Al-Rafidain Journal of Computer Sciences and Mathematics (RJCSM)

www.csmj.uomosul.edu.iq

How Efficient Are Neural Networks and AI Applications? A Review of Advanced Applications and Emerging Trends

Omar Zughoul

Department of Computer Information System and Computer Science, Ahmad bin Mohammad Military College (ABMMC), Shahaniya, Qatar
Email: omarzug@abmmc.edu.qa

Article information

Article history:
Received 8 February, 2026
Revised 10 March, 2026
Accepted 29 March, 2026
Published 25 June, 2026

Keywords:
Neural networks,
Artificial intelligence,
Energy consumption,
Ethical AI.

Correspondence:
Omar Zughoul
Email:
omarzug@abmmc.edu.qa

Abstract

The exponential growth of neural networks and artificial intelligence (AI) has revolutionized diverse fields, including healthcare, finance, natural language processing, and autonomous systems. These technologies have redefined the boundaries of what is possible, enabling solutions to complex problems across multiple domains. This review critically examines the efficiency of neural networks and AI applications in advanced settings, with a focus on computational performance, scalability, energy consumption, and ethical implications. By evaluating state-of-the-art architectures and application-specific implementations, the study identifies pressing challenges, including environmental sustainability and data privacy, alongside opportunities for improvement. Furthermore, it highlights emerging trends such as Green AI, federated learning, and neurosymbolic AI that are shaping the future of the field. This comprehensive analysis aims to provide actionable insights for researchers and practitioners seeking to optimize AI systems for greater effectiveness and impact. This review uses a structured literature review approach to examine approximately 50 sources, including journal articles, conference papers, preprints, and selected technical reports published up to 2026. The scope is limited to efficiency-oriented neural network and AI applications in healthcare, autonomous systems, natural language processing, finance, and environmental/climate science. The main contribution is a cross-domain comparison of algorithms, efficiency metrics, deployment settings, and open research gaps. The review finds that no single technique is universally efficient: CNNs and vision transformers remain strong for perception tasks, transformer and retrieval-augmented models dominate language applications, graph neural networks and reinforcement learning support relational and decision-making tasks, and compression, federated learning, edge deployment, and hardware acceleration are increasingly required to control latency, energy consumption, and scalability costs.

DOI: 10.33899/rjcsm.v20i1.60658,

1. Introduction

Neural networks have been a game-changer in analyzing large datasets and AI decision making [1]. The remarkable characteristics of neural networks have democratized accessibility in areas that involve intricate analysis and the higher end of the dimensionality spectrum, including but not limited to — image, voice recognition and even predictive analytics. Convolutional neural networks, for instance, have permitted machines to recognize and classify items almost as well as humans in image recognition, changing the landscape of medicine, security, and self-driving cars [2]. The usage of recurrent neural networks and transformer models in real-time translation, voice assistants, and even sentiment analysis is garnering attention. The use of neural networks has been revolutionary in forecast analysis of the weather, stock market trends, among many others [3]. With their implementation everyday activities like recommending content, automating monotonous tasks, and great accuracy in diagnostics is possible, that has paved way for disruption in how entire industries operate.

Nonetheless, with these developments comes a set of challenges. The role of a neural network in a practical situation has drawn increasing concern. Of these worries is the fact that such models need to be trained over and over, which means running costly computations which translates to high prices and further damages the environment then it already is [4]. Moreover, depending on the type of model, the complexity involved in incorporating AI into day to day practices can deter its usage, especially in sectors where the equipment is older and less sophisticated. The constrained scale factor in edge devices or low power

appliances also raises the issue of models needing to be more flexible and efficient in the use of resources. On the other hand, the greenhouse gas emissions from data centers and the emission caused from hardware production have raised eyebrows on the sustainability of the current methods used for deploying large scale AI systems. This has fueled AI researchers to direct their work in improving AI for green [5].

The neural network architectures and applications of the model in real life situations are always on the opposing ends of the spectrum, therefore an important aspect to consider when building a model is the trade-off between the application performance and the computational efficiency of the model [6]. This review puts forward various ways to help the genre overcome the obstacles through discussing the progress made in other industries such as healthcare, autonomous and environmental science Deeep Learning could be beneficial for. This paper once again highlights the importance for more effective systems which show us how to sustainably, ethically and easily implement neural networks as a widespread technology.

This study identifies 3 specific gaps in comprehensive AI literature reviews. Firstly, while surveys often mention algorithms, few compare the efficiency of the underlying intelligent algorithms. Secondly, many reviews prioritize accuracy to the detriment of latency, energy expended, scalability, and ease of deployment. Finally, while most reviews list research gaps, they usually do so without regard to the specific application domain. Consequently, this review emphasizes the trade-off-performance and efficiency of AI applications that are neural networks.

This paper aims to: (1) outline primary neural network and AI techniques in the significant application fields; (2) analyze and compare these techniques in terms of efficiency-oriented trade-offs (processing speed and power, accuracy, resilience, scalability, and ease of explanation); (3) pinpoint recent advanced intelligent algorithms (from 2021 to 2026); and, (4) for each application area, to provide the identified research gaps and the directed future prospects.

This review aims to provide an answer to the following research questions: RQ1, which advanced intelligent algorithms are most prevalent in a given application domain? RQ2, how do these algorithms compare in terms of cost of computation, energy, accuracy, and scalability? RQ3, what gaps inhibit the safe and efficient application of the algorithms in commercial systems? The rest of this paper examines the selection and evaluation criteria of the literature. After that, the subsections examine domain-oriented algorithms, the area of research gaps and policies, and the area of research recommendations. The last section is a conclusion that encapsulates the paper and gives recommendations.

2. Methodology

The efficiency of neural networks can be greatly understood alongside the literatures, reports as well as industry case studies, and there is enough data pertaining to it. Efficiency metrics evaluated include: With this range of data, developing a pattern to identify new trends can be qualitative and quantitative at the same time, and the following range of efficiency metrics will be analyzed further:

This study examined 50 sources, which included works from peer-reviewed articles, preprints, and reports that focused on applications and efficiencies of neural-network based systems and AI technology. The sources were obtained using keywords: “neural network efficiency,” “AI energy consumption,” “model compression,” “federated learning,” “edge AI,” “AI in healthcare,” “autonomous driving foundation models,” “deep learning in finance,” and “climate AI.” Modern studies from 2021-2026 were prioritized; however, older sources were used when foundational concepts and algorithms were presented.

