**Journal of Advanced Health Informatics Research (JAHIR)**
Vol. 2, No. 2, (2024), pp. 85-94
ISSN: 2985-6124, DOI: 10.59247/jahir.v2i2.292

# Understanding Transformers: A Comprehensive Review

Berliana Rahmadhani[1,*], Purwono[2] , Safar Dwi Kurniawan[3]
[1,2] *Department of Informatics, Universitas Harapan Bangsa, Purwokerto, Indonesia*
[3]*Department of Computer Engineering, Politeknik Harapan Bersama, Tegal, Indonesia*

**ARTICLE INFO**

**Article history:**

Received March 18, 2024
Revised August 8, 2024
Published August 28, 2024

**Keywords:**

Transformer;
Self Attention;
Deep Learning;
Visual Transformer;
Positional Encoding;

**ABSTRACT**

Transformers have been recognized as one of the most significant innovations in the development of deep learning technology, with widespread application to Natural Language Processing (NLP), Computer Vision (CV), and multimodal data analysis. The self-attention mechanism, which is at the core of this architecture, is designed to capture global relationships in sequential and spatial data in parallel, enabling more efficient and accurate processing than Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN)-based approaches. Models such as BERT, GPT, and Vision Transformer (ViT) have been used for a variety of tasks, including text classification, translation, object detection, and image segmentation. Although the advantages of this model are significant, the high computing power requirements and reliance on large datasets are major challenges. Efforts to overcome these limitations have been made through the development of lightweight variants, such as the MobileViT and Swin Transformer, which are designed to improve efficiency without sacrificing accuracy. Further research is also directed at the application of transformers for multimodal data and specific domains, such as medical image analysis. With its high flexibility and adaptability, transformers continue to be regarded as a key component in the development of more advanced and far-reaching artificial intelligence.



**Corresponding Author**:

Berliana Rahmadhani, Universitas Harapan Bangsa, Purwokerto, Indonesia
Email: berlianardn05@gmail.com

## 1. INTRODUCTION

Transformers are recognized as one of the most significant innovations in the development of deep learning technology [1]. Since its introduction by Vaswani et al. in 2017, it has replaced traditional Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN)-based approaches in a variety of tasks [2]. Its advantage lies in its ability to process sequential and spatial data in parallel, which provides higher efficiency and accuracy than previous approaches. Its application to Natural Language Processing (NLP) has resulted in flagship models such as BERT and GPT that are changing the way various applications work, including text classification, translation, and summarization [3].

Transformers have been adopted in fields such as Computer Vision (CV) and medical image analysis. Models such as the Vision Transformer (ViT) and Swin Transformer demonstrate excellence in image classification, object detection, and image segmentation tasks [4]. In the medical domain, Transformers make significant contributions to image segmentation, image-based diagnosis, and analysis of complex medical data. This demonstrates the flexibility and adaptability of the Transformer in a wide range of applications [5][6].

The challenges faced by Transformers are the high computing power requirements and the reliance on large datasets during training. This is a major obstacle to deployment on resource-constrained devices, such as mobile applications or environments with minimal compute infrastructure [7]. The research is focused on optimizing Transformers through the approach and development of lightweight models, which aims to reduce computational complexity without sacrificing model performance [8].

This article aims to provide a comprehensive overview of Transformers, covering their core concepts, variants, and applications across various domains [9]. The discussion included efficiency solutions and strategies to overcome existing limitations, as well as exploration of future development potential. A narrative approach and systematic literature exploration are used to provide in-depth references for researchers and practitioners [10][11].

This article discusses how Transformers are adapted to meet specific needs in various domains, including medical image analysis and other data-driven tasks. Future research directions are identified to drive innovation in the development of efficient and adaptive Transformers. This strategy provides insight to the research community on how to overcome the challenges of implementing Transformers at various application scales [12].

## 2. ARCHITECTURE

### 2.1. Base Architecture

Encoder-decoders are one of the fundamental frameworks in deep learning architectures that are often used for a variety of tasks, from image segmentation to language translation. The ecmoder and decoder models can be seen in Figure 1 [13]. In this architecture, the encoder is tasked with extracting information from the input data by converting the input into a more compact and meaningful representation of the feature. This representation is then passed to the decoder, who is responsible for reconstructing the output according to the needs of the task, for example producing segmentation maps or text in other languages. This approach ensures that the model can capture important information while ignoring irrelevant data, improving efficiency and accuracy [14].

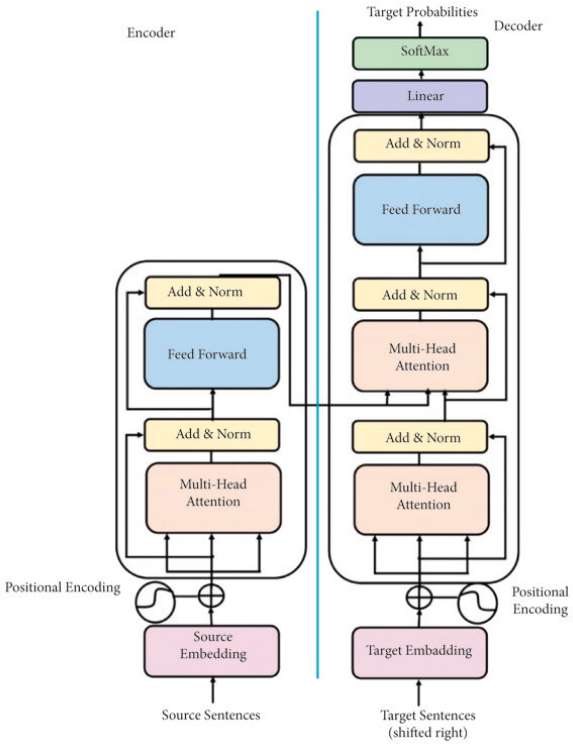


**Fig. 1.** The transformer encoder-decoder model

Another important component in this architecture is multi-head attention, which is designed to capture the global relationships between elements in the data. Unlike regular attention mechanisms, multi-head

attention utilizes multiple attention in parallel, allowing the model to focus on different aspects of the data simultaneously [15]. For example, in an image segmentation task, this component can help the model understand the relationship between a particular area in an image and its overall context. This advantage makes multi-head attention a key element in modern models, such as transformers, which excel at processing sequential or spatial-based data [16].

In addition, feed-forward layers complement the architecture by providing non-linear processing capabilities to the representation of data generated by encoders or attention mechanisms. This layer consists of a series of linear transformations followed by a non-linear activation function, such as ReLU [17]. Feed-forward layers are tasked with reinforcing important features while reducing information redundancy, so that the model can detect more complex patterns. The combination of these layers with other components creates an efficient processing line, guaranteeing optimal performance for a variety of challenging tasks [18][19].

By integrating encoder-decoders, multi-head attention, and feed-forward layers, the architecture is capable of handling a wide range of data types, including text-based, imagery, or sequence-based data. This processing flexibility and power makes it the foundation for many advanced models in deep learning, such as BERT for natural language processing and SegFormer for image segmentation [20]. The integration of these elements allows the model to leverage local and global relationships simultaneously, resulting in precise and accurate results in a wide range of applications.

## 2.2. Self-Attention Mechanism

Self-attention in transformers is a core mechanism that allows models to understand global relationships between elements in sequential data, such as words in a sentence or pixels in an image. This mechanism works by comparing each element of data against the others to determine its relevance. This is done through query *(Q),* key *(K),* and value *(V)* operations, where each is generated from an input vector with a trainable weight matrix. The process is summarized in the attention formula as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

In this formula, Q is the query matrix, K is the key matrix, V is the value matrix, and $d_k$ is the key dimension. Normalization with $\sqrt{d_k}$ helps avoid excessive product dot values, thus maintaining Softmax computing stability.

The self-attention mechanism in transformers uses a multi-head attention approach, where multiple attentions *(h)* are performed in parallel to capture various relationship patterns in the data. Each attention operates a separate version of *Q, K,* and *V*, and then the results are combined to produce the final representation. The formula for multi-head attention is:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_n)W^O \quad (2)$$

Here, $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, with $QW_i^Q, KW_i^K$ and $VW_i^V$ as a weight matrix for *query*, *key*, and *value* on attention at step *i* and $W^O$ is the output weight matrix. Self-attention in transformers is also equipped with a positional encoding mechanism to capture positional information in sequential data, which is not directly noticed by the attention mechanism. The combined self-attention, multi-head attention, and feed-forward layers allow the transformer to handle complex tasks, such as NLP and CV, with high efficiency and flexibility.

## 2.3. Positional Encoding

Positional Encoding on Transformers is a mechanism used to insert information about the sequence or position of elements in sequential data [21]. This is important because the self-attention mechanism does not explicitly consider the order of the data elements, so positional encoding is necessary to provide positional context in the input representation. Positional encoding adds positional information directly to the input embedding through a mathematical function [22][23].

Positional encoding is usually defined using sinusoidal functions, with the following formula for *the post* position and embedding dimension - *i*:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000\frac{2i}{d_{model}}}\right)$$
$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000\frac{2i}{d_{model}}}\right) \quad (3)$$

Where post is the position index in the input sequence, $i$ is the dimension index of the embedding and $d_{model}$ is the total dimension of embedding. This sinusoidal function was chosen because it provides periodic properties that allow the model to easily extrapolate position information to a larger sequence length of the training data.

Once the positional encoding is calculated, these values are added to the input embedding, resulting in a new representation for the data element that incorporates the positional information. Mathematically, the updated input representation is:

$$Z = X + PE \quad (2)$$

Where $X$ is input embedding, $PE$ is positional encoding, and $Z$ is an input embedding that has been enriched with position information. The advantage of this sinusoidal approach is that it does not require any additional parameters that can be trained, so it remains lightweight and allows the model to easily understand positional relationships even beyond the length of the training data sequence. With the integration of positional encoding, transformers become better able to handle sequential data, such as text, which is highly dependent on the context of the order of elements.

## 3. DEVELOPMENT OF TRANSFORMER VARIANTS

### 3.1. Natural Language Processing

Natural Language Processing is a branch of artificial intelligence that studies how computers can understand, generate, and process human language. The development of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and T5 (Text-to-Text Transfer Transformer), has brought about a revolution in this field [24]. BERT is designed with a bidirectional approach, which allows it to understand the context of words by looking from both directions, making it superior for tasks such as text classification and information extraction. In contrast, GPT, which adopts a generative approach, shows excellent performance in generating cohesive and natural text, making it popular in creative tasks such as autowriting and interactive dialogue. T5, on the other hand, summarizes all NLP tasks in a text-to-text format, making it a versatile tool for a wide range of applications, from translation to text summarization [25][26].

In recent years, the popularity of NLP-based generative models has increased with the emergence of GPT-based applications, such as OpenAI's ChatGPT, Google's DeepMind's Gemini, and Microsoft's CoPilot [27]. ChatGPT has become one of the most well-known NLP applications thanks to its ability to answer questions, assist with writing, and support interactive and informative dialogue, making it a very useful tool for individuals as well as businesses [28]. Google DeepMind introduces Gemini as a strong competitor, focusing on integrating multimodal capabilities capable of processing text, images, and other data simultaneously. On the other hand, Microsoft CoPilot leverages the integration of generative models with productivity apps like Microsoft Office, helping users create documents, analyze data, and even present more efficiently.

The application of these models has expanded to a wide range of sectors, from customer service and education to the development of AI-based productivity tools and personal assistants. The success of models such as ChatGPT, Gemini, and CoPilot not only demonstrates the technical superiority of NLP technology but also marks an important transition in the way humans interact with computers. With the continued development of this technology, the future of NLP is predicted to focus more on personalization, multimodality, and cross-domain efficiency [29].

### 3.2. Computer Vision

Transformers in Computer Vision have brought significant innovations in visual data processing approaches, which were previously dominated by CNN networks [30]. Architectures such as Vision Transformer (ViT) and Swin Transformer have proven that transformer-based approaches are capable of competing with CNNs in a variety of tasks, including image classification, object detection, and image segmentation. Vision Transformer as seen in Figure 2 [31], designed to take advantage of the self-attention mechanism, which functions to capture the global relationships between pixels in the image. In its implementation, images are broken down into small patches that are treated as tokens, resemble words in text, and then processed through a transformer architecture to produce a contextual and immersive visual representation.

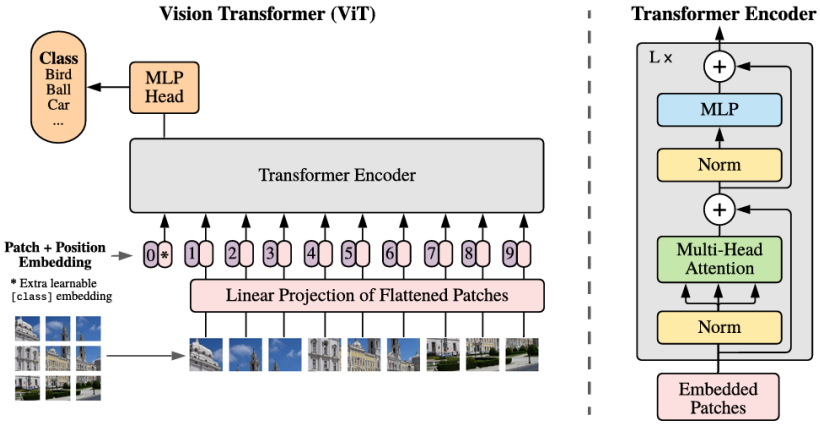


**Fig. 2.** The Vision Transformer (ViT) Architecture

Swin Transformer (Shifted Window Transformer) expands ViT capabilities by implementing a local window-based attention mechanism that shifts gradually [32]. This approach allows self-attention calculations to be performed locally in a small window, before the scope is expanded by shifting the window. This strategy improves computational efficiency on high-resolution images while maintaining an understanding of the global relationships between features. In addition, the hierarchical structure applied to Swin Transformer makes it more flexible and compatible with a variety of tasks, including image segmentation and object detection, and is more efficient than a fully global approach.

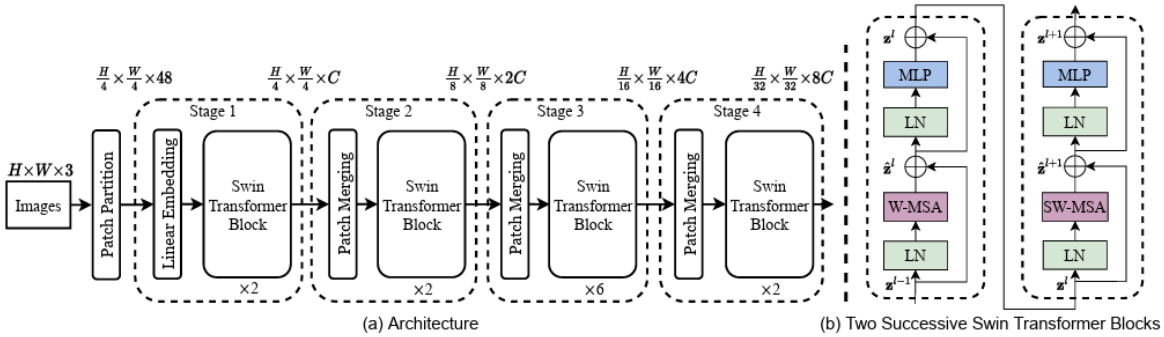


**Fig. 3.** (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks [33]

Both architectures have demonstrated competitive performance, even surpassing CNNs in several well-known benchmarks, such as ImageNet, COCO, and ADE20K. Another advantage offered is the flexibility of the transformer architecture, which allows the integration of multimodal data such as text and image combinations. This transformation opens up new opportunities for wider applications in the field of computer vision, demonstrating significant application potential across various domains of technology and research [33].

### 3.3. Medical Applications

Transformers have been widely applied in a variety of medical applications, including medical image segmentation, image-based diagnosis, and clinical data classification [34]. In the task of segmenting medical images, the ability of Transformers to distinguish complex structures such as organs, tissues, or lesions has been shown to be superior to CNN-based methods. Models such as TransUNet and MedT are designed to utilize the attention mechanism to produce more accurate segmentation.

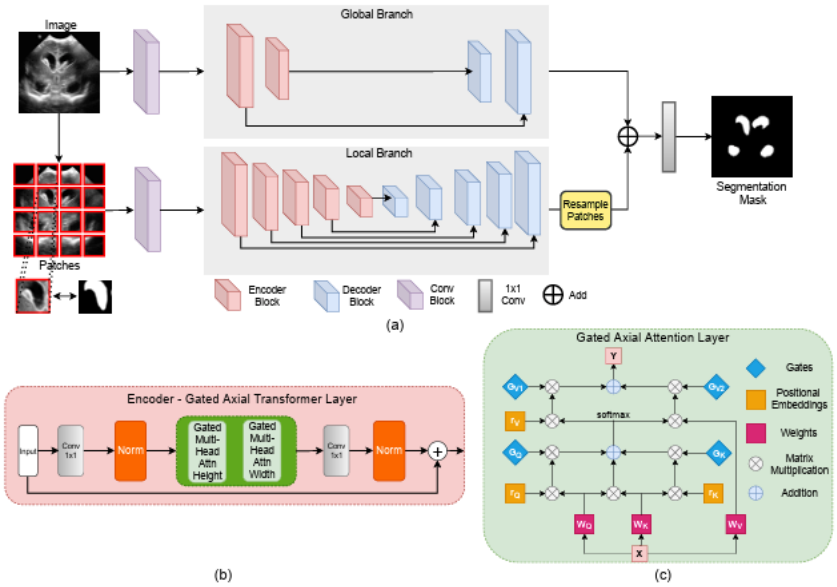


**Fig. 4.** (a) The main architecture diagram of MedT which uses LoGo strategy for training. (b) The gated axial transformer layer which is used in MedT. (c) Gated Axial Attention layer which is the basic building block of both height and width gated multi-head attention blocks found in the gated axial transformer layer.

Based on Figure 4 [35] MedT (Medical Transformer) uses a transformer-based approach specifically designed for medical image segmentation, with an architecture that combines Global Branch and Local Branch to capture information holistically. The medical imaging data is broken down into patches, which are then processed by the Global Branch using the Gated Axial Transformer Layer to understand the spatial and contextual relationships between patches on a global scale. Meanwhile, the Local Branch focuses on processing local spatial details using convolution to capture small patterns that are important in medical imagery.

At the encoder stage, these two branches work in parallel to extract local and global features, which are then passed on to the decoder to reconstruct the image with more precise segmentation. One of the key components in MedT is the Gated Axial Attention Layer, which is designed to improve the efficiency of the self-attention mechanism by reducing computational complexity, allowing for a more directed focus on important elements in medical imaging. This combination results in a rich and in-depth representation of features, which is ideal for tasks such as tumor segmentation or organ structure identification.

In image-based diagnosis, Transformers are used to analyze different types of medical images, such as X-rays, MRIs, and CT scans. Relevant patterns, including tumor or anomaly detection, can be identified with a high degree of accuracy. The processing of spatial relationships in medical images is carried out by Transformer to ensure that high-resolution features can be well understood, thus supporting clinical decision-making.

In the task of classifying clinical data, Transformer has been applied to process multimodal data that includes a combination of medical images, clinical texts, and laboratory data. This approach allows information from multiple sources to be analyzed simultaneously to produce more comprehensive and accurate predictions. For example, the classification of cancer types is carried out by utilizing histopathological images combined with genomic data [35].

Transformer-based decision support systems are also developed to provide clinical recommendations automatically. Patient data analysis is carried out by integrating information from various sources, so that efficiency and accuracy in the diagnosis and treatment process can be improved. This transformation shows how artificial intelligence technology can be applied in improving the quality of healthcare services.

### 3.4. Lightweight Applications

The development of lightweight transformer models is currently focused on answering the challenges of computing efficiency and limited hardware resources. Models such as MobileViT, Tiny Transformers, and Lite Transformers are designed to maintain the advantages of self-attention mechanisms in understanding global relationships, while reducing the high computational complexity of conventional transformers. These innovations are achieved through the reduction of the number of parameters, the simplification of the self-attention mechanism with the sparse attention approach, and the application of hierarchical features that take advantage of the advantages of the convolutional neural network architecture. With this approach, the deployment of transformers on edge or mobile devices becomes more possible, resulting in increased efficiency for tasks such as image classification, object detection, and text analysis [36].

A hybrid approach is also being used to combine lightweight transformers with other architectures to improve efficiency without sacrificing accuracy. For example, local attention mechanisms applied to models such as Swin Transformer Lite and MobileViT are designed to reduce complexity while still retaining the ability to capture global relationships in data [37]. With this approach, the use of lightweight transformers in resource-constrained devices is expanding, including in Internet of Things (IoT) device applications and real-time data processing scenarios.

## 4. ADVANTAGES AND LIMITATIONS OF TRANSFORMER

### 4.1. Advantages

One of the main advantages of transformer-based architecture is its ability to effectively capture long context generalizations. The self-attention mechanism at the heart of the transformer allows every element in the data to interact with other elements, regardless of their positional distance. This makes transformers excel in tasks that require understanding global relationships, such as analyzing the order of words in text or spatial patterns in images. This ability to understand long contexts gives models a better ability to capture complex relationships that may be missed by traditional approaches such as convolutional neural networks (CNNs) or recursive models [38].

Domain flexibility is also a significant advantage of transformer architecture. Unlike other models that are often designed for specific applications, transformers can be easily adapted for a variety of tasks, such as NLP, CV, and multimodal data. For example, models like BERT and GPT show tremendous success in NLP, while Vision Transformer and Swin Transformer prove their superiority in CVs. This ability to handle data from a variety of formats and domains makes transformers a very versatile architecture in the development of artificial intelligence technology.

The combination of long-context generalization capabilities and domain flexibility has made transformers one of the most significant innovations in modern deep learning. This architecture not only allows for wider application but also provides more effective and efficient solutions to various challenges in the field of technology. The improved performance in various benchmarks and practical applications shows that transformers are able to push the limits of the capabilities of artificial intelligence models, while opening up new opportunities for further exploration and development in the future [39].

### 4.2. Limitations

Although transformer-based architectures have shown excellent performance in a variety of applications, some limitations still need to be noted, one of which is the high complexity of computing. The self-attention mechanism, which is at the heart of the transformer, is designed with a time complexity of $O(n^2)$, where $n$ is the length of the data sequence. As a result, memory consumption and computing time increase significantly, especially when large amounts of data must be processed, such as high-resolution images or long text. These challenges often hinder the deployment of transformers on devices with limited resources.

In addition, the need for large amounts of data is also a major obstacle to this architecture. To achieve optimal performance, transformer models require a very large dataset during the pretraining process. Rich representations can only be obtained if the patterns in the data are thoroughly explored, which requires a sufficient amount of data [40]. In some domains, such as medical or social sciences, large-scale data collection is often hampered by privacy, cost, or accessibility limitations, making it more difficult to use this model.

These problems of computational complexity and the need for big data have become a major concern in the development of more efficient transformers. Various efforts have been made to overcome these limitations, including the development of lighter variants, such as the Swin Transformer or MobileViT [41]. However,

these challenges continue to affect the widespread adoption of transformer architectures, especially for applications that require high efficiency on devices with low computing power.

## 5. FUTURE DIRECTIONS

Future research on transformer architectures can be directed to improve computing efficiency, reduce reliance on large datasets, expand multimodal integration, optimize real-world applications, pay attention to ethical and transparency aspects, and explore quantum and neuromorphic computing. Computing efficiency can be improved through the development of lighter self-attention mechanisms, such as sparse attention or approximative algorithms, which are able to reduce memory and computational time requirements. In addition, the design of lightweight models such as the MobileViT can facilitate the use of transformers in devices with limited resources. Dependency on large datasets can be minimized by utilizing methods such as self-supervised learning, transfer learning, or synthetic data generation, which allow exploration of domains with limited data, such as medical or local language processing [42].

The development of transformers for multimodal data, such as text, images, and audio, is also a promising research direction. Models such as Vision-Language Transformers (VLTs) can be extended with more sophisticated attention mechanisms to understand the relationships between modalities in depth. On the other hand, adapting transformers for real-world applications, such as healthcare, automotive, or cybersecurity, requires optimization in terms of inference speed, energy efficiency, and noise resistance. Cross-disciplinary collaboration with specific domains will result in more relevant practical solutions. In addition, attention to ethical aspects and transparency is increasingly important. Research in the area of interpretability, such as attention visualization, can help understand model decisions, while evaluations of potential biases should be strengthened to ensure fair and responsible application.

The exploration of quantum computing and neuromorphics is also opening up new opportunities for transformers to achieve higher efficiency and capabilities. Integration with these innovative hardware technologies has the potential to change the way transformers operate, enabling more complex and sophisticated applications. By addressing these challenges, future transformer research can further improve its efficiency, flexibility, and impact in various technology and industrial domains [43].

## 6. Conclusion

Transformers have been recognized as one of the most significant innovations in the development of deep learning technology since it was introduced by Vaswani et al. in 2017. This architecture has replaced traditional RNN and CNN-based approaches in a variety of tasks, such as natural language processing, computer vision, and medical data analysis. The main advantage of transformers lies in the ability of self-attention mechanisms to capture global relationships between data elements in parallel, which provides higher efficiency and accuracy than conventional approaches. Its flexibility and adaptability allow for wide applicability across a wide range of domains, from text classification to medical image segmentation.

Despite having significant advantages, transformers face several challenges, including high computing requirements and reliance on large datasets during training. This obstacle has prompted further research to develop lightweight models, hybrid approaches, and other efficiency solutions to expand the adoption of transformers in resource-constrained scenarios. With the development of variants such as MobileViT and Swin Transformer, steps towards greater efficiency have been taken, although challenges related to complexity remain.

Future research will focus on optimizing computing efficiency, multimodal data integration, and applications in specific domains, such as health and cybersecurity. In addition, attention to the ethical aspects, interpretability, and reduction of model bias is becoming increasingly important to ensure the responsible use of transformers. With continuous innovation, transformers are projected to continue to be the foundation in the development of more adaptive, efficient, and far-reaching artificial intelligence technologies in various sectors.

## REFERENCES

[1] A. Ma'Arif, A. I. Cahyadi, S. Herdjunanto, and O. Wahyunggoro, "Tracking control of high order input reference using integrals state feedback and coefficient diagram method tuning," *IEEE Access*, vol. 8, pp. 182731–182741, 2020, doi: 10.1109/ACCESS.2020.3029115.

[2] Y. Yang *et al.*, "Transformers Meet Visual Learning Understanding: A Comprehensive Review," pp. 1–20, 2022.

[3] H. Jiang and Q. Li, "Approximation Rate of the Transformer Architecture for Sequence Modeling," no. NeurIPS, pp. 1–30, 2023.
[4] T. Ergen, B. Neyshabur, and H. Mehta, "Convexifying Transformers: Improving optimization and understanding of transformer networks," pp. 1–22, 2022.
[5] Y. Bondarenