IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Role of Deep Learning in Artificial Intelligence: A Systematic Review of Architectures, Applications, Challenges and Future Directions

Pooja Patankar[1], Rasika Lonkar[2], Samruddhi Khalokar[3], Mr. Sanghesh Bele[4]
Students, Department of Master of Computer Applications[1,2,3]
Guide, Department of Master of Computer Applications[4]
Vidya Bharti Mahavidyalaya, Amravati
Sant Gadge Baba Amravati University

Abstract: *Deep learning (DL) has become the leading approach in artificial intelligence (AI) because multilayer neural networks learn complex patterns directly from large amounts of data. This paper reviews the role of DL in AI through a systematic, qualitative literature review of academic publications. It describes the relationship between AI, machine learning and DL, and examines the major architectures: convolutional neural networks, recurrent and long short-term memory networks, Transformers, generative models and state space models. It then analyses applications in computer vision, speech and language, healthcare, finance and cybersecurity, robotics and recommendation systems. The review finds that no single architecture suits every task, and that the strongest results carry the largest caveats. DL still depends on large datasets and heavy computation, and it faces open problems in interpretability, bias, privacy, security and energy use. Emerging directions include foundation and multimodal models, explainable AI, edge AI and efficient sequence architectures. The paper closes with research gaps and recommendations for building efficient, reliable, explainable and responsible DL systems.*

Keywords: Artificial Intelligence, Deep Learning, Machine Learning, Neural Networks, CNN, RNN, LSTM, Transformers, Explainable AI.

I. INTRODUCTION

Deep learning (DL) has become the dominant approach for building high-performing artificial intelligence (AI) systems since 2012, when a deep convolutional network won the ImageNet image-classification challenge by a wide margin over conventional methods [4]. Artificial intelligence is the broad field concerned with computational systems that perform tasks normally requiring human intelligence, such as perception, language understanding, reasoning, prediction and decision-making. Machine learning (ML) is the part of AI in which systems learn patterns from data instead of relying only on hand-written rules [7]. Deep learning is a subfield of ML that stacks many layers of artificial neurons so that a model can learn increasingly abstract representations directly from raw data [1], [2].

Three developments explain the speed of this progress: large labelled and unlabelled datasets, parallel hardware such as graphics processing units (GPUs), and architectural advances including deep convolutional networks, residual connections [5] and the Transformer [6]. As a result, DL now underpins image recognition, speech processing, natural language processing (NLP), medical imaging, fraud detection, recommendation systems and autonomous driving. The same progress has raised concerns about data hunger, computational and energy cost, opacity, bias, privacy and security, which limit how safely DL can be deployed.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Although many surveys cover individual architectures or single application areas, students and practitioners often need one consolidated account that links architectures, applications, limitations and research directions. This paper provides that account.

1.1 Aim and objcctivcs

The aim is to examine the role of deep learning within artificial intelligence. The objectives are to:

1. Describe the concepts and evolution of AI, ML and DL.
2. Examine the main DL architectures: CNNs, RNNs, LSTMs and Transformer-based models.
3. Analyse applications of DL across major domains.
4. Evaluate the advantages and limitations of DL approaches.
5. Identify challenges in data requirements, computation, interpretability, privacy, security and bias.
6. Identify future research opportunities for efficient, reliable, explainable and responsible DL systems.

1.2 Scope

The review covers the fundamentals of AI, ML and DL; artificial neural networks; CNNs, RNNs, LSTMs and Transformers; generative and multimodal DL; applications in computer vision, NLP, healthcare, finance, cybersecurity, robotics, autonomous systems and recommendation systems; and explainable and responsible AI. It is a secondary-source study. It does not report new experiments or benchmark any model.

1.3 Contributions

A structured comparison of the main DL architectures by data type, strengths and weaknesses.
A domain-wise synthesis of how DL is applied, with the evidence behind each application.
A classification of open challenges into technical, ethical and practical groups.
A set of research gaps and future directions covering efficient, explainable, privacy-preserving and responsible DL.

1.4 Organisation

Section 2 describes the review method. Section 3 gives background on AI, ML and DL. Section 4 reviews architectures and Section 5 applications. Section 6 discusses challenges and Section 7 emerging trends. Section 8 compares approaches and identifies research gaps, and Section 9 concludes.

II. METHODOLOGY

This study follows a qualitative, analytical and systematic literature-review design built only on secondary sources. No surveys or experiments were conducted.

2.1 Sourccs and scarch stratcgy

Literature was searched in IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, Google Scholar and arXiv. Search strings combined the following keywords with Boolean operators: "deep learning", "deep neural networks", "convolutional neural networks", "recurrent neural networks", "LSTM", "Transformer", "generative AI", "explainable AI", "deep learning applications" and "deep learning challenges". Foundational works were included regardless of date, and application and challenge studies were drawn mainly from 2012 onward, the period in which modern DL took hold.

2.2 Sclcction critcria

Criterion	Include	Exclude
Topic	Addresses AI or deep learning and fits at least one review objective	Unrelated to the research objectives
Source	Peer-reviewed journals, conference proceedings,	Sources without scholarly or technical standing

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

	books, recognised preprints and technical documentation	
Content	Gives enough methodological or analytical detail to judge the finding	Lacks sufficient technical information
Novelty	Adds a distinct finding	Repeats findings already represented in the selected set

2.3 Classifiication and analysis

Selected studies were classified along eight dimensions: architecture, application domain, advantages and performance, computational requirements, limitations, explainability, privacy and security, and bias and fairness. A comparative analysis across these dimensions identifies the strengths, weaknesses and suitable uses of each approach, and the gaps reported in Section 8.

2.4 Limitations of the method

Because the review is qualitative and rests on published work, it cannot verify reported results independently. Performance figures quoted from the literature depend on each study's datasets and settings and are not directly comparable across studies. The field also moves quickly, so recent preprints may not yet be peer reviewed.

III. BACKGROUND: AI, MACHINC LEARNING AND DEEP LEARNING

Deep learning is best understood as the innermost of three nested fields. AI is the broadest, covering any technique that lets machines show intelligent behaviour. Machine learning is the subset of AI that improves performance on a task from data [7]. Deep learning is the subset of ML that uses neural networks with many layers to learn representations directly from raw inputs such as pixels, audio waveforms or text [1].

3.1 From hand-crafted ficaturcs to learned representations

Classical ML pipelines depend on engineers designing features by hand, such as edge descriptors for images, before a shallow classifier is trained. Deep networks replace this step. Early layers learn simple patterns, and deeper layers combine them into higher-level concepts, so a network for images can progress from edges to textures to object parts to whole objects [1]. This hierarchical learning is the main reason DL outperforms feature-engineered methods when large datasets are available [2], [3].

3.2 How a network is trained

A neural network is a stack of layers, each applying a weighted sum followed by a non-linear activation. Training adjusts the weights to minimise a loss function that measures prediction error. Gradients of the loss with respect to every weight are computed by backpropagation [37] and the weights are updated by gradient descent or a variant of it:

Here θ denotes the weights, η the learning rate and L the loss. Learning can be supervised (labelled examples), unsupervised or self-supervised (structure found in unlabelled data), or reinforcement-based (rewards from interaction). Self-supervised pre-training followed by task-specific fine-tuning now underlies most large language and vision models [16], [36].

3.3 Why deep learning took off

Neural networks were studied for decades before they became dominant. Deep belief networks showed in 2006 that deep models could be trained effectively [38], but three conditions had to coincide for wide adoption. Large datasets such as ImageNet supplied enough examples, GPUs supplied the parallel compute, and better training techniques and architectures made very deep models stable [4], [5]. Empirical work on scaling later showed that language-model loss falls predictably as model size, data and compute grow [41], which encouraged ever larger models.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

3.4 Millstones

Year	Milestone	Ref.
2021	Vision Transformer applies Transformers to image patches	[9]
2020	GPT-3 shows few-shot learning in a large language model	[10]
2017	Transformer replaces recurrence with self-attention	[6]
2016	Residual networks make very deep CNNs trainable	[5]
2014	Generative adversarial networks introduced	[13]
2012	AlexNet wins ImageNet with a deep CNN trained on GPUs	[4]
2006	Deep belief networks demonstrate layer-wise training of deep models	[38]
1998	LeNet-style CNNs applied to document recognition	[12]
1997	Long short-term memory introduced	[11]
1986	Backpropagation popularised for training multilayer networks	[37]

IV. MAJOR DEEP-LEARNING ARCHITCCTURES

Deep-learning architectures differ mainly in how they exploit the structure of their input: grids for images, sequences for text and speech, or all pairs of elements through attention. Table 1 summarises the families reviewed in this section.

4.1 Artificial neural networks and multilayer perceptions

The multilayer perceptron (MLP) is the basic feed-forward network, in which every neuron in one layer connects to every neuron in the next. MLPs handle tabular data and serve as the classifier or projection layers inside larger models. They ignore spatial and temporal structure, so they scale poorly to raw images or long sequences [2].