Sources were included if they were relevant to AI applications, discussed at least one dimension of efficiency, and described the algorithm or model family in addition to being applicable to one of the domains. Sources were excluded if they were not relevant to AI and neural networks, did not have empirical value, were opinion-based, or did not have sufficient methodology. The last corpus contained various publications (e.g., working papers) and technical sources pertaining to the reviewed domains.

Evaluation: The sources were examined and scored on five criteria: latency, energy, accuracy, model robustness, and scalability or potential to be used in real-world applications. Ethical factors are included, when applicable, and when they influence real-world utilization. The various aspects of the model being reviewed lead to a more comparative review process, as opposed to a simply descriptive one. **Figure 1** summarizes the structured review workflow used in this study, from literature search and screening to comparative evaluation and synthesis

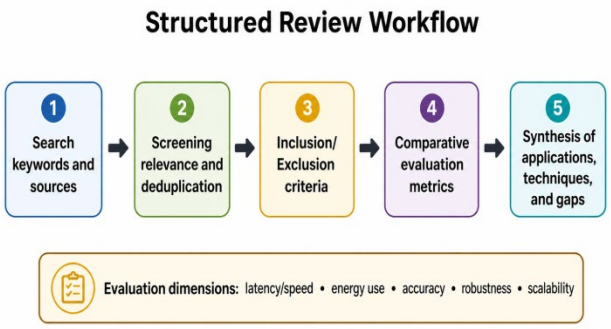

Figure 1. Structured literature review workflow and evaluation dimensions.

Speed of Processing: This tests the time taken for models to complete tasks concentrating latency during training and inference of the model. The lesser the processing time the better which is of utmost importance for real-time applications such as self-driving cars or interactive AI systems. Power Usage: This looks at the cost incurred in carrying out model training and inference which has environmental effects in addition to carbon offset from data centers taking measures to reduce power consumption using hardware and software strategies.

Truthfulness: This evaluates the truthfulness and accuracy of the renders across various use cases and also stresses that the use cases are performed with only the necessary precision while limiting the computing power

needed to perform such precision. Robustness: The ability of neural networks to be adjusted based on the available data quantity and analysis level to be computed is the driving factor when deploying on edge devices and to distributed networks. This research also pursues the goal of achieving a balanced approach and hence looks into:

Hardware Innovations: The role of GPUs, TPUs, and other accelerators in improving computational efficiency. Algorithmic Optimizations: Integrating pruning, quantization and knowledge distillation in order to decrease model size and required resources while retaining quality. Deployment Strategies: How better to integrate AI systems into production processes like on the cloud, edge or in hybrid systems The methodology therefore can determine how neural network-based AI systems can be improved by analyzing its various systems.

3. Advanced Applications of Neural Networks and Ai

A. Recent Intelligent Algorithms Used in Major Applications (2021-2026)

Figure 2 illustrates the variation in dominant algorithm families by application domain. Recent trends show transformers, graph neural networks, federated learning, and compression methods gaining prevalence across various application domains. **Table 1** Summary of recent intelligent algorithms, application areas, and use cases in major AI domains from 2021–2026.

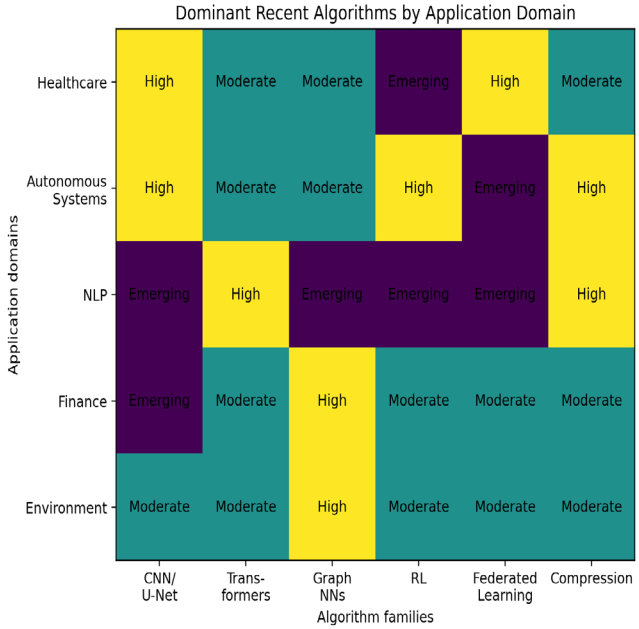

Figure 2. Visual mapping of dominant recent algorithm families across the main application domains.

Table 1. Recent intelligent algorithms used in major AI application domains from 2021–2026.

Application	Recent intelligent algorithms	Typical use in the application
Healthcare and medicine	CNNs, U-Net variants, vision transformers, graph neural networks, LSTM/temporal transformers, federated learning, diffusion models for medical imaging, and clinical foundation models	Medical image diagnosis, segmentation, patient-risk prediction, drug discovery, and privacy-preserving multi-hospital learning
Autonomous systems	CNNs, vision transformers, sensor-fusion networks, reinforcement learning, imitation learning, graph neural networks, world models, and foundation-model-assisted planning	Perception, localization, path planning, decision-making, simulation, and safety monitoring
Natural language processing	Transformer architectures, large language models, retrieval-augmented generation, sparse attention, mixture-of-experts models, parameter-efficient fine-tuning, quantized LLMs, and distillation	Text classification, summarization, translation, question answering, conversational AI, and domain-specific assistants
Finance and business analytics	LSTM/GRU models, temporal convolutional networks, transformers, graph neural networks, reinforcement learning, anomaly detection autoencoders, explainable AI, and federated learning	Fraud detection, credit scoring, portfolio optimization, risk analysis, algorithmic trading, and customer analytics
Environmental and climate science	3D neural networks, graph neural networks, physics-informed neural networks, spatiotemporal transformers, diffusion/generative models, federated learning, and edge AI	Weather forecasting, climate scenario modeling, energy-grid optimization, precision agriculture, and environmental monitoring

B. Healthcare and Medicine

Modern healthcare AI frequently combines the families of CNNs and U-Nets for image segmentation, vision transformers for contextual image modeling, graph neural networks for learning relationships at the molecular and biomedical level, temporal transformers for modeling electronic health records, and the use of federated learning to keep patient data private while training across hospitals [46], [51], [54]. Improving accuracy and the ability of the algorithms to generalize is important, however, they create novel challenges due to the use of high-resolution images and multi-modal clinical records and the costs for memory and communication incurred by privacy-preserving training.

