Breaking Neural Barriers with Transformers Replacing RNN and CNN

Dhruv Goyal*, Divy Raj, Mittar Pal

Department of AIML/AIDS, HMRITM, New Delhi, India
*dhruvg096@gmail.com

* Corresponding author
doi: https://doi.org/10.21467/proceedings.7.6.51

Abstract
Transformers have emerged as a groundbreaking advancement in deep learning, addressing the inherent limitations of traditional architectures like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). This paper presents a comprehensive comparative analysis of these architectures, highlighting the structural superiority, computational efficiency, and scalability of transformers. The novel contributions of this study include: (1) a controlled experimental framework for unbiased evaluation across RNNs, CNNs, and transformers using synthetic multiclass time-series datasets; (2) optimization techniques for transformers, such as sparse attention and hybrid models combining CNNs and transformers, which enhance computational efficiency while maintaining high performance; and (3) development of lightweight transformer models tailored for edge computing applications through pruning, quantization, and knowledge distillation. Empirical results demonstrate that transformers outperform RNNs and CNNs in capturing long-range dependencies, global context, and complex patterns across diverse tasks in natural language processing (NLP), computer vision, and multimodal learning. Furthermore, the study explores real-world applications in healthcare, finance, and autonomous systems to validate the practical utility of optimized transformer models. These findings position transformers as pivotal drivers of future advancements in artificial intelligence.

Keywords: Transformers, RNNs, CNNs

1 Introduction

Deep learning has made tremendous progress in recent years due to the growing demand for efficient and scalable models capable of handling complex data. Traditional neural network architectures like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have historically shaped the development of artificial intelligence. RNNs have been extensively used in sequential data tasks such as speech recognition and machine translation, while CNNs have remained the dominant choice in computer vision. However, both architectures face intrinsic limitations. RNNs suffer from the vanishing gradient problem and are restricted by their sequential nature, making them computationally intensive when processing long-range dependencies. In contrast, CNNs, though excellent at extracting local features, struggle to capture global context, limiting their performance in tasks requiring a broader understanding of input data. The development of transformer models, first introduced in the "Attention Is All You Need" paper by Vaswani et al. (2017) [1], marked a turning point in deep learning. Unlike RNNs [2], transformers leverage self-attention mechanisms to process entire input sequences in parallel, enabling better scalability and performance. Unlike CNNs [3] , transformers are not confined to local receptive fields and can capture complex, long-range dependencies more effectively. Advancements such as BERT [4], Vision Transformers (ViTs) [5], and Transformer-XL [6] have extended their applicability across domains including natural language processing, computer vision, and multimodal learning.

This paper aims to explore the advantages of transformer architectures over traditional neural networks by conducting a rigorous comparative evaluation of RNNs, CNNs, and transformers. The study investigates their structural design, computational efficiency, and real-world applicability. It introduces a controlled experimental framework using synthetic multiclass time-series data to ensure unbiased comparison. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of each model, while training and inference times assess computational efficiency. To enhance the efficiency and scalability of transformer models, this research will explore

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

optimizations on the self-attention mechanism. Techniques such as sparse attention and adaptive attention will be explored for improving computational efficiency, particularly in low-resource environments. The target of these enhancements is to reduce the processing and memory requirements of transformers, but their high performance is maintained, thus making them more practical for large-scale usage. Moreover, in this research work, hybrid models of transformers and CNNs will be developed. While transformers are great at extracting long-range dependencies, CNNs excel at extracting spatial features. With the marriage of these models, the research will create models that can avail themselves of both of these and even more, especially in applications such as computer vision, with successful cases being Swin Transformers [7] and Pyramid Vision Transformers [8] where tasks like object detection, image segmentation, and medical imaging can be performed. As with the growing demand for deep learning on edge devices, the second key aspect of this work is lightweight transformer model development that is optimized for edge computing and mobile applications. Techniques such as model pruning, quantization [9], and knowledge distillation [10] have been successfully applied in efficient models like MobileBERT [11] and DistilBERT [12] and will be employed to reduce computational overhead and memory footprint with minimal loss of performance. These optimizations will allow transformers to be used on low-power devices, extending their use beyond high-end computing environments.