4.2 Convolutional neural networks

Convolutional neural networks (CNNs) apply small learned filters across a grid, which shares weights and captures local patterns with far fewer parameters than an MLP [12]. AlexNet showed that a deep CNN trained on GPUs could win the ImageNet benchmark by a large margin [4]. Residual connections later allowed networks with many more layers to be trained without degradation [5]. Variants serve object detection in a single pass [40] and biomedical image segmentation [39]. CNNs remain a strong default for image data, especially when compute is limited.

4.3 Recurrent networks and LSTMs

Recurrent neural networks (RNNs) process a sequence one element at a time while carrying a hidden state, which suits speech, text and time series. Plain RNNs struggle to learn long-range dependencies because gradients vanish or explode over many steps [30]. Long short-term memory (LSTM) networks add gated memory cells that preserve information over long spans [11], and the gated recurrent unit offers a lighter alternative [31]. Attention added to recurrent encoder-decoder models improved machine translation by letting the decoder look back at relevant source words [42]. Recurrence is inherently sequential, which limits parallel training on modern hardware.

4.4 Transformers

The Transformer removes recurrence and relates every position in a sequence to every other through self-attention [6]. Scaled dot-product attention is computed as:

```
\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\!\left(\frac{QK^{\top}}
{\sqrt{d_k}}\right)V
```

where Q, K and V are query, key and value matrices and d_k is the key dimension. Because all positions are processed in parallel, Transformers train efficiently on large corpora. Pre-trained encoders such as BERT transfer to many language tasks after fine-tuning [16], and very large decoder models such as GPT-3 perform new tasks from a few examples given in the prompt [10]. The Vision Transformer applies the same design to image patches and matches or

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

exceeds CNNs when pre-trained on large datasets [9]. The main cost is that attention grows quadratically with sequence length.

4.5 Generative models

Generative models learn a data distribution and sample new examples from it. Variational autoencoders learn a compressed latent space and decode samples from it [15].

Generative adversarial networks train a generator against a discriminator and produce sharp images, but training can be unstable [13]. Diffusion models generate data by learning to reverse a gradual noising process and have become a leading approach for image synthesis [14].

4.6 State space models as an alternative to attention

To address the quadratic cost of attention, selective state space models such as Mamba process sequences with computation that grows linearly in sequence length while remaining competitive with Transformers on several tasks [44]. A recent survey notes that this makes them attractive for long sequences, and that hybrids combining attention and state space layers are an active research direction [43]. Whether they can replace Transformers more broadly is still debated.

4.7 Comparison

Table 1. Comparison of major deep-learning architectures

Architecture	Typical data	Main strengths	Main weaknesses	Typical uses
MLP	Tabular, feature vectors	Simple, flexible	Ignores spatial or temporal structure	Tabular prediction, classifier heads
CNN	Images, grids, audio spectrograms	Weight sharing, local pattern detection, efficient	Limited global context, needs many labelled images	Classification, detection, segmentation
RNN / LSTM	Sequences, time series	Models order, handles variable length	Sequential training, long-range memory still limited	Speech, forecasting, earlier NLP
Transformer	Text, images, audio, multimodal	Parallel training, long-range context, transfers well	Quadratic attention cost, large data and compute	Language models, vision, multimodal systems
VAE / GAN / diffusion	Images, audio, text	Generate new samples	Unstable training (GAN), slow sampling (diffusion)	Image and audio synthesis, data augmentation
State space (Mamba)	Long sequences	Linear-time sequence processing	Younger ecosystem, fewer large-scale results	Long-sequence modelling

V. APPLICATIONS OF DEEP LEARNING

Domain	Representative task	Dominant architectures	Ref.
Computer vision	Image classification, detection, segmentation	CNN, Vision Transformer	[4], [5], [9], [39], [40]
Speech	Acoustic modelling, speech recognition	Deep networks, RNN/LSTM, Transformer	[32]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Natural language processing	Translation, question answering, text generation	Transformer, attention-based RNN	[6], [10], [16], [42], [53]
Healthcare	Skin-lesion and chest X-ray analysis, protein structure	CNN, attention-based models	[17], [18], [52]
Finance and cybersecurity	Fraud, anomaly and intrusion detection	MLP, LSTM, autoencoders	[8]
Robotics and autonomous systems	Driving, game playing, control	CNN with reinforcement learning	[19], [50]
Recommendation	Ranking and candidate generation	Deep feed-forward networks	[51]

5.1 Computer vision

Image recognition was the first domain transformed by DL. AlexNet cut ImageNet classification error sharply in 2012 [4], and deeper residual networks followed [5]. Detection and segmentation models, including single-pass detectors [40] and encoder-decoder networks for biomedical images [39], extended the approach beyond classification. Vision Transformers now compete with CNNs on large-scale benchmarks [9].

5.2 Speech and natural language processing

Deep networks replaced Gaussian mixture models for acoustic modelling in speech recognition and gave clear accuracy gains [32]. In NLP, attention-based translation [42] gave way to Transformer models [6]. Pre-trained language models transfer to many tasks [16], and large models can follow instructions or learn tasks from a few examples [10], [53]. These advances underpin chatbots, search, summarisation and code assistants.

5.3 Healthcare and science

A CNN trained on a large set of clinical images classified skin lesions with accuracy comparable to dermatologists in a published study [17]. A deep network for chest radiographs reported pneumonia-detection performance comparable to or above radiologists on a benchmark set [52]. AlphaFold predicted protein structures with accuracy often close to experimental methods, an advance with wide implications for biology [18]. These results come from specific datasets, and clinical deployment needs prospective validation, regulatory review and attention to bias (Section 6).

5.4 Finance and cyber security

Financial institutions use DL for fraud detection, credit scoring and risk modelling, where sequence models and autoencoders can flag unusual transaction patterns. In cybersecurity, DL supports malware and intrusion detection. Security applications face an adversary who adapts, and deep models can be fooled by small deliberate input changes known as adversarial examples [8]. Published evidence in finance is limited because transaction data and model details are usually proprietary.

5.5 Robotics and autonomous systems

Deep reinforcement learning combined with search produced a system that defeated a professional Go player, showing that DL can learn complex decision policies [19]. An early end-to-end driving system mapped camera images directly to steering commands [50].

Safety-critical deployment still depends on reliability, interpretability and testing far beyond benchmark accuracy.

5.6 Recommendation systems

Large-scale recommenders use deep networks to generate candidate items and rank them for each user, as in a widely cited video-recommendation system [51]. Such systems depend on very large behavioural datasets, which raises the privacy and fairness issues discussed next.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

VI. CHALLENGES AND LIMITATIONS

Despite its record of success, deep learning faces six linked challenges. Table 3 summarises them with the main mitigation strategies found in the literature.

Table 3. Challenges of deep learning and common mitigations

Challenge	Why it matters	Common mitigations	Ref.
Data requirements	Supervised models need large labelled datasets, which are costly and sometimes unavailable	Transfer learning, self-supervised pre-training, data augmentation	[2], [16]
Computational and energy cost	Training and serving large models needs specialised hardware and large amounts of electricity	Compression, distillation, efficient architectures	[20], [27], [34], [45]
Interpretability	Decisions are hard to explain, which limits trust in high-stakes settings	Post-hoc explanation methods, inherently interpretable models	[21]–[25]
Bias and fairness	Models can reproduce and amplify bias in data	Auditing, balanced datasets, fairness-aware training	[29]
Privacy	Models can leak information about training data	Differential privacy, federated learning	[28], [33], [35]
Security	Small crafted perturbations can change predictions	Adversarial training, robust evaluation	[8]

6.1 Data requirements

Deep models improve with more data, but labelling is expensive and many domains, such as rare diseases, have few examples. Transfer learning and self-supervised pre-training reduce the need for labels by reusing representations learned from large unlabelled corpora [16]. Data quality matters as much as quantity, since noisy or unrepresentative data limits what a model can learn [2].

6.2 Computational and environmental cost

Training large models is compute-intensive. A widely cited analysis estimated that training one large NLP model with architecture search produced carbon emissions comparable to several cars over their lifetimes, and argued for reporting energy cost alongside accuracy [20]. The wider picture is now larger. The International Energy Agency projects that electricity demand from data centres will more than double by 2030 to around 945 TWh, with AI the most significant driver [45]. Model-level efficiency techniques include distillation, which trains a small model to imitate a large one [27], and compression by pruning and quantisation, which reduced the storage of AlexNet by about 35 times without loss of accuracy in one study [34].

