in near-term deployments. For researchers, it delineates concrete evidence gaps—standardized efficiency metrics, lifecycle carbon accounting at deployment, inference-stage methods and benchmarks, and policy–design alignment—thereby providing a focused agenda for rigorous, comparable studies.

Despite the fact that the EU AI Act and the related Code of Practice necessitate reporting of energy and resource consumption in model cards of general-purpose AI models with systemic risk, many open questions may be raised concerning the practical validation, such as the necessary measurement tools, sampling windows and statistical strength, audit scope, third-party verification protocol and enforcement.

5.5. Future directions

Looking into the future, it becomes increasingly important to find a new way to combine AI with environmental sustainability. Here, a discussion of some new patterns in green AI is provided, along with an idea of how the new segment is bringing a more sustainable and nature-sensitive future.

Explainable AI (XAI) promises to become a topical green AI use case since it increases transparency and responsibility. As the role of AI in environmental decision-making grows, someone must know the rationale behind the decisions made by the AI device. This is particularly relevant to the green-by AI applications, where the XAI models might provide insights into the mechanisms influencing a more sustainable world. Some of the early uses of XAI are in agriculture [117] and the industrial sector [118]. These directions are already evidenced in the reviewed studies (e.g., agriculture and industrial case studies), where XAI improved interpretability without materially increasing compute overhead (see §4.5), supporting its practical relevance to sustainability-focused deployments [117,118].

Eco-Friendly AI Hardware Accelerators: Hardware design innovation is now concerned with the development of such hardware accelerators that are energy-efficient rather than power-hungry [119]. As summarized earlier (see §4.4), specialized accelerators paired with model-level efficiency (e.g., pruning/quantization) reduce device-level energy and enable lower-power inference at scale without compromising utility. In parallel, structurally efficient architectures—such as graph neural networks and hypergraph-based models—offer a model-design route to Green AI by encoding high-order relations in the topology and reducing redundant feature propagation, thereby lowering computational cost for structure-rich domains such as energy and commodity markets [106].

Neuromorphic Computing: Neuromorphic computing is a revolutionary change in the technology used in computers that works in the same way as the human brain. The human brain is considered highly energy-efficient and can perform complex tasks using much less energy than conventional computers. Neuromorphic computing attempts to achieve this same efficiency by approximating real-life neural networks in the form of spiking neural networks (SNNs). This will achieve a substantial level of energy efficiency and performance, especially in AI applications where low reaction time is needed. Another sustainability proposal on the future of neural networks is by Oh et al. [120], who propose to create biodegradable neural network hardware, thus reducing any waste induced by technology. This aligns with the evidence reviewed for event-driven, spiking hardware in §4.4, where workload-appropriate neuromorphic platforms demonstrated high performance-per-watt in edge scenarios.

Energy-Harvesting AI Devices: AI devices capable of harvesting energy from the surrounding environment to power AI systems are the focus of research efforts [121]. As an example, ambient light can be harvested and energy can be generated by capturing vibrations or harvesting heat [122]. Further integration of energy harvesting components will help AI equipment become more energy independent, reducing its dependence on external power sources and making AI usage more sustainable. These concepts complement the low-power edge patterns highlighted in §4.4, where ultra-efficient sensing and inference reduce the energy budget required for practical energy-harvesting operation.

AI has been used in environmental conservation through its use in crop pollinators, by bot robots, and ocean AI [123,124]. Some of the robotic pollinators are designed to replicate the process of natural pollination [125] and are helping to resolve the problem of the falling number of natural pollinators, thus ensuring food security and sustainable farming practices. To conserve marine life, underwater drones powered by AI and advanced sensors are being deployed to survey the marine environment, assess its well-being, and warn against imminent attacks on marine life [126]. These drones monitor life in the ocean, detect contamination [127], report unlawful fishing [128], and assist in protecting biodiversity and limiting environmental concerns in the seas as well as on land. These applications illustrate how Green-AI techniques from our review (e.g., lightweight models on edge hardware) can be operationalized in field robotics with strict energy budgets (see §4.5).

Combined, these developments reflect the idea that AI is emerging as a necessity in the drive to implement sustainable practices, conserve and safeguard biodiversity, and the environment in the various diverse ecosystems worldwide. Nature of BERT, a transformer-based bidirectional encoder, and its impact on Natural Language Processing tasks like sentence boundary detection, tokenization, and grammar correction, aiming to enhance its capabilities [129]. Advanced technology impacts the environment, leading to the need for better-designed buildings. The concept of "green buildings" promotes eco-friendly activities, and the Internet of Things (IoT) is transforming this. This chapter surveys major IoT concepts for green building management, summarizing scientific contributions and discussing open research challenges in this emerging area [130].

The demand for AI chips has increased significantly, especially in 2023, with the AI boom, which has seen companies such as NVIDIA record high revenues. The result of such a surge would be a significant increase in the energy requirement of AI. As an example, the use of AI, such as ChatGPT, in Google searches may require using enormous server resources, which would be up to 29.2 TWh per year. This worst-case scenario is, in some ways, the illustration of the increasing energy footprint, but it is unlikely to occur soon due to the shortcomings of equipment and costs. These are scenario-based estimates from the reviewed source and not forecasts of actual operations; realized values depend on hardware efficiency, duty cycles, and siting (see §4.7) [14]. Conversely, the estimated AI server shipments of NVIDIA in 2023 would take a comparatively small amount of 5.7 TWh to 8.9 TWh. No matter how difficult it can be to produce chips, the advancements in hardware efficiency and AI model algorithms can potentially alleviate the energy concerns, despite Jevons Paradox hinting that the greater the efficiency, the stronger the necessity may become. Moreover, the trend of reusing GPUs used to mine cryptocurrencies in performing AI applications, driven by the rise in model efficiency, is likely to change a large portion of the electricity consumption over to AI [14]. Figure S4: Distribution of AI Domains by Risk Level.

Fig. 10 compares the estimated energy consumption per request for AI-powered systems versus Google search. It shows that AI models like ChatGPT and AI-powered Google search (New State Research) consume significantly more energy than traditional Google search and models like BLOOM.

Connecting future directions to reviewed evidence: Hybrid algorithm–hardware approaches (e.g., depthwise-separable networks and accelerator-optimized pipelines) demonstrated efficiency gains in our case studies (§4.4–§4.5); energy-harvesting/ultra-low-power edge devices align with event-driven platforms discussed in §4.4; and refined data-center operations (low-PUE cooling, siting with 24/7 clean power) are evidenced in §§2 and 4.7.

The systematic review highlights the massive potential of Green AI methods, including model compression, a 60 % reduction of energy with DistilBERT, low-precision computing, 50 % savings by quantization, neural architecture search, and specialized hardware such as TPUs and neuromorphic computing, to considerably lower the energy demands of AI relative to conventional models. Based on 131 studies, most of which were published between 2017 and 2025, gathered on IEEE, Elsevier, and arXiv, these methods prove better energy efficiency, especially when applied to NLP and computer vision. Nevertheless, the existence of hurdles such as the intermittency of renewable energy, lack of regulation, and incomplete domain coverage requires subsequent research on hybrid systems, more sophisticated hardware, and optimized data centers to capitalize on Green AI to bring about sustainable practice of AI development.

6. Conclusion

This systematic review has provided an overview of the existing state of Green AI methods of cutting energy use in AI systems. The strongest and most well-publicized efficiency improvements, which are usually 2090 percentage reductions in algorithm techniques (compression, quantization, pruning, distillation) and several-fold efficiency with specialized hardware, are most often found in inference, and less common, less reproducible, and much more hardware-sensitive large-scale pre-training. New cooling, virtualization and renewable at the infrastructure level, enhanced cooling, virtualization and renewable can achieve facility scale savings, and PUE values are minimized at 1.051.1 in liquid-cooled systems. Potentially most effective actions in the near- and medium-term, such as (i) inference-based optimizations (quantization to 8-bit/4-bit, structured sparsity, and optimizer-specific kernels) in production pipelines, (ii) requiring standard energy reporting in model cards and benchmarks, (iii) focusing on reproducible measurement protocols and open datasets to characterize energy consumption, and (iv) including liquid/immersion cooling and renewable purchasing in the design of new data-centers. Practitioners are advised to be more realistic about savings in the training phase and test methods on their workload and hardware stack. Significant drawbacks still exist non-uniform reporting standards, a literature bias to inference workloads, non-uniform methodologies in measuring, regulatory frameworks (e.g., the EU AI Act energy-transparency requirements) whose practical validation procedures are yet to be determined. This review is in turn limited by the heterogeneity of the studies and partial transparency artifacts of the primary sources. Subsequent studies need to be conducted on hybrid algorithm-hardware co-design, scalable neuromorphic and energy-harvesting, robust cross-study benchmarks (e.g., extensions of MLPerf Power), and policy to impose verifiable energy reporting. It is only with a combined technical, methodological and regulatory

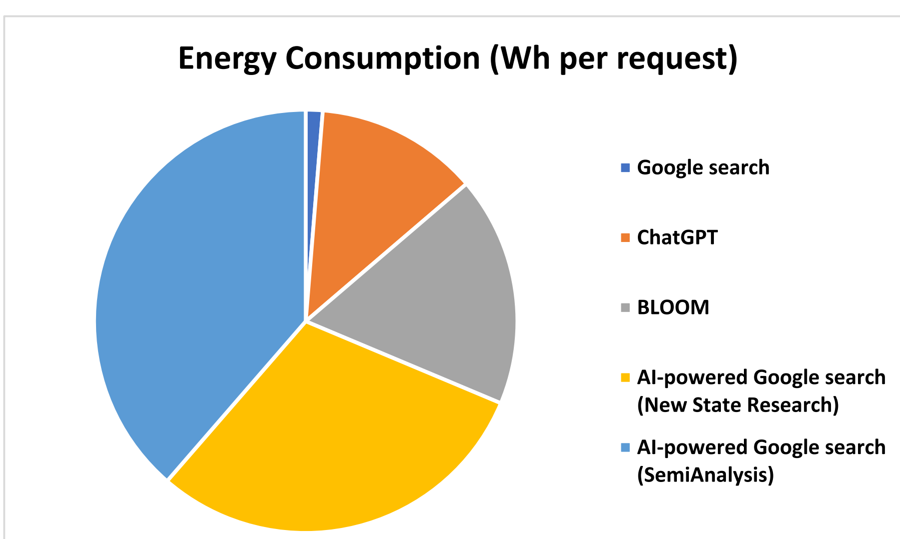

Fig. 10. Estimated energy consumption per request for AI-powered systems vs. Google search [14].

advancement that the AI community can provide computationally strong models that are both globally responsible and sustainable.

CRediT authorship contribution statement

Sunawar Khan: Resources, Investigation, Formal analysis, Data curation. **Naila Sammar Naz:** Validation, Resources, Investigation. **Tehseen Mazhar:** Writing – review & editing, Writing – original draft, Visualization, Validation. **Muhammad Usman Tariq:** Visualization, Validation, Resources, Investigation. **Tariq Shahzad:** Writing – review & editing, Visualization, Validation. **Sghaier Guizani:** Writing – original draft, Visualization, Validation. **Habib Hamam:** Validation, Software, Funding acquisition, Formal analysis.

Funding

No.

Declaration of competing interest

The authors have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.array.2025.100652.

Data availability

Data will be made available on request.

References

[1] Patterson D, et al. The carbon footprint of machine learning training will Plateau, then shrink. Computer 2022;55(7):18–28.
[2] Agency IE. Data centres and data transmission networks. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks. [Accessed 30 July 2025].
[3] Makridakis S. The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. Futures 2017;90:46–60.
[4] Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Gener Comput Syst 2022;135:364–81.
[5] Liu S, Liu L, Tang J, Yu B, Wang Y, Shi W. Edge computing for autonomous driving: opportunities and challenges. Proc IEEE 2019;107(8):1697–716.
[6] Miao Q, Zheng W, Lv Y, Huang M, Ding W, Wang F-Y. DAO to HANOI via DeSci: AI paradigm shifts from AlphaGo to ChatGPT. IEEE/CAA J Automat Sin 2023;10(4):877–97.
[7] Ouyang L, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730–44.
[8] Gholami A, Yao Z, Kim S, Hooper C, Mahoney MW, Keutzer K. Ai and memory wall. IEEE Micro 2024;44(3):33–9.
[9] Energy M. Why is energy efficiency important?. https://www.medianenergy.com/blog/why-is-energy-efficiency-important.
[10] I. Agdestein. "AI energy efficiency: reducing power consumption in AI models." Focalx logo. https://focalx.ai/ai/ai-energy-efficiency/(accessed July 30, 2025.
[11] Bolón-Canedo V, Morán-Fernández L, Cancela B, Alonso-Betanzos A. A review of green artificial intelligence: towards a more sustainable future. Neurocomputing 2024;599:128096.
[12] Verdecchia R, Sallou J, Cruz L. A systematic review of green AI. Wiley Interdiscip Rev: Data Min Knowl Discov 2023;13(4):e1507. https://doi.org/10.1002/widm.1507.
[13] SemiAnalysis. "The inference cost of search disruption – large language model cost analysis." semianalysis.com. https://semianalysis.com/2023/02/09/the-inference-cost-of-search-disruption/(accessed July 30, 2025.
[14] De Vries A. The growing energy footprint of artificial intelligence. Joule 2023;7(10):2191–4. https://doi.org/10.1016/j.joule.2023.09.004.
[15] Schwartz R, Dodge J, Smith NA, Etzioni O. Green ai. Commun ACM 2020;63(12):54–63. https://doi.org/10.1145/338183.
[16] Goals UNSD. Goal 10: reduce inequality within and among countries [Online]. Available: https://www.un.org/sustainabledevelopment/inequality/; 2025.
[17] Yigitcanlar T, Mehmood R, Corchado JM. Green artificial intelligence: towards an efficient, sustainable and equitable technology for smart cities and futures. Sustainability 2021;13(16):8952. https://doi.org/10.3390/su13168952.
[18] Chen J, Zeng G-Q, Zhou W, Du W, Lu K-D. Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. Energy Convers Manag 2018;165:681–95 [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0196890418303261.
[19] Díaz–Vico D, Torres–Barrán A, Omari A, Dorronsoro JR. Deep neural networks for wind and solar energy prediction. Neural Process Lett 2017;46(3):829–44. https://doi.org/10.1007/s11063-017-9613-7.
[20] Abdel-Nasser M, Mahmoud K. Accurate photovoltaic power forecasting models using deep LSTM-RNN. Neural Comput Appl 2019;31(7):2727–40. https://doi.org/10.1007/s00521-017-3225-z.
[21] Qolomany B, et al. Leveraging machine learning and big data for smart buildings: a comprehensive survey. IEEE Access 2019;7:90316–56. https://doi.org/10.1109/ACCESS.2019.2926642.
[22] Pham A-D, Ngo N-T, Truong TTH, Huynh N-T, Truong N-S. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. J Clean Prod 2020;260:121082. https://doi.org/10.1016/j.jclepro.2020.121082.
[23] Zekić-Sušac M, Mitrović S, Has A. Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. Int J Inf Manag 2021;58:102074. https://doi.org/10.1016/j.ijinfomgt.2020.102074.
[24] Ghazal TM, Hasan MK, Ahmad M, Alzoubi HM, Alshurideh M. Machine learning approaches for sustainable cities using internet of things. In: The effect of information technology on business and marketing intelligence systems. Springer; 2023. p. 1969–86.
[25] Milojevic-Dupont N, Creutzig F. Machine learning for geographically differentiated climate change mitigation in urban areas. Sustain Cities Soc 2021;64:102526. https://doi.org/10.1016/j.scs.2020.102526.
[26] Zhang M, et al. Impact of urban expansion on land surface temperature and carbon emissions using machine learning algorithms in Wuhan, China. Urban Clim 2023;47:101347. https://doi.org/10.1016/j.uclim.2022.101347.
[27] Renggli C, Ashkboos S, Aghagolzadeh M, Alistarh D, Hoefler T. SparCML: High-performance sparse communication for machine learning. In: Proceedings of the international conference for high performance computing, networking, storage and analysis; 2019. p. 1–15. https://dl.acm.org/doi/abs/10.1145/3295500.3356222.
[28] Tay Y, Bahri D, Yang L, Metzler D, Juan D-C. Sparse sinkhorn attention. In: International conference on machine learning. PMLR; 2020. p. 9438–47 [Online]. Available: https://arxiv.org/abs/2002.11296.
[29] Bibikar S, Vikalo H, Wang Z, Chen X. Federated dynamic sparse training: computing less, communicating less, yet learning better. In: Proceedings of the AAAI conference on artificial intelligence, vol. 36; 2022. p. 6080–8. no. 6, https://arxiv.org/abs/2112.09824.
[30] Elgabli A, Park J, Bedi AS, Issaid CB, Bennis M, Aggarwal V. Q-GADMM: quantized group ADMM for communication efficient decentralized machine learning. IEEE Trans Commun 2020;69(1):164–81. https://doi.org/10.1109/TCOMM.2020.3026398.
[31] Khirirat S, Magnússon S, Aytekin A, Johansson M. A flexible framework for communication-efficient machine learning. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35; 2021. p. 8101–9. no. 9, https://arxiv.org/abs/2003.06377.
[32] Xiao G, Lin J, Seznec M, Wu H, Demouth J, Han S. Smoothquant: accurate and efficient post-training quantization for large language models. In: International conference on machine learning. PMLR; 2023. p. 38087–99 [Online]. Available: https://arxiv.org/abs/2211.10438.
[33] Yang T-J, Chen Y-H, Sze V. Designing energy-efficient convolutional neural networks using energy-aware pruning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 5687–95 [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Yang_Designing_Energy-Efficient_Convolutional_CVPR_2017_paper.html.
[34] Zafrir O, Boudoukh G, Izsak P, Wasserblat M. Q8bert: quantized 8bit bert. In: 2019 fifth workshop on energy efficient machine learning and cognitive Computing-NeurIPS edition (EMC2-NIPS). IEEE; 2019. p. 36–9 [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9463531.
[35] Liu Z, Wang Y, Han K, Zhang W, Ma S, Gao W. Post-training quantization for vision transformer. Adv Neural Inf Process Syst 2021;34:28092–103. https://doi.org/10.48550/arXiv.2106.14156.
[36] Suárez-Marcote S, Morán-Fernández L, Bolón-Canedo V. Logarithmic division for green feature selection: an information-theoretic approach. In: European symposium on artificial neural networks, computational intelligence and machine learning; 2023 [Online]. Available: https://www.esann.org/sites/default/files/proceedings/2023/ES2023-77.pdf.
[37] Castillo-García G, Morán-Fernández L, Bolón-Canedo V. Efficient feature selection for domain adaptation using mutual information maximization. In: European symposium on artificial neural networks, computational intelligence and machine learning; 2023 [Online]. Available: https://www.esann.org/sites/default/files/proceedings/2023/ES2023-61.pdf.
[38] Lourenço A, Ferraz C, Meira J, Marreiros G, Bolón-Canedo V, Alonso-Betanzos A. Automated green machine learning for condition-based maintenance. In: Proceedings of the European symposium on artificial neural networks. Computational Intelligence and Machine Learning; October, 2023. p. 4–6 [Online]. Available: https://www.researchgate.net/profile/Afonso-Lourenco-3/publication/373996780_Automated_green_machine_learning_for_condition-based_maintenance/links/650cc64f61f18040c211689b/Automated-green-machine-learning-for-condition-based-maintenance.pdf.
[39] Walton J. Nvidia unveils its next-generation 7nm ampere A100 GPU for data centers, and it's absolutely massive. Tom's Hardware; May 14, 2020 [Online]. Available: https://www.tomshardware.com/news/nvidia-ampere-A100-gpu-7nm. [Accessed 6 September 2025].

[40] Jouppi N, et al. Tpu v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In: Proceedings of the 50th annual international symposium on computer architecture; 2023. p. 1–14 [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3579371.3589350.
[41] Osta M, Alameh M, Younes H, Ibrahim A, Valle M. Energy efficient implementation of machine learning algorithms on hardware platforms. In: 2019 26th IEEE international conference on electronics, circuits and systems (ICECS). IEEE; 2019. p. 21–4 [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8965157.
[42] Stamoulis D, et al. Single-path nas: designing hardware-efficient convnets in less than 4 hours. In: Joint european conference on machine learning and knowledge discovery in databases. Springer; 2019. p. 481–97 [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-46147-8_29.
[43] Anthony LFW, Kanding B, Selvan R. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. arXiv preprint arXiv: 03051 2020. https://doi.org/10.48550/arXiv.2007.03051.
[44] Singh S, Sulthana R, Shewale T, Chamola V, Benslimane A, Sikdar B. Machine-learning-assisted security and privacy provisioning for edge computing: a survey. IEEE Internet Things J 2021;9(1):236–60. https://doi.org/10.1109/JIOT.2021.3098051.
[45] Wan Y, Qu Y, Gao L, Xiang Y. Privacy-preserving blockchain-enabled federated learning for B5G-Driven edge computing. Comput Netw 2022;204:108671. https://doi.org/10.1016/j.comnet.2021.108671.
[46] Novoa-Paradela D, Fontenla-Romero O, Guijarro-Berdiñas B. Fast deep autoencoder for federated learning. Pattern Recogn 2023;143:109805. https://doi.org/10.1016/j.patcog.2023.109805.
[47] Devarakota P, Tsesmetzis N, Alpak FO, Gala A, Hohl D. AI and the net-zero journey: energy demand, emissions, and the potential for transition. arXiv preprint arXiv: 10750 2025. https://doi.org/10.48550/arXiv.2507.10750.
[48] Ghasemi A, Toroghi Haghighat A. A multi-objective load balancing algorithm for virtual machine placement in cloud data centers based on machine learning. Computing 2020;102(9):2049–72. https://doi.org/10.1007/s00607-020-00813-w.
[49] Dong H, Munir A, Tout H, Ganjali Y. Next-generation data center network enabled by machine learning: review, challenges, and opportunities. IEEE Access 2021;9:136459–75. https://doi.org/10.1109/ACCESS.2021.3117763.
[50] Li Y, Wen Y, Tao D, Guan K. Transforming cooling optimization for green data center via deep reinforcement learning. IEEE Trans Cybern 2019;50(5):2002–13. https://doi.org/10.1109/TCYB.2019.2927410.
[51] Cao Z, Zhou X, Hu H, Wang Z, Wen Y. Toward a systematic survey for carbon neutral data centers. IEEE Commun Surv Tutor 2022;24(2):895–936. https://doi.org/10.1109/COMST.2022.3161275.
[52] Moreno-Vozmediano R, Montero RS, Huedo E, Llorente IM. Efficient resource provisioning for elastic cloud services based on machine learning techniques. J Cloud Comput 2019;8(1):1–18. https://doi.org/10.1186/s13677-019-0128-9.
[53] Vianova. Data centers. https://www.vianova.it/en/data-center/.
[54] Kulkarni U, et al. AI model compression for edge devices using optimization techniques. In: Modern approaches in machine learning and cognitive science: a walkthrough: latest trends in AI, vol. 2. Springer; 2021. p. 227–40.
[55] Liu Z, Sun M, Zhou T, Huang G, Darrell T. Rethinking the value of network pruning. arXiv preprint arXiv: 05270 2018. https://doi.org/10.48550/arXiv.1810.05270.
[56] Jacob B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 2704–13 [Online]. Available: https://arxiv.org/abs/1712.05877v1.
[57] Bukhari AH, et al. Predictive analysis of stochastic stock pattern utilizing fractional order dynamics and heteroscedastic with a radial neural network framework. Eng Appl Artif Intell 2024;135:108687. https://doi.org/10.1016/j.engappai.2024.108687.
[58] Reddy MI, Rao PV, Kumar TS, R. K S. Encryption with access policy and cloud data selection for secure and energy-efficient cloud computing. Multimed Tool Appl 2024;83(6):15649–75. https://doi.org/10.1007/s11042-023-16082-6.
[59] Frankle J, Carbin M. The lottery ticket hypothesis: finding sparse, trainable neural networks. arXiv preprint arXiv: 03635 2018. https://doi.org/10.48550/arXiv.1803.03635.
[60] Gray RM, Neuhoff DL. Quantization. IEEE Trans Inf Theor 2002;44(6):2325–83. https://doi.org/10.1109/18.720541.
[61] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv: 01108 2019. https://doi.org/10.48550/arXiv.1910.01108.
[62] Liu C, et al. Progressive neural architecture search. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 19–34 [Online]. Available: https://arxiv.org/abs/1712.00559v3.
[63] Cai H, Zhu L, Han S. Proxylessnas: direct neural architecture search on target task and hardware. arXiv preprint arXiv: 00332 2018. https://doi.org/10.48550/arXiv.1812.00332.
[64] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4510–20 [Online]. Available: https://arxiv.org/abs/1801.04381v4.
[65] Moore GE. Cramming more components onto integrated circuits. Proc IEEE 1998; 86(1):82–5. https://doi.org/10.1109/JPROC.1998.658762.
[66] Aslam S, Herodotou H, Mohsin SM, Javaid N, Ashraf N, Aslam S. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. Renew Sustain Energy Rev 2021;144:110992. https://doi.org/10.1016/j.rser.2021.110992.
[67] D. Patterson. "Good news about the carbon footprint of machine learning training" Google. https://research.google/blog/good-news-about-the-carbon-footprint-of-machine-learning-training/(accessed.
[68] Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S. Sparse representation for computer vision and pattern recognition. Proc IEEE 2010;98(6):1031–44. https://doi.org/10.1109/JPROC.2010.2044470.
[69] Zhang Z, Xu Y, Yang J, Li X, Zhang D. A survey of sparse representation: algorithms and applications. IEEE Access 2015;3:490–530. https://doi.org/10.1109/ACCESS.2015.2430359.
[70] Han J, Orshansky M. Approximate computing: an emerging paradigm for energy-efficient design. In: 2013 18th IEEE european test symposium (ETS). IEEE; 2013. p. 1–6 [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6569370.
[71] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big data 2016;3(1):9. https://doi.org/10.1186/s40537-016-0043-6.
[72] Dhar P. The carbon impact of artificial intelligence. London: Nature Publishing Group UK; 2020.
[73] Tabbakh A, Al Amin L, Islam M, Mahmud GI, Chowdhury IK, Mukta MSH. Towards sustainable AI: a comprehensive framework for green AI. Disc Sustain 2024;5(1):408. https://doi.org/10.1007/s43621-024-00641-4.
[74] Vandendriessche J, Wouters N, da Silva B, Lamrini M, Chkouri MY, Touhafi A. Environmental sound recognition on embedded systems: from FPGAs to TPUs. Electronics 2021;10(21):2622. https://doi.org/10.3390/electronics10212622.
[75] Hascoet T, Zhuang W, Febvre Q, Ariki Y, Takiguchi T. Reducing the memory cost of training convolutional neural networks by CPU offloading. J Software Eng Appl 2019;12(8):307–20. https://doi.org/10.4236/jsea.2019.128019.
[76] Zawish M, Davy S, Abraham L. Complexity-driven model compression for resource-constrained deep learning on edge. IEEE Trans Artif 2024;5(8):3886–901. https://doi.org/10.1109/TAI.2024.3353157.
[77] Tschand A, et al. MLPerf power: benchmarking the energy efficiency of machine learning systems from μWatts to MWatts for sustainable AI. 2024. arXiv: 2410.12032vol. 1.
[78] MLCommons. MLPerf inference v4.0 results. Press release; Mar 27, 2024. arXiv.
[79] He A, et al. HLSTransform: energy-efficient Llama-2 inference on FPGAs via high-level synthesis. arXiv:2405 2024:00738.
[80] Isik M, et al. Accelerating sensor fusion in neuromorphic computing: a case study on Loihi-2. 2024, 16096. arXiv:2408.
[81] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); 2019. p. 4171–86 [Online]. Available: https://aclanthology.org/N19-1423/.
[82] Dettmers T, Lewis M, Belkada Y, Zettlemoyer L. Gpt3. int8 8-bit matrix multiplication for transformers at scale. Adv Neural Inf Process Syst 2022;35: 30318–32. https://doi.org/10.48550/arXiv.2208.07339.
[83] Mohammed A, Schmidt B, Wang L, Gao L. Minimizing energy consumption for robot arm movement. Proced CIRP 2014;25:400–5. https://doi.org/10.1016/j.procir.2014.10.055.
[84] Shuvo MB, Ahommed R, Reza S, Hashem M. CNL-UNet: a novel lightweight deep learning architecture for multimodal biomedical image segmentation with false output suppression. Biomed Signal Process Control 2021;70:102959. https://doi.org/10.1016/j.bspc.2021.102959.
[85] Liu D, Zhao F-Y, Tang G-F. Active low-grade energy recovery potential for building energy conservation. Renew Sustain Energy Rev 2010;14(9):2736–47. https://doi.org/10.1016/j.rser.2010.06.005.
[86] Gullbrand J, Luckeroth MJ, Sprenger ME, Winkel C. Liquid cooling of compute system. J Electron Packag 2019;141(1):010802. https://doi.org/10.1115/1.4042802.
[87] Pambudi NA, et al. Preliminary experimental of GPU immersion-cooling. E3S web of conferences, vol. 93. EDP Sciences; 2019, 03003 [Online]. Available: https://www.e3s-conferences.org/articles/e3sconf/abs/2019/19/e3sconf_cgeee2018_03003/e3sconf_cgeee2018_03003.html.
[88] Pambudi NA, Sarifudin A, Firdaus RA, Ulfa DK, Gandidi IM, Romadhon R. The immersion cooling technology: current and future development in energy saving. Alex Eng J 2022;61(12):9509–27. https://doi.org/10.1016/j.aej.2022.02.059.
[89] Zhang H, Shao S, Xu H, Zou H, Tian C. Free cooling of data centers: a review. Renew Sustain Energy Rev 2014;35:171–82. https://doi.org/10.1016/j.rser.2014.04.017.
[90] Zhang Y, Wei Z, Zhang M. Free cooling technologies for data centers: energy saving mechanism and applications. Energy Proc 2017;143:410–5. https://doi.org/10.1016/j.egypro.2017.12.703.
[91] Le Sueur E, Heiser G. Dynamic voltage and frequency scaling: the laws of diminishing returns. In: Proceedings of the 2010 international conference on power aware computing and systems; 2010. p. 1–8 [Online]. Available: https://www.usenix.org/legacy/events/hotpower/tech/full_papers/LeSueur.pdf.
[92] Kumar R, Khatri SK, Diván MJ. Power usage efficiency (PUE) optimization with counterpointing machine learning techniques for data center temperatures. Int J Mathemat Eng Manag Sci 2021;6(6):1594. https://doi.org/10.33889/IJMEMS.2021.6.6.095.
[93] Mukherjee D, Chakraborty S, Sarkar I, Ghosh A, Roy S. A detailed study on data centre energy efficiency and efficient cooling techniques. Int J 2020;9(5) [Online]. Available: https://www.researchgate.net/profile/Dibyendu-Mukherjee-3/publication/344100890_A_Detailed_Study_on_Data_Centre_Energy_Efficiency_and_Efficient_Cooling_Techniques/links/5f9bddca299bf1b53e51465b/A-De

tailed-Study-on-Data-Centre-Energy-Efficiency-and-Efficient-Cooling-Techniques.pdf.
[94] Helali L, Omri MN. A survey of data center consolidation in cloud computing systems. Comp Sci Rev 2021;39:100366. https://doi.org/10.1016/j.cosrev.2021.100366.
[95] Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: a survey. Int J Comput Vis 2021;129(6):1789–819. https://doi.org/10.1007/s11263-021-01453-z.
[96] Hoefler T, Alistarh D, Ben-Nun T, Dryden N, Peste A. Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. J Mach Learn Res 2021;22(241):1–124. https://doi.org/10.48550/arXiv.2102.00554.
[97] Zhao T, et al. A survey of deep learning on mobile devices: applications, optimizations, challenges, and research opportunities. In: Proceedings of the IEEE, vol 110; 2022. p. 334–54. https://doi.org/10.1109/JPROC.2022.3153408. no. 3.
[98] Mittal S, Vaishay S. A survey of techniques for optimizing deep learning on GPUs. J Syst Architect 2019;99:101635. https://doi.org/10.1016/j.sysarc.2019.101635.
[99] Hodak M, Gorkovenko M, Dholakia A. Towards power efficiency in deep learning on data center hardware. In: 2019 IEEE international conference on big data (big data). IEEE; 2019. p. 1814–20 [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9005632.
[100] Ai H. High-level expert group on artificial intelligence, vol 6. Ethics guidelines for trustworthy AI; 2019 [Online]. Available: https://www.aepd.es/sites/default/files/2019-09/ai-definition.pdf.
[101] Commision E. Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342.
[102] Forum WE. Achieving 24/7 clean power procurement and consumption in industry. Modernizing Industrial Energy Demand; 2024 [Online]. Available: https://www3.weforum.org/docs/WEF_MODERNIZING_INDUSTRIAL_ENERGY_DEMAND_2024.pdf.
[103] Li W, et al. A general framework for unmet demand prediction in on-demand transport services. IEEE Trans Intell Transport Syst 2018;20(8):2820–30. https://doi.org/10.1109/TITS.2018.2873092.
[104] Hacker P. Sustainable AI regulation. Common Mark Law Rev 2024;61(2). https://doi.org/10.54648/cola2024025.
[105] Reguero ÁD, Martínez-Fernández S, Verdecchia R. Energy-efficient neural network training through runtime layer freezing, model quantization, and early stopping. Comput Stand Interfac 2025;92:103906. https://doi.org/10.1016/j.csi.2024.103906.
[106] Zhu P, et al. AI-driven hypergraph neural network for predicting gasoline price trends. Energy Econ 2025;151:108895. https://doi.org/10.1016/j.eneco.2025.108895.
[107] J. Hiller. "Three New York cities' worth of power: AI is stressing the grid." Wall St J. https://www.wsj.com/business/energy-oil/ai-data-center-boom-spurs-race-to-find-power-87cf39dd?mod=Searchresults_%20pos1&page=1#comments_sector. (accessed on December 5, 2025).
[108] Economist T. Oil bosses have big hopes for AI Boom [Online]. Available: https://www.economist.com/business/2024/11/07/oilbosses-have-big-hopes-for-the-ai-boom; 2024.
[109] J. Lindemann. "With load growth and fear of rising utility bills, are low-income customers protected?" DSIRinsights. https://www.dsireinsight.com/blog/2024/7/9/with-load-growth-and-fear-of-rising-utility-bills-are-low-income-customers-protected (accessed.
[110] Energy UDo. Environmental impacts of clean energy. https://www.energy.gov/eere/environmentalimpacts-clean-energy.
[111] C. f. C. a. E. S. (C2ES). "Carbon capture." C2ES. https://www.c2es.org/content/carbon-capture/#:~:text=Carbon%20capture%2C%20use%2C%20and%20storage,decarbonization%20in%20the%20industrial%20sector (accessed.
[112] Energy UDo. At COP28, countries launch declaration to triple nuclear energy capacity by 2050, recognizing the key role of nuclear energy in reaching net zero. 2023 [Online]. Available: https://www.energy.gov/articles/cop28-countries-launch-declaration-triple-nuclear-energy-capacity-2050-recognizing-key.
[113] Hasan SM, Islam T, Saifuzzaman M, Ahmed KR, Huang CH, Shahid AR. Carbon emission quantification of machine learning: a review. IEEE Trans Sustain Comp 2025. https://doi.org/10.1109/TSUSC.2025.3578834.
[114] Wang H-z, Li G-q, Wang G-b, Peng J-c, Jiang H, Liu Y-t. Deep learning based ensemble approach for probabilistic wind power forecasting. Appl Energy 2017;188:56–70. https://doi.org/10.1016/j.apenergy.2016.11.111.
[115] Lin K-P, Pai P-F, Ting Y-J. Deep belief networks with genetic algorithms in forecasting wind speed. IEEE Access 2019;7:99244–53. https://doi.org/10.1109/ACCESS.2019.2929542.
[116] Sorkun MC, Paoli C, Incel ÖD. Time series forecasting on solar irradiation using deep learning. In: 2017 10th international conference on electrical and electronics engineering (ELECO). IEEE; 2017. p. 151–5 [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8266215.
[117] Delaney E, Greene D, Shalloo L, Lynch M, Keane MT. Forecasting for sustainable dairy produce: enhanced long-term, milk-supply forecasting using k-NN for data augmentation, with prefactual explanations for XAI. In: International conference on case-based reasoning. Springer; 2022. p. 365–79. https://doi.org/10.1007/978-3-031-14923-8_24.
[118] Turner C, Okorie O, Oyekan J. XAI sustainable human in the loop maintenance. IFAC-PapersOnLine 2022;55(19):67–72. https://doi.org/10.1016/j.ifacol.2022.09.185.
[119] Rahmani H, et al. Next-generation IoT devices: sustainable eco-friendly manufacturing, energy harvesting, and wireless connectivity. IEEE J Microwav 2023;3(1):237–55. https://doi.org/10.1109/JMW.2022.3228683.
[120] Oh S, Kim H, Kim SE, Kim M-H, Park H-L, Lee S-H. Biodegradable and flexible polymer-based memristor possessing optimized synaptic plasticity for eco-friendly wearable neural networks with high energy efficiency. Adv Intell Syst 2023;5(5):2200272. https://doi.org/10.1002/aisy.202200272.
[121] Divya S, et al. Smart data processing for energy harvesting systems using artificial intelligence. Nano Energy 2023;106:108084. https://doi.org/10.1016/j.nanoen.2022.108084.
[122] Ali A, Shaukat H, Bibi S, Altabey WA, Noori M, Kouritem SA. Recent progress in energy harvesting systems for wearable technology. Energy Strategy Rev 2023;49:101124. https://doi.org/10.1016/j.esr.2023.101124.
[123] Hiraguri T, Kimura T, Endo K, Ohya T, Takanashi T, Shimizu H. Shape classification technology of pollinated tomato flowers for robotic implementation. Sci Rep 2023;13(1):2159. https://doi.org/10.1038/s41598-023-27971-z.
[124] Shivaprakash KN, et al. Potential for artificial intelligence (AI) and machine learning (ML) applications in biodiversity conservation, managing forests, and related services in India. Sustainability 2022;14(12):7154. https://doi.org/10.3390/su14127154.
[125] Joy P, Thanka R, Edwin B. Smart self-pollination for future agricultural-a computational structure for micro air vehicles with man-made and artificial intelligence. Int J Intell Syst Appl Eng 2022;10(2):170–4.
[126] Mandal A, Ghosh AR. AI-driven surveillance of the health and disease status of ocean organisms: a review. Aquac Int 2024;32(1):887–98. https://doi.org/10.1007/s10499-023-01192-7.
[127] Vardar N, et al. Autonomous network system with specialized and integrated multi-sensor technology for dynamic monitoring of marine pollution (SMARTPOL). J ETA Marit Sci 2023;11(3). https://doi.org/10.4274/jems.2023.61687.
[128] Angus J, Bouvier-Brown M, Robinson A. Solar powered uncrewed surface vehicles (USVs) for marine protected area (MPA) monitoring. In: OCEANS 2022, Hampton roads. IEEE; 2022. p. 1–5 [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9977292/.
[129] Gardazi NM, Daud A, Malik MK, Bukhari A, Alsahfi T, Alshemaimri B. BERT applications in natural language processing: a review. Artif Intell Rev 2025;58(6):1–49. https://doi.org/10.1007/s10462-025-11162-5.
[130] Nadeem MW, Goh HG, Hussain M, Hussain M, Khan MA. Internet of things for green building management: a survey. In: Role of IoT in green energy systems. IGI Global; 2021. p. 156–70.