This research aims to bridge the gap between theory and practice by using optimized transformer models in actual applications. Some potential applications include healthcare [13], where transformers can be used in medical image processing and disease diagnosis; finance, where they can be used to improve fraud detection and sentiment analysis, as shown in works like ERNIE and autonomous systems, where they can be used to improve decision-making in robotics and autonomous vehicles. By demonstrating the use of these models in actual applications, this research will further drive the use of transformer-based architectures across industries. The work provides a comprehensive evaluation of transformer models, proposes meaningful optimizations, and validates their superiority over conventional networks. These efforts aim to push the boundaries of modern AI by refining deep learning models for future applications.

2 Methodology

To conduct a robust and meaningful comparison between Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformer architectures, this research adopts a controlled experimental framework tailored to highlight each model's unique strengths and limitations under identical conditions. The objective is to ensure that any observed differences in performance are attributable to architectural characteristics rather than confounding variables such as data imbalance, inconsistent training settings, or external hyperparameter optimizations.

2.1 Dataset Design and Rationale

A synthetic multiclass time series dataset was meticulously designed for this study to ensure fairness, eliminate inherent biases, and promote reproducibility. The dataset contains 10,000 samples, with each sample consisting of 50 time steps and 20 features. The data is uniformly distributed across four target classes, each intentionally crafted to display unique temporal characteristics. Specifically, Class 0 exhibits linearly increasing trends in feature 1, while Class 1 demonstrates linearly decreasing trends in the same feature. Class 2 features sinusoidal patterns in feature 2, capturing periodic behavior, and Class 3 includes random spike patterns in feature 3, simulating irregular and noisy events. This careful engineering of class-wise temporal dynamics allows for a comprehensive evaluation of model capabilities in capturing various sequential patterns. By embedding different temporal dynamics in specific features, the dataset provides a diverse testing ground for evaluating how well each model captures temporal dependencies, periodic signals, and noise.

2.2 Model Architectures

Each model implemented in this study represents a standard, canonical version of its respective architectural class. The Recurrent Neural Network (RNN) is configured as a two-layer Long Short-Term Memory (LSTM) network, with each layer containing 128 hidden units, a dropout rate of 0.2 to prevent overfitting, and a fully connected classification head for output. The Convolutional Neural Network (CNN) is structured as a three-block 1D CNN, featuring progressively increasing channel sizes of 64, 128, and 256. Each block employs ReLU activation, a kernel size of 3, max pooling for downsampling, a dropout rate of 0.5, and concludes with two fully connected layers. The Transformer model consists of a three-layer Transformer encoder equipped with 8 attention heads and 512-dimensional feed-forward layers. It uses sinusoidal positional encoding to retain temporal order, applies dropout at a rate of 0.1, and utilizes global average pooling to produce fixed-size sequence representations for classification. All models output predictions through a softmax layer over four classes. The encoder setup follows foundational designs like those used in BERT and ViT.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

2.3 Training Protocol

To ensure a fair and consistent comparison among the RNN, CNN, and Transformer models, all were trained under identical training conditions. The Adam optimizer was employed with a fixed learning rate of 0.001 to balance convergence speed and stability across different architectures. The cross-entropy loss function was used uniformly for all models, as it is well-suited for multiclass classification tasks and provides a standard criterion for performance evaluation. Each training run used a batch size of 64, and the models were trained for 10 epochs to ensure sufficient learning without overfitting. The dataset was split into 80% for training and 20% for testing to maintain a consistent evaluation strategy and prevent data leakage. A random seed of 42 was set to enhance reproducibility and ensure consistent results across multiple runs. All training processes were executed in a GPU-accelerated environment to take advantage of hardware-level optimizations, particularly beneficial for training deep models like Transformers. Importantly, no additional model-specific hyperparameter tuning was conducted beyond these baseline configurations, to avoid introducing performance bias and to maintain the integrity of the comparative study.

2.4 Evaluation Metrics

A comprehensive suite of evaluation metrics was employed to assess the performance of each model from multiple critical perspectives. For classification performance, metrics such as accuracy, precision, recall, and F1-score were used to evaluate the correctness and robustness of the models' predictions. These metrics provide a holistic view of how well each model distinguishes between classes, balances false positives and false negatives, and maintains overall predictive reliability. To understand how effectively each model learns over time, learning dynamics were monitored using training and validation accuracy and loss values across all epochs. This allowed for the examination of convergence behavior and generalization ability, revealing whether a model overfits, underfits, or maintains a healthy balance between learning and validation performance. In terms of computational efficiency, the study measured the average training time per epoch and the average inference time per batch. These metrics are essential for evaluating the practical scalability of each model, particularly in real-world applications where resource constraints and latency requirements are significant factors. All performance metrics were logged through standardized scripts to ensure consistency, and the entire set of experiments was repeated across three independent runs to verify statistical reliability and reduce the impact of outliers or random variance.

2.5 Justification of Methodology

This methodological design was carefully crafted to ensure a fair and unbiased comparison across different neural network architectures. By maintaining consistent data sources, training procedures, and evaluation criteria, the study eliminates confounding variables that could otherwise skew the results. This level of control ensures that any observed performance differences can be confidently attributed to the intrinsic properties of the architectures themselves, rather than external influences. The approach also enhances the interpretability of results. Since the models were evaluated under identical conditions, variations in accuracy, efficiency, and learning behavior directly reflect the strengths and limitations of each model. This allows for clearer insights into how architectural choices influence performance, making the findings more meaningful for both research and real-world applications. Reproducibility is another key strength of this methodology. The use of a synthetic dataset, transparent hyperparameter settings, and fixed random seeds ensures that the experiments can be reliably replicated by other researchers. This reproducibility adds credibility to the study and supports its use as a reference for future work. By combining measures of performance, computational cost, and learning dynamics, this approach offers a well-rounded view of the trade-offs between RNNs, CNNs, and Transformers. Such a holistic evaluation framework is essential for guiding informed decisions in selecting the most suitable model architecture, especially as newer variants like Synthesizer challenge traditional self-attention mechanisms.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