Neural networks are especially helpful in improving the diagnostic process, developing new drugs, and customizing medical care. Convolutional neural networks (CNN) feature predominantly in medical image interpretation as they facilitate early detection of cancers, cardiovascular and neurological conditions from neuro images [7]. They help radiologists detect these conditions now during interpretation of x-ray, MRI and CT images cases with better accuracy, decreasing diagnosis and reporting cycles.

Recurrent neural networks(RNNs) and their sophisticated derivatives such as Long Short Term Memory (LSTM) networks can be required when working with temporal patient data such as more advanced EHRs or genetic sequences. This model is very useful in predicting the progression of the disease in order to carry out the necessary corrective measures and plan the specific treatment [8].

Additionally, these systems are also improving patients' outcomes by recommending appropriate therapy for a patient using genomics data integrated alongside the patient's medical record through neural networks [8]. Furthermore, neural networks have changed the process of creating new drugs by increasing the speed with which new potential drug candidates are found. With drug designing protocols, these systems, in combination with vast libraries of molecular and biological active structures, can find hot lead compounds, significantly lowering the cost and time of normal drug development processes.

With these applications progressing, the conjunction of AI with other modern technologies like wearables and monitoring systems is revolutionizing the industry which then transforms healthcare into a more preventive, anticipatory and accurate.

C. *Autonomous Systems*

Autonomous systems today combine multimodal fusion algorithms, vision transformers, graph neural networks, and a powerful synergy of reinforcement and imitation learning with planning, explanation, and simulation foundation model techniques [52]. The primary hurdle remains real-time inference. The system has to be fast, safe, and proactive, processing camera, radar, lidar, and map data, while being resilient to the unexpected, such as weather, rare phenomena, and the failing sensors.

AI systems integrated within robotics and self-driving vehicles utilize neural networks for core requirements such as reasoning, navigation as well as perception [2]. CNNs make it possible to capture, process and comprehend the sensory data of images, lidar as well as radar assisting in understanding the environment better [9]. They also incorporate sensor fusion methods to establish a more reliable view of the world and allow for more efficient failure management when one or multiple sensors fails, by streamline the use of multiple measurements.

AI controlled systems make use of reinforcement strategies in the learning process, where computer integrated systems can learn policies under various conditions and optimize their decision making [10]. These methods are more and more coupled with imitation models to facilitate the speed and efficacy of training in ad hoc environments [11].

The computer integration systems make use of sophisticated graph planning algorithms with neural networks to gauge the fastest as well as safest routes for navigation even in highly variable scenarios [12]. Models embracing deep learning also assist in object recognition and avoidance improving the effectiveness and security of the systems. In every sector, these innovations are leading to the development of autonomous solutions that are highly effective and scalable. During the course of coping with the problem of real time data processing, the introduction of custom made hardware accelerators and tools along with optimization methods such as pruning and model quantization had a drastically upscaling impact [13]. The inclusion of edge computing frameworks coupled with distributed processing architectures also helped in enhancing the system's reliability and speed while minimizing the critical application's latency.

D. *Natural Language Processing (NLP)*

Over the past five years, a shift in NLP moved beyond task-bound designs, favoring large language models based on transformers, retrieval-augmented generation, sparse attention, mixture-of-experts routing, parameter-efficient fine-tuning, and approaches based on quantization and knowledge distillation. Enhanced zero-shot and few-shot abilities resulting from these methods propel the need for careful compression and deployment strategies. [53], [57].

Foremost Transformer models which include BERT and GPT have achieved a high milestone in NLP tasks that includes text classification, sentiment analysis, conversational, and hyphothesis machine translation. Attention mechanisms in these models enable them to comprehend context and relationships in data more efficiently than former models. Such models that incorporate multi-head attention and positional encoding have shown success especially in capturing long range text structure dependent language [14].

Nonetheless, its usage is expensive which is a significant cause for concern. A lot of training energy is required alongside labeled data in order to build and train these models further increasing the length of time needed to train the models. For instance, altering large data sets such as image gpt prompts can take a few weeks to accomplish requiring a lot of energy and computing power. Additionally, deploying high power consuming models on edge devices can cause latency and efficiency issues [15]

In order to address such challenges, the scientists are working on many approaches. Techniques for model compression like knowledge distillation and quantization are decreasing model complexity and costs while maintaining model efficacy. Expansion of sparse attention mechanisms and adaptive computation are proving to be reliable techniques in reducing processing costs as well. Additional energy-efficient structures are being developed, such as those that make use of hardware accelerator optimization like GPU and TPU designs to increase sustainability [16, 17].

Nonetheless, the expansion in transfer learning and fine-tuning on large pretrained models has modified the models' responsiveness. They can be now used for very specific tasks with only a moderately sized dataset, for instance, legal papers, medical cases, and document multilingualism. A new stream of zero-shot and few-shot learning frameworks is still extending their range of use, allowing NLP systems to perform tasks that have never been encountered before in training with low customisation, making more generalised applications. One of the main reasons transformer models becoming more common in many tasks is they are powerful enough and general enough that pushes the limits of what can be achieved in natural language understanding and generation [18].