6.3 Interpretability and explainability

Deep networks are often called black boxes because their internal computations are hard to inspect. Post-hoc methods approximate a model locally with a simpler one [22], attribute a prediction to input features using Shapley values [23], or highlight the image regions that drove a decision [24]. A broad review organises these methods and links them to responsible AI [25]. Some researchers argue that for high-stakes decisions inherently interpretable models should be preferred over explanations of black boxes, because explanations can be unfaithful to what the model actually computes [21].

6.4 Bias and fairness

A widely cited audit of commercial gender-classification systems found error rates of up to 34.7% for darker-skinned women, against 0.8% for lighter-skinned men [29]. The cause was unrepresentative training and benchmark data. Bias can enter through data collection, labelling and objective design, so auditing across subgroups is needed before deployment.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

6.5 Privacy and security

Large models can memorise training data. Researchers have extracted verbatim training examples from a deployed language model [33]. Differentially private training limits what a model can reveal about any one record at some cost in accuracy [28], and federated learning keeps raw data on user devices while sharing only model updates [35]. On the security side, adversarial examples show that imperceptible changes to an input can alter a model's output [8].

6.6 Regulation and responsible deployment

Regulation is starting to address these risks. The EU Artificial Intelligence Act sets obligations for providers of general-purpose AI models that began to apply on 2 August 2025, including documentation and transparency duties [48]. The 2026 AI Index reports that responsible-AI practice is not keeping pace with capability growth [46].

Organisations deploying DL therefore need governance processes alongside technical safeguards.

VII. EMCRGING TRCNDS

Five trends shape current research: generative and foundation models, multimodal learning, explainable AI, efficient and edge AI, and new sequence architectures.

7.1 Generative AI and foundation models

Foundation models are large models trained on broad data and adapted to many downstream tasks. Their capabilities and risks, including homogenisation and concentration of power, have been analysed in a major report [36]. Their rise rests on the Transformer [6], large-scale self-supervised pre-training [16] and the observation that performance improves predictably with scale [41]. Instruction tuning with human feedback made such models follow user intent more reliably [53], and diffusion models extended generation to images [14]. The 2026 AI Index also reports large gains on benchmarks of agents that operate computers, showing that these models are moving from text generation toward task execution [46].

7.2 Multimodal learning

Multimodal models learn from several data types at once. Contrastive training on image-text pairs produced a model that transfers to many vision tasks without task-specific labels [26]. The same shared-representation idea underlies systems that combine text, images, audio and video, and it is a main route toward more general AI systems.

7.3 Explainable and trustworthy AI

Explainable AI (XAI) is moving from visualisation tools toward explanations that are faithful, evaluated with users and tied to regulation [21], [25]. Open problems include measuring whether an explanation reflects the model's real reasoning and extending explanation methods to very large language and multimodal models.

7.4 Efficient and edge AI

Edge AI runs models on phones, sensors and microcontrollers instead of in the cloud, which lowers latency, reduces dependence on connectivity and keeps data on the device. Recent surveys describe a move from tiny machine learning toward tiny deep learning on severely resource-constrained hardware, supported by compression, hardware co-design and on-device learning [47], [49]. Federated learning extends this by training across devices without centralising data [35]. Efficiency also addresses the energy concerns raised in Section 6.2 [45].

7.5 New sequence architectures

State space models such as Mamba offer linear-time sequence processing and are being combined with attention in hybrid designs [43], [44]. The longer-term question is which architecture gives the best trade-off between quality, cost and long-context capability.

VIII. DISCUSSION AND RESEARCH GAPS

The literature shows that no single architecture is best for every task. CNNs remain efficient for image data, recurrent models are being displaced for language by Transformers, and Transformers dominate large-scale language and multimodal work at a high cost in data and compute (Table 1). The practical choice depends on data type, available data volume, latency and energy limits, and the cost of an error in the target domain.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

A second finding is that the strongest results come with the largest caveats. Benchmark successes in healthcare and vision [17], [52] have not removed the need for external validation, fairness audits [29] and explanation [21]. Capability growth is also outpacing governance and efficiency work [45], [46]. Accuracy alone is therefore an insufficient measure of progress, and future evaluations should report cost, robustness, fairness and explainability together.

8.1 Research gaps and directions

Gap	Why it matters	Promising direction	Ref.
Faithful explanation of very large models	Post-hoc explanations may not reflect real model behaviour	Interpretable-by-design models, user-evaluated explanations	[21], [25]
Data-efficient learning	Labels are scarce in many domains	Self-supervised and transfer learning	[16]
Energy-efficient training and inference	Data-centre demand is growing quickly	Compression, distillation, efficient architectures, edge deployment	[27], [34], [45], [47]
Privacy-preserving learning	Models can memorise and leak data	Differential privacy, federated learning	[28], [33], [35]
Robustness to adversarial and shifted inputs	Small changes can alter predictions	Robust training and evaluation	[8]
Fairness across groups	Unequal error rates cause real harm	Subgroup auditing, representative data	[29]
Long-context sequence modelling	Attention cost grows quadratically	State space and hybrid models	[43], [44]

8.2 Implications

For students and practitioners, the main implication is to match the model to the problem rather than to the trend: start with a simple or pre-trained model, measure cost and fairness as well as accuracy, and document limitations. For researchers, the gaps above point to work on efficient, explainable and privacy-preserving methods, which is where practical deployment is most constrained.

IX. CONCLUSION

Deep learning has become the main engine of modern artificial intelligence because multilayer networks learn useful representations directly from data. CNNs, recurrent networks and LSTMs, Transformers and generative models each suit different data types and tasks, and the field is now exploring state space and hybrid designs. Applications in vision, language, healthcare, science, finance, robotics and recommendation show the breadth of the approach.

The same literature shows that deep learning depends on large data, heavy computation and careful governance. Interpretability, bias, privacy, security and energy cost remain open problems, and capability is growing faster than the practices that make systems safe and trustworthy. Future work should therefore pursue data-efficient, energy-efficient, explainable, privacy-preserving and fair deep-learning systems, and evaluate them on all of these dimensions rather than accuracy alone.

This review is limited to published secondary sources and does not benchmark models directly. Future work could add controlled experiments that compare architectures on common tasks under equal compute and data budgets.

REFERENECS

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
2. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
3. J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, 2015.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2012.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
6. A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017.
7. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260, 2015.
8. I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
9. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
10. T. B. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020.
11. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
12. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
13. I. Goodfellow et al., "Generative adversarial nets," in Proc. NeurIPS, 2014.
14. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. NeurIPS, 2020.
15. D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. ICLR, 2014.
16. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
17. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
18. J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, pp. 583–589, 2021.
19. D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, pp. 484–489, 2016.
20. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2019, pp. 3645–3650.
21. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Mach. Intell., vol. 1, pp. 206–215, 2019.
22. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.
23. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017.
24. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
25. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020.
26. A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021.
27. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.
28. M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM Conf. Comput. Commun. Secur. (CCS), 2016, pp. 308–318.
29. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proc. Conf. Fairness, Accountability and Transparency (PMLR vol. 81), 2018, pp. 77–91.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

30. Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Netw., vol. 5, no. 2, pp. 157–166, 1994.
31. K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014.
32. G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.
33. N. Carlini et al., "Extracting training data from large language models," in Proc. USENIX Security Symp., 2021.
34. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. ICLR, 2016.
35. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. Int. Conf. Artif. Intell. Statist. (AISTATS), 2017.
36. R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv:2108.07258, 2021.
37. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp. 533–536, 1986.
38. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.
39. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241.
40. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE CVPR, 2016, pp. 779–788.
41. J. Kaplan et al., "Scaling laws for neural language models," arXiv:2001.08361, 2020.
42. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.
43. H. Qu et al., "A survey of Mamba," arXiv:2408.01129, 2024 (rev. Jun. 2025).
44. A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv:2312.00752, 2023.
45. International Energy Agency, Energy and AI. Paris, France: IEA, 2025. [Online]. Available: https://www.iea.org/reports/energy-and-ai
46. AI Index Steering Committee, The AI Index 2026 Annual Report. Stanford, CA, USA: Stanford Institute for Human-Centered AI, Apr. 2026. [Online]. Available: https://hai.stanford.edu/ai-index/2026-ai-index-report
47. S. Somvanshi et al., "From tiny machine learning to tiny deep learning: A survey," arXiv:2506.18927, 2025.
48. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024.
49. S. Heydari and Q. H. Mahmoud, "Tiny machine learning and on-device inference: A survey of applications, challenges, and future directions," Sensors, vol. 25, no. 10, Art. no. 3191, 2025, doi: 10.3390/s25103191.
50. M. Bojarski et al., "End to end learning for self-driving cars," arXiv:1604.07316, 2016.
51. P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in Proc. ACM Conf. Recommender Syst. (RecSys), 2016, pp. 191–198.
52. P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv:1711.05225, 2017.
53. L. Ouyang et al., "Training language models to follow instructions with human feedback," in Proc. NeurIPS, 2022.