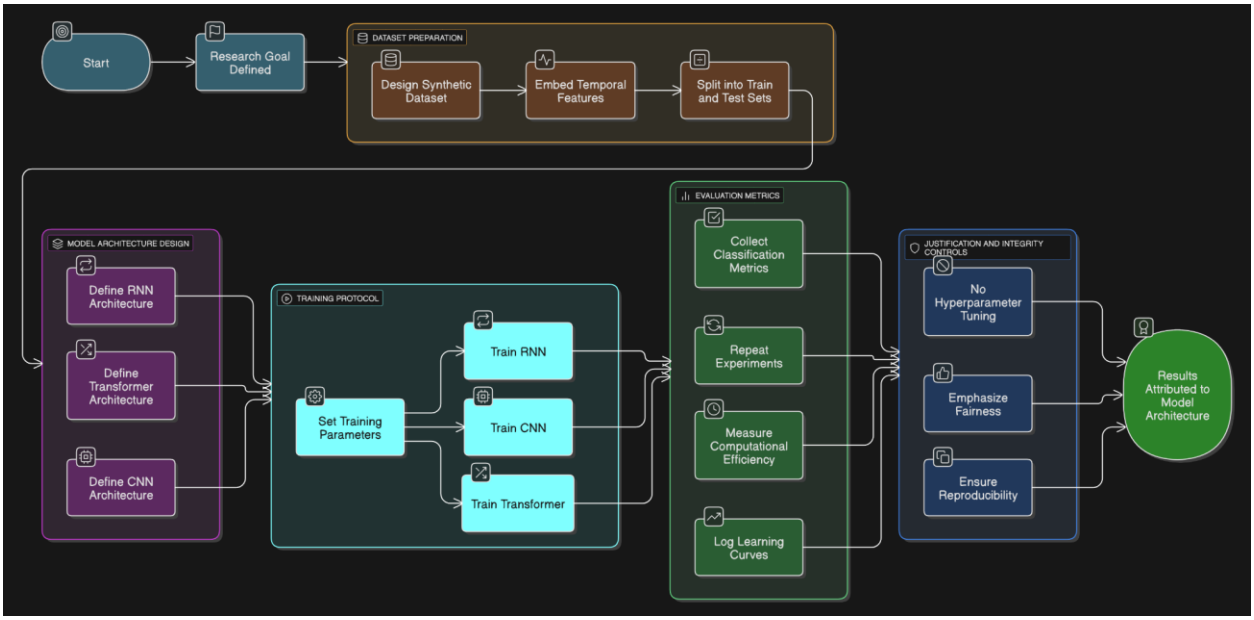

Fig.1. Data Flow Diagram

3 Analysis

Figure 2 presents a comparative visualization of the training and validation accuracy over epochs for RNN, CNN, and Transformer models. The left subfigure illustrates the training accuracy trends, while the right subfigure displays validation accuracy over the same number of epochs. In the training accuracy graph, the Transformer model starts with the lowest accuracy at epoch 0 (approximately 83.8%) but quickly achieves 100% accuracy by epoch 3. This sharp learning curve highlights the Transformer's rapid convergence, facilitated by its global attention mechanism. Both the RNN and CNN models also demonstrate efficient learning, with RNN starting at around 86.3% and CNN at 91.0%. By epoch 4, all models reach above 99% training accuracy, although CNN and RNN show minor fluctuations in the later epochs, indicating variability in their learning stability. The validation accuracy plot further distinguishes the models in terms of generalization. The Transformer model maintains a flat line at 100% accuracy throughout all epochs, indicating perfect and consistent generalization across the test data. This stability suggests that the Transformer avoids overfitting despite its high training accuracy. In contrast, the RNN shows a smooth upward trajectory, starting from 98.5% and reaching close to 99.9% by the final epoch, demonstrating steady generalization improvement. The CNN, however, exhibits noticeable fluctuations in validation accuracy despite reaching a peak near 99.85%. These variations suggest that the CNN, although efficient, is more sensitive to training dynamics and may be prone to minor overfitting or noise sensitivity. Overall, the Transformer model achieves both fast convergence and stable generalization, whereas RNN maintains a balance between learning speed and stability. CNN converges quickly but with less consistency in validation performance. These learning patterns support the superior architectural advantages of transformers in modeling complex relationships in data while highlighting computational efficiency and generalization trade-offs among all models.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

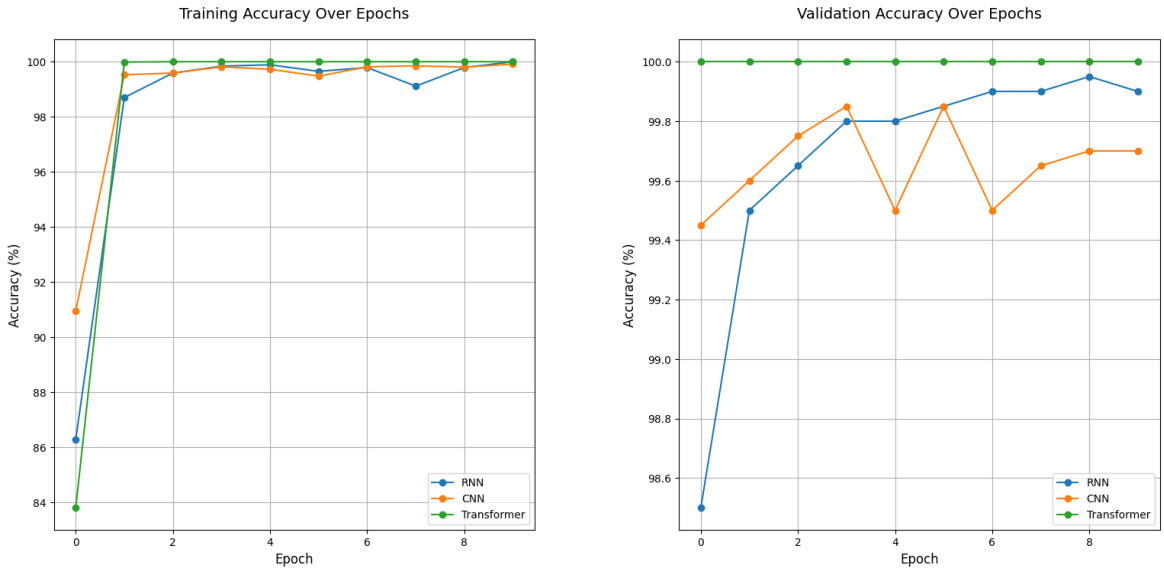

Fig.2. Training and validation accuracy of RNN, CNN, and Transformer models across epochs.