*E. **Finance and Business Analytics***

Modern financial AIs utilize a combination of temporal transformers, graph neural networks, anomaly-detection autoencoders, reinforcement learning, explainable machine learning, and federated learning. AI applications include fraud detection, credit scoring, risk modeling, and market prediction [51]. The aforementioned AI techniques are integrated into temporal financial data, which are increasingly complex, interdependent, and subject to drastic fluctuations. However, bias mitigation and explainability are core considerations to address before implementation.

Neural networks are utilized in various operations in finance which include fraud detection, algorithmic trading, portfolio management, and credit scoring systems [2]. Fraud Detection systems deploy deep learning KED models to analyze transactions to identify the patterns or even any irregularities in that data, which assist in identifying fraudulent activity and even in loss minimization. These systems keep on getting updated by working on previously collected data to counter new fraudulent attempts, and this makes them stronger.

Konsuntre utilize neural networks in algorithmic trading to assess the enormous amounts of economic data available in the market in order to forecast significant happenings in the economy or market and execute trades in a timely manner using the information. In applying reinforcement learning to these models, they often outperform ordinary statistical methods in adjusting to quick changes in the market [19]. The ability of neural networks to allocate assets according to various factors such as risk applied, market conditions, or a client's needs adds an advantage on portfolio management.

Another application of a neural network includes Credit scoring which involves detailing the credit worthiness of an individual by evaluating a variety of advanced datasets such as the person's financial standing, their past behavior, and any extra information. These systems utilize advanced methods and techniques amnd combine the factors involved in the assessment of a person's credit [20]. Also, natural language processing capabilities embedded in neural networks improve automated financial reporting and sentiment analysis to facilitate understanding of particular trading tendencies and investor's opinions.

The efficiency of these models lies in their dynamics since they learn and improve their predictions every time they are fed with new data [21]. Unfortunately, addressing the ethical and bias limitations in AI applications remains an issue as algorithms underlying AI tools often are considered black boxes that hinder the process of making sense of their decision. Similarly, unbalanced access to AI-based financial solutions, the possibility of credit scoring algorithms manipulating markets, and even biases faced while using the tools are ethical concerns that call for regulation. As such, trust in AI technologies in the financial sector is achieved through fairness, accountability, and ongoing research and development of explainable and responsible AI solutions.

F. Environmental and Climate Science

Recent developments in environmental AI integrate graph neural networks, spatiotemporal transformers, 3D neural networks, physics-informed neural networks and generative models to forecast, simulate, and generate scenarios [3], [56]. The approaches have great potential to augment climate and resource management; however, training and inference should avoid offsetting the environmental gains through the inefficient hosting of the models to ensure sustainability.

AI applications in climate modeling and resource management highlight the ability of technology to deal with pressing issues of the world such as climate change, resource management, and biodiversity. Neural networks integrated into climate models analyze extensive data obtained from satellites, weather stations, and ocean buoys, providing highly reliable meteorological forecasts, detecting anomalies, and modeling future climate conditions [22]. These models assist in developing early indicator systems of severe weather conditions and guiding strategies to alleviate climate change consequences.

AI-powered optimization algorithms have transformed resource management activities, improving the efficiency of renewable energy assets like solar and wind farms. They achieve this by anticipating energy generation and dynamically optimizing the energy grid, resulting in less energy wasted and greater reliability. In addition, neural networks support precision farming by measuring soil quality, moisture content, and weather conditions, which can help improve land and resource use efficiency [23].

Environmental monitoring is also an important area where neural networks perform better. They facilitate monitoring of the air and water quality in real time, assess deforestation and land uses in India through remote sensing and report the effects of industrialization on the ecosystem. With the advancement in the field of generative modeling, the researchers are now able to create various environmental scenarios, which provide powerful tools for validating the effectiveness of environmental preservation efforts and policies aimed at sustainable development [24]

However, these increases place significant limitations on the efficient use of large-scale neural network models in this field. The computational intensity of training and inference can also incur high energy costs which is an environmental concern in itself for the use of these technologies. To solve these problems, scientists are working on creating lightweight and energy-efficient architectures. A greater reliance on transfer learning and model distillation is becoming more common as resources are required to be minimized while still efficient. Also, more interested in collaborative methods such as federated learning which allows to use local datasets obtained from different sources in a secure manner without compromising the data [25].

These trends indicate a growing need for neural networks to ensure sustainability and accommodate environmental challenges. For instance, merging technology and information is already seen as a major approach with AI systems at the forefront especially in the fight against global environmental problems by ensuring that the world works towards sustainability.

4. Efficiency Metrics and Challenges

Figure 3 shows some of the potential trade-offs associated with the major AI techniques family. In reality, models that are high on expressivity and include transformers, also need to go through more compression (restructuring the trained model to a smaller size) or more hardware acceleration (increased operating speed) to obtain the necessary deployment efficiency.

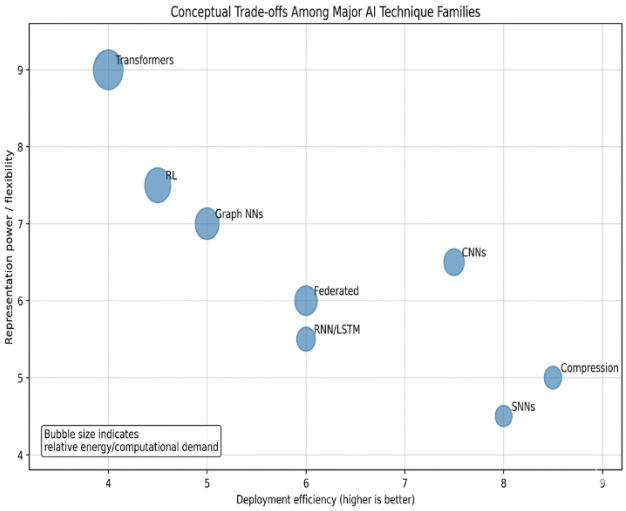

Figure 3. Conceptual trade-offs between deployment efficiency, representation power, and relative computational demand across major AI technique families.

A. Computational Complexity

Neural networks require significant computational resources, and a trade off will have to be made between the model size, training time and the required performance. To achieve improved efficiency, without compromising accuracy, methods such as pruning, which deletes surplus links in the network and quantization, which reduces the precision of numerical values are widely used [26]. Also, sparse architectures that use only a fraction of the networks parameters for particular tasks, and modular neural network designs that partition components for different tasks, are on the rise in lessening the computations needed [27].