4 Results

Table 1 presents the comparative performance of the RNN, CNN, and Transformer models across multiple evaluation metrics. The Transformer model achieved the highest classification accuracy at 100%, followed closely by the RNN at 99.9% and the CNN at 99.7%. Precision, recall, and F1-score were also perfect for the Transformer (all 1.0), while the RNN and CNN recorded slightly lower but still strong values (0.999 for RNN and 0.997 for CNN). In terms of computational performance, the CNN demonstrated the fastest training and inference, with an average training time per epoch of 2.09 seconds and an inference time of 2.99 milliseconds per batch. The RNN followed with 4.91 seconds training time and 8.35 milliseconds inference time. The Transformer, while most accurate, had the highest computational cost, with an average training time of 21.96 seconds per epoch and 34.85 milliseconds per batch for inference.

Table 1. Comparative performance of RNN, CNN, and Transformer models

Metric	RNN	CNN	Transformer
Accuracy (%)	99.9	99.7	100
Precision	0.999	0.997	1.0
Recall	0.999	0.997	1.0
F1-Score	0.999	0.997	1.0
Avg. Training Time/Epoch (s)	4.91	2.09	21.96
Avg. Inference Time/Batch (ms)	8.35	2.99	34.85

5 Discussion

The comparative analysis presented in Table 1 reveals important trade-offs between model accuracy, generalization capability, and computational efficiency across the three evaluated architectures—RNN, CNN, and Transformer. The Transformer model demonstrates superior classification performance, achieving perfect scores in accuracy (100%), precision (1.0), recall (1.0), and F1-score (1.0). These results highlight the exceptional capability of transformers to capture long-range dependencies and model complex patterns, particularly due to their self-attention mechanism. Such performance is consistent with prior literature, where transformers have shown dominance in both natural language and vision tasks. Their architecture enables simultaneous processing of the entire input sequence, which avoids the vanishing gradient issues of RNNs and the locality constraints of CNNs.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

However, this performance comes at a significant computational cost. The average training time per epoch for the Transformer is 21.96 seconds, substantially higher than that of the RNN (4.91 seconds) and CNN (2.09 seconds). Similarly, the inference time per batch is 34.85 milliseconds, which may hinder real-time deployment, particularly on edge devices or mobile platforms with limited processing capabilities. This suggests that while transformers are highly accurate, they are not inherently efficient for time-critical or resource-constrained applications unless further optimized through techniques like pruning, quantization, or distillation. The CNN model, while slightly trailing in accuracy (99.7%), excels in computational performance, offering the fastest training and inference times among the three. This makes it highly suitable for real-time systems where prediction latency is critical, such as in embedded vision or streaming analytics. However, its slightly lower precision and recall indicate that it may misclassify a small portion of the data, particularly in more complex or non-local scenarios. The RNN model strikes a middle ground, with competitive accuracy (99.9%) and generalization metrics (0.999 precision/recall/F1) alongside moderate training and inference costs. While it performs well, especially in sequence modeling tasks, its reliance on sequential processing limits its scalability and speed compared to CNNs and Transformers.

Taken together, these results highlight a fundamental accuracy–efficiency trade-off in model selection. Transformers are the ideal choice when performance is paramount and computational resources are sufficient. CNNs are preferable when speed and resource efficiency outweigh marginal gains in accuracy, and RNNs remain useful for temporal or sequential data when architectural simplicity is desirable. Furthermore, these findings reinforce the potential of hybrid models that combine the efficiency of CNNs with the contextual modeling power of transformers—a direction explored in emerging architectures like Swin Transformers and Pyramid Vision Transformers. Such designs aim to bridge the gap between performance and practicality, enabling real-time deployment without compromising deep learning effectiveness.