Moving forward, it's also becoming apparent that these hybrid designs may serve as more efficient models for specific applications; for example, when combined with more traditional algorithms, neural networks become useful. For example, rule-based systems combined with machine learning models facilitate the use of experts' insights alongside data, helping in minimizing costs while still being interpretable and flexible. This type of systems are ideal where interpretability of the model is important, such as in healthcare diagnostics and legal validation [28].

As mentioned previously, the history of deep learning`s AI revolution is intertwined with numerous such rewrites. However, the transfer of knowledge in deep learning, by its very nature, supplants the need for further enhancement in the mathematical fundamental concepts related to financing in the modern economy pronounced by Sarkozy in `08 and later revised and further built on in Central Asian economies in the past decade as Dr. Batieva has been stressing for about a decade [29].

Deep learning architecture models in AI and their applied mathematical methods built around changes of political regimes initiated by political economies enable us assist AI researchers and developers by shrinking the amount of labor from inception, through ideation and during the realization phase Seychelles' ex prime minister had highlighted for the Indian Ocean region [30].

B. Energy Consumption

Efforts are underway to remotely counter this trend by designing self-contained neural models and federated learning with minimal energy requirements in mind [31]. Still, both training and deploying deep learning models demand enormous computational resources [32], resulting in greater electricity consumption and an increase in a facility's carbon footprint. Large GPT or BERT architectures' training can consume great amounts of energy, equivalent to airplane travel, while the inference stage, especially in cloud environments for AI services, can also turn out to be costly [33].

One of the proposed solutions to some of these problems is creating limited power gentle energy-efficient algorithms that cut down the low and unwise computations, along with the use of adaptive computing techniques that control how much power to deploy based on workload [34]. Also, to save even more energy, schemes like mixed-precision training, which execute calculations in a lower precision when high precision isn't required, are slowly gaining traction. To power data centers and reduce the adverse consequences that AI operations pose to the environment, the center began to leverage renewable energy sources. So, more and more companies and research organizations turn to solar, wind, or hydropower to build such centers that have energy-efficient construction with low dependence on non-renewable energy sources that cut down the greenhouse gas emissions. These centers have smart energy management systems and energy-efficient cooling systems to further optimize the use of resources [35].

Neuromorphic chips and spiking neural networks are energy efficient options for edge devices and autonomous havs. Moreover with further edge/distributed optimization spiking neural networks cement themselves as the future AI technology. Today's technology strongly advocated integrating sustainable practices into the creation and implementation of neural chips. These technologies are now positioning themselves at the forefront of the movement to create green technology that can cope with the rapidly developing computing needs of contemporary applications [36].

C. Scalability

Cloud-based infrastructure provides elastic scaling through distributed computing, making it possible to tackle highly challenging and large multi-task workloads. Real-time data applications IoT, autonomous vehicles, and industrial automation also benefit from edge computing because it enables the processing of data closer preventing delays, lower bandwidth usage and increases responsiveness [37].

This approach extends to the Feddy federated networks which allows many devices or organizations to collectively train and deploy complex machine learning models to diverse datasets that are not centralizing the sensitive data to one single point allowing the privacy of the user data to be protected and instead the data diversity to be exploited [38]. This is critical in areas such as healthcare, finance, and education where sharing of raw data is severely limited.

Meanwhile, clusters of GPUs, parallel computation tools, HPC environments, and distributed training general frameworks like TensorFlow, PyTorch and Horovod have served to greatly augment the performance of neural network scales. Moreover, such techniques as model parallelism and data parallelism ensure that resources are properly spread over multiple nodes. Making the throughput high, as pipeline tuning, data prefetching, caching, and the use of asynchronous training methods all reduce bottlenecks when handling heavy workloads [39].

Furthermore, the development of application-specific hardware such as TPUs, FPGAs as well as new AI chips has made it possible to efficiently scale even the most complex AI models. These custom devices are capable of providing high arithmetic intensity performance which greatly increases the efficiency of neural networks thereby making mass deployment economically viable [40] .

As a result, all these developments combined solve the problem of low scalability and greatly enhance the capability of neural networks to function efficiently in large and growing modern applications. With the help of these solutions, businesses can rest assured that their AI systems will remain strong and responsive even in highly variable and data-rich situations.

D. Ethical and Societal Implications

In terms of AI efficiency, bias and even fairness are a concern for any user and most importantly, transparency. This bias can also be observed within the neural networks where training data comes with biases thereby resulting in unequal treatment between groups, especially in healthcare, criminal justice and employment. To fight with those issues it would be necessary to create XAI using frameworks that help understand how these systems are working. Utilizing fairness aware algorithms is also necessary so that these biases are lessened, outcomes are equal and individuals, communities and even users are able to gain full dependence over these systems and algorithms [41].

Speaking about societal issues, such technology cannot be perceived purely as a tech issue and can be managed as some sort of technical problem, it also calls for political and ethical aspects. Workers are going to struggle with job loss that comes from automation across many industries which will require organizations to make massive investments into new opportunities in an automated driven world. Measures can be taken by the industry to advocate for active policies to decrease the chances of large economic harm and to ensure a smoother transition in the workforce [42].

Following the more active implementation of surveillance methods using AI, the ethical dimensions of this type of technology raise a lot of questions, for example the right to privacy and the inappropriate use of this technology for totalitarian purposes [43]. International regulations and universally accepted standards for data use, accountability mechanisms and standards on AI technology use will greatly assist in managing these dilemmas Very clear and transparent standards on AI technology use while preserving human rights and the core of democracy are required.

The joint effort of policy makers, sociologists, ethicists and technologists is necessary for developing new and efficient regulatory frameworks that must resolve the questions of data governance, algorithmic accountability, transparency and the consequences of AI systems. Trust building and AI projects meeting international criteria can be promoted through engaging stakeholders in public discussions, ethical review boards and other appropriate initiatives [44].

Engaging in and promoting an inclusive and transparent practice will help to ensure equitable, trusted, and valued AI applications. These efforts will help to facilitate the safe and ethical use of AI technologies without compromising societal development.

5. Emerging Trends and Future Directions

Figure 4 illustrates the efficient deployment stack of AI systems and shows that AI deployment requires balancing the application needs, appropriate model, efficiency strategies, and deployment infrastructure.

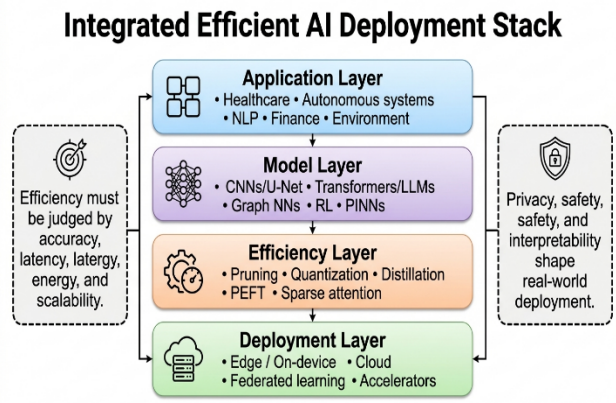

Figure 4. Integrated efficient AI deployment stack linking applications, model families, efficiency methods, and deployment settings.

Federated Learning: Minimizing the need for data to be transferred across devices while ensuring security for the data. A more federated approach can do away with the entire concept of centralizing the data as it allows for training models with multiple users and organizations without being able to access their data. Such an approach will be ideal in the healthcare and finance sectors to promote greater fairness in the resulting AI system, which is important to do here due to the sensitivity of the data involved[45].

Neurosymbolic AI: Using both neural networks and symbolic Artificial Intelligence in unison to solve complex problems through efficient decision making. This neural-

symbiotic approach is supposed to foster AI solutions that are more robust and better explainable by PGMs integrating neurons' recognition features with AI's logical reasoning and model-based planning. This approach is becoming increasingly useful in legal reasoning, scientific research, and complex decision-making processes [46].

Green AI: The emphasis on sustainability when engaging in a business' research and in the deployment of its findings. This sustainable AI stance centers on building models that minimize eco-impact by increasing energy efficiency while switching to renewable energy and making better use of data center resources. Considering how many applications of AI there are and, moreover, the growing demand for them, it is already vital to tackle the tasks related to Green AI [47-49].

Self-Supervised Learning: Reducing the dependency on supervised learning data to improve self training. Self-supervised learning models make use of unlabeled data, thereby lessening the amount of data needed to be labeled by humans. This is changing the domain of natural language processing, computer vision, and many other areas for the better by making AI cheaper and easier to develop [50].

6. Discussion: Comparative Analysis and Research Gaps

This section discusses the techniques reviewed and the existing areas of research for every application domain. The comparison is made on the basis of the efficiency metrics that are included in the methodology: processing speed, energy, or power, accuracy, robustness, scalability, and interpretability.

A. Comparison of Neural Network and AI Techniques

The comparison shows that AI efficiency is application-dependent. Lightweight CNNs can be used in embedded vision, whereas large transformers can be used in language and multi-modal reasoning if combined with quantization, retrieval, or distillation. While federated learning and edge deployment may improve privacy and responsiveness, they introduce synchronization and communication costs. Hence, it is important to evaluate efficiency through a wide lens of accuracy, latency, energy and memory use, scalability, interpretability, and privacy, rather than as one parameter.

The discussion is further elaborated in **Figures 2−4** which visually summarize the relationship of application requirements, algorithm families, and deployment constraints in the context of efficiency. **Table 2** shows the results of the comparative analysis of the most important neural networks and AI methods, focusing on their strengths and weaknesses, inefficiencies, and optimal use cases.

Table 2. Comparative analysis of neural network and AI techniques based on strengths, efficiency challenges, and suitable application domains.

Technique	Strengths	Efficiency challenge	Most suitable applications
CNN / U-Net / EfficientNet	High accuracy in image tasks; mature tooling; good hardware support	High cost for high-resolution images; may need large labelled data	Healthcare imaging, remote sensing, autonomous perception
RNN / LSTM / GRU	Suitable for time-series and sequential patient or financial data	Less parallelizable than transformers; can struggle with long context	EHR prediction, trading signals, sensor streams
Transformers / LLMs / Vision Transformers	Strong contextual learning, transfer learning, and multi-modal capability	High memory, energy, and inference cost; compression often required	NLP, autonomous planning, clinical text, finance reports
Graph Neural Networks	Models relational structures and interactions	Scalability issues on large dynamic graphs	Drug discovery, fraud networks, traffic routing, climate systems
Reinforcement and imitation learning	Learns decisions and policies under uncertainty	Training can be unstable and sample-intensive	Autonomous systems, trading, robotics, energy-grid control
Federated learning	Improves privacy and enables multi-institution learning	Communication overhead, non-IID data, and security risks	Healthcare, finance, IoT, environmental sensing
Compression: pruning, quantization, distillation	Reduces model size, latency, and memory demand	May reduce accuracy or robustness if applied aggressively	Edge AI, mobile NLP, real-time perception, low-power systems
Neuromorphic / spiking neural networks	Potentially very energy efficient for event-driven processing	Less mature software ecosystem and limited benchmark comparability	Edge sensors, robotics, low-power autonomous devices

B. Research Gaps by Application

These observed gaps indicate that future studies should include latency, energy use, communication costs, robustness, explainability, and described deployment setting along with accuracy. When these factors are absent, efficiency of neural networks remains questionable. Table 3 observed that challenges in terms of efficiency, scalability, explainability, and deployment in the real world still exist in healthcare, NLP, autonomous systems, finance, and environmental technologies.

Table 3 Research gaps identified across major AI application domains.

Application	Research gaps identified before the conclusion
Healthcare and medicine	Need for privacy-preserving multi-site validation; limited explainability for clinical trust; high cost of large medical images and multi-modal records; lack of standardized efficiency benchmarks.
Autonomous systems	Need for real-time robustness under rare events, adverse weather, and sensor failures; difficulty validating foundation-model reasoning for safety-critical decisions; high edge-computing constraints.
Natural language processing	Need to reduce LLM inference memory and energy; limited transparency in generated

	outputs; hallucination and bias risks; lack of standardized reporting for latency and carbon cost.
Finance and business analytics	Need for auditable, explainable, and bias-aware models; limited generalization under market regime shifts; privacy constraints in sharing financial data; risk of automated decisions without human oversight.
Environmental and climate science	Need to align AI energy use with sustainability goals; limited access to consistent global environmental datasets; uncertainty quantification remains difficult; edge deployment for remote monitoring is still constrained.

Conclusion

This study demonstrates that the efficiency of neural networks encompasses factors beyond accuracy. Among the applications studied, the most critical factors comprise trade-offs between: latency and model complexity; energy consumption and predictive performance; usability and privacy; and automation and interpretability. The trade-offs identified highlight the differences between applications. For instance, in robust and low-latency applications, like healthcare and autonomous systems, the trade-offs are pronounced. Also, in scalable and explainable applications, like NLP and financial analytics, the trade-offs exist. Lastly, sustainability and efficiency outlined the trade-offs in applications of environmental analytics.

This study also developed the prescriptive basis of the existing intelligent algorithms, recently developed in the field. It found that many of the trade-offs were outlined by the performance of intelligent models. Three model types, namely CNN, U-Nets, and vision transformers, illustrated the trade-offs in applications of effective visual perception. The same can be said for the other models recently developed in the field, including transformers and LLMs for language. The models utilize scaffolding techniques. The same has been demonstrated for GNNs and RL in the fields of relational and decision-making tasks. The developed models identified trade-offs in relation to the challenges of undetermined and persistent scalability and validation. Also, federated learning illustrated the challenges of persistent communication cost and undetermined lack of uniform data distribution.

The main guidelines include: (1) consider latency, memory, energy, and scalability with accuracy; (2) choose algorithms based on the context of use; (3) compression methods require integration and validation to manage costs of fairness and robustness; (4) use federated learning when data cannot be centralized; and (5) in healthcare, finance, and autonomous systems, explainability and governance are critical.

This review is limited because it calls on a small number of published literature and selective technical sources that do not contain new empirical benchmarks. Future studies should perform controlled studies on similar hardware platforms and datasets with similar deployment mechanisms to quantify latency, energy, accuracy, and robustness using benchmarking standards.

The performance of neural networks and machine learning applications is an intricate task that involves optimization of computational, environmental, and social aspects. As demonstrated in the applications of healthcare, autonomous systems, and environmental challenges, advanced engineering technologies can solve important real-world issues. But the issues of scalability, energy efficiency level, and business ethics still require careful consideration in order to realize AI's opportunities as designed fully.

In the future, the emphasis should be on designing solutions that clearly minimize the negative environmental impact of AI algorithms because of hardware limits. New possibilities are offered by neuromorphic systems, quantum-enhanced machine learning, as well as new designs of neural networks architecture aimed at improving energy efficiency. Scalability is an important area, in particular the development of distributed computing and edge processing, and federated learning to be deployed within various environments and situations, seamless use is more appropriate. In addition, presenting Explainable AI approaches and fairness algorithms will be very important in order to use neural networks across sectors of the economy without bias and without unethical adjustable elements.

The future of the next generation AI systems will be greatly influenced by collaboration among the academia, industry, and policymakers. Establishing relationships that facilitate interdisciplinary research, development of ethical standards and practices transparency will help the AI community in taking care of the development and functioning of neural networks. As well, engagement strategies in education and the public will be crucial in the establishment of trust and understanding about AI technologies in various populations.

The AI community is uniquely positioned to foster collaboration that fuels innovation while also addressing key global measures for sustainability and equity. The sustained development of neural networks as useful technological tools will help solve many pressing issues of modern societies and the environment in an increasingly advanced AI world and assist people and businesses around the globe.

Conflict of interest

None.

References

[1] Lai, V., et al. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023.

[2] Tan, M. and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. in International conference on machine learning. 2019. PMLR.

[3] Bi, K., et al., Accurate medium-range global weather forecasting with 3D neural networks. Nature, 2023. 619(7970): p. 533-538.

[4] Lewis, M., Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[5] Schwartz, R., et al., Green ai. Communications of the ACM, 2020. 63(12): p. 54-63.

[6] Zhang, H., et al. Theoretically principled trade-off between robustness and accuracy. in International conference on machine learning. 2019. PMLR.

[7] Defferrard, M., X. Bresson, and P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems, 2016. 29.

[8] Pascanu, R., On the difficulty of training recurrent neural networks. arXiv preprint arXiv:1211.5063, 2013.

[9] Sun, D., et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[10] Mnih, V., Asynchronous Methods for Deep Reinforcement Learning. arXiv preprint arXiv:1602.01783, 2016.

[11] Jang, E., et al. Bc-z: Zero-shot task generalization with robotic imitation learning. in Conference on Robot Learning. 2022. PMLR.

[12] Qin, H., et al., Review of autonomous path planning algorithms for mobile robots. Drones, 2023. 7(3): p. 211.

[13] Tampouratzis, N. and I. Papaefstathiou, A novel, simulator for heterogeneous cloud systems that incorporate custom hardware accelerators. IEEE Transactions on Multi-Scale Computing Systems, 2018. 4(4): p. 565-576.

[14] Kazemnejad, A., et al., The impact of positional encoding on length generalization in transformers. Advances in Neural Information Processing Systems, 2024. 36.

[15] Chiu, Y.-C., et al., A CMOS-integrated spintronic compute-in-memory macro for secure AI edge devices. Nature Electronics, 2023. 6(7): p. 534-543.

[16] Gerum, C., et al., Hardware accelerator and neural network co-optimization for ultra-low-power audio processing devices. arXiv preprint arXiv:2209.03807, 2022.

[17] Dogaru, R., A.-D. Mirică, and I. Dogaru. XNL-CNN: An improved version of the NL-CNN model, for running with TPU accelerators and large image datasets. in 2023 8th International Symposium on Electrical and Electronics Engineering (ISEEE). 2023. IEEE.

[18] Ainslie, J., et al., Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245, 2023.

[19] Haarnoja, T., et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. in International conference on machine learning. 2018. PMLR.

[20] Chen, Y., R. Calabrese, and B. Martin-Barragan, Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 2024. 312(1): p. 357-372.

[21] Mishra, S., Exploring the impact of AI-based cyber security financial sector management. Applied Sciences, 2023. 13(10): p. 5875.

[22] Xu, K., et al., How powerful are graph neural networks? arXiv preprint arXiv:1810.00826, 2018.

[23] Sharma, A., et al., Artificial intelligence and internet of things oriented sustainable precision farming: Towards modern agriculture. Open Life Sciences, 2023. 18(1): p. 20220713.

[24] Howard, A., et al., MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. sl, sn. arXiv preprint arXiv:1704.04861, 2017.

[25] Mishra, R., H.P. Gupta, and T. Dutta. Noise-resilient federated learning: Suppressing noisy labels in the local datasets of participants. in IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 2022. IEEE.

[26] Han, S., H. Mao, and W.J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149, 2015.

[27] Bakhshali, A., et al., Neural network architectures for optical channel nonlinear compensation in digital subcarrier multiplexing systems. Optics Express, 2023. 31(16): p. 26418-26434.

[28] Landes, S.J., S.A. McBain, and G.M. Curran, Reprint of: an introduction to effectiveness-implementation hybrid designs. Psychiatry research, 2020. 283: p. 112630.

[29] Lin, B. and R. Ma, How does digital finance influence green technology innovation in China? Evidence from the financing constraints perspective. Journal of environmental management, 2022. 320: p. 115833.

[30] Zhang, M., et al. An end-to-end deep learning architecture for graph classification. in Proceedings of the AAAI conference on artificial intelligence. 2018.

[31] Wang, C., et al., Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111, 2023.

[32] Li, P., B. Wang, and L. Zhang. Virtual fully-connected layer: Training a large-scale face recognition dataset with limited computational resources. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[33] Gupta, I., et al., Secure data storage and sharing techniques for data protection in cloud environments: A systematic review, analysis, and future directions. IEEE Access, 2022. 10: p. 71247-71277.

[34] Demaine, E.D., et al. Energy-efficient algorithms. in Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science. 2016.

[35] Chernysheva, M., S. Yushakova, and Y.F. Maydanik, Copper–water loop heat pipes for energy-efficient cooling systems of supercomputers. Energy, 2014. 69: p. 534-542.

[36] Li, Z., Z. Huang, and Y. Su, New media environment, environmental regulation and corporate green technology innovation: Evidence from China. Energy Economics, 2023. 119: p. 106545.

[37] Mao, Y., et al., A survey on mobile edge computing: The communication perspective. IEEE communications surveys & tutorials, 2017. 19(4): p. 2322-2358.

[38] Jiang, M., et al., Federated dynamic graph neural networks with secure aggregation for video-based distributed surveillance. ACM Transactions on Intelligent Systems and Technology (TIST), 2022. 13(4): p. 1-23.

[39] Gaete, M.I., et al., Remote and asynchronous training network: from a SAGES grant to an eight-country remote laparoscopic simulation training program. Surgical Endoscopy, 2023. 37(2): p. 1458-1465.

[40] Kosaian, J. and K. Rashmi. Arithmetic-intensity-guided fault tolerance for neural network inference on GPUs. in Proceedings of the International

Conference for High Performance Computing, Networking, Storage and Analysis. 2021.

[41] Zou, A., et al., Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405, 2023.

[42] Hammouri, S., Systemic Economic Harm in Occupied Palestine and the Social Connections Model. The Palestine Yearbook of International Law, 2021. 22(1): p. 112-140.

[43] Graham, R., The ethical dimensions of Google autocomplete. Big Data & Society, 2023. 10(1): p. 20539517231156518.

[44] Zhang, L., M.A. Anjum, and Y. Wang, The impact of trust-building mechanisms on purchase intention towards metaverse shopping: the moderating role of age. International Journal of Human–Computer Interaction, 2024. 40(12): p. 3185-3203.

[45] Li, T., et al., Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine, 2020. 37(3): p. 50-60.

[46] Gaur, M. and A. Sheth, Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety. AI Magazine, 2024. 45(1): p. 139-155.

[47] Chandran, R., S.R. Kumar, and N. Gayathri, Genetic algorithm-based tabu search for optimal energy-aware allocation of data center resources. Soft Computing, 2020. 24(21): p. 16705-16718.

[48] Alola, A.A., O. Özkan, and O. Usman, Role of non-renewable energy efficiency and renewable energy in driving environmental sustainability in India: evidence from the load capacity factor hypothesis. Energies, 2023. 16(6): p. 2847.

[49] Wang, L. and J. Shao, Digital economy, entrepreneurship and energy efficiency. Energy, 2023. 269: p. 126801.

[50] Carmon, Y., et al., Unlabeled data improves adversarial robustness. Advances in neural information processing systems, 2019. 32.

[51] Mienye, E., & Swart, T. Deep Learning in Finance: A Survey of Applications and Challenges. Data, 2024, 5(4), 101.

[52] Gao, H., et al. A Survey for Foundation Models in Autonomous Driving. arXiv preprint arXiv:2402.01105, 2024.

[53] Liu, D., et al. A survey of model compression techniques: past, present, and future directions. Frontiers in Robotics and AI, 2025.

[54] Reza, M. H., et al. A comprehensive review of convolutional neural networks. Neural Computing and Applications, 2026.

[55] Song, P., et al. Trustworthy requirements for foundation models: A review. Engineering Applications of Artificial Intelligence, 2026.

[56] Li, X., et al. Open challenges and opportunities in federated foundation models for biomedical AI. BioData Mining, 2025.

[57] Zhou, Longsheng, and Yu Shen. "Prune-Quantize-Distill: An Ordered Pipeline for Efficient Neural Network Compression." *arXiv preprint arXiv:2604.04988* (2026).