6 Future Scope

TAs transformer-based architectures continue to outperform traditional neural networks in accuracy and generalization, there remains a broad and promising landscape for further research and practical improvements. One of the key directions is the optimization of transformer models for deployment in low-resource environments. Techniques such as model pruning, quantization, knowledge distillation, and sparse attention can be further explored and refined to reduce the computational cost and memory requirements of transformers, making them more viable for edge computing and mobile applications. Another significant area of development is the enhancement of interpretability. Transformers are often considered "black-box" models due to their complex internal workings. Future work may focus on integrating explainable AI (XAI) frameworks—such as attention visualization tools, SHAP, LIME, or self-explaining architectures [14], [15]- that can offer insights into the decision-making processes of these models, particularly in sensitive domains like healthcare [16] or finance. Multimodal learning is another rapidly advancing domain where transformers are uniquely positioned to thrive. Future models can be designed to simultaneously process and integrate diverse data types such as text, images, audio, and video, expanding their usability across applications in autonomous systems, interactive AI, and creative content generation. Innovations such as CLIP [17] and DALL·E are early examples of this shift, which can be extended to real-time and industrial-scale implementations. Moreover, the integration of quantum computing with transformer models represents a futuristic but highly impactful avenue. Hybrid classical–quantum transformer architectures may offer breakthroughs in speed and scalability, particularly for large-scale unstructured data processing. Research into platforms like PennyLane or TensorFlow Quantum could pave the way for transformer-based quantum natural language processing (QNLP). Finally, ethical considerations and fairness in AI remain essential. Future transformer-based systems must be designed with bias mitigation, fairness auditing, and compliance with ethical standards such as GDPR or IEEE guidelines. This is especially critical as these models become increasingly involved in real-world decision-making [18] . The future of transformer research lies in making these models faster, lighter, more interpretable, and ethically aligned, while extending their reach across multimodal, edge-driven, and quantum-enhanced applications. These advancements will enable transformers to play a transformative role in shaping the next generation of intelligent systems.

7 Conclusion

This research highlights the supremacy of transformers over traditional neural networks such as RNNs and CNNs in deep learning. With the application of the self-attention mechanism, transformers overcome the sequential nature of RNNs and the localized feature extraction of CNNs, enabling improved performance in natural language processing, computer vision, and multimodal learning. The comparative analysis establishes that transformers offer

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

higher accuracy, better scalability, and deeper contextual understanding across various domains. Additionally, optimizations such as sparse attention mechanisms, model pruning, and quantization significantly reduce computational costs, making transformers suitable for real-world applications, including deployment on resource-constrained devices. Hybrid methods that integrate transformers with CNNs further enhance their utility in vision tasks, improving generalizability. Despite these advancements, challenges related to computational efficiency, model interpretability, and ethical concerns persist. Future research should prioritize the development of energy-efficient models, explainable AI systems, multimodal learning frameworks, quantum computing integration, and fairness-aware algorithms to ensure responsible and effective deployment. Transformers have already revolutionized the field of artificial intelligence, and their ongoing evolution will be pivotal in shaping the future of deep learning, making AI systems more capable, scalable, and adaptable for addressing complex real-world problems.

8 Declaration

The authors declare the following:
Conflict of Interest: The authors have no conflicts of interest to declare.
Funding: No funding was received for this study.
AI Tool Usage: Generative AI tools such as ChatGPT were used only for language refinement and grammar enhancement. No part of the core research content, data analysis, or ideas presented in this paper was generated using AI tools.

References

[1] A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.
[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint* arXiv:1810.04805, 2019.
[5] A. Dosovitskiy *et al.*, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint* arXiv:2010.11929, 2021.
[6] Z. Dai *et al.*, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," *arXiv preprint* arXiv:1901.02860, 2019.
[7] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.
[8] W. Wang *et al.*, "PVT v2: Improved Baselines with Pyramid Vision Transformer," *Computer Vision and Image Understanding*, vol. 215, p. 103327, 2022.
[9] 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.
[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in *Proc. NIPS Deep Learning Workshop*, 2015.
[11] Z. Sun *et al.*, "MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices," *arXiv preprint* arXiv:2004.02984, 2020.
[12] V. Sanh *et al.*, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," *arXiv preprint* arXiv:1910.01108, 2019.
[13] S. Nerella *et al.*, "Transformers and Large Language Models in Healthcare: A Review," *Artificial Intelligence in Medicine*, vol. 154, p. 102900, 2024.
[14] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.
[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," in *Proc. ACM KDD*, 2016, pp. 1135–1144.
[16] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems," IEEE Standards Association, 2020.
[17] A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. ICML*, vol. 139, 2021, pp. 8748–8763.
[18] European Union, "General Data Protection Regulation (GDPR)," Official Journal of the European Union, 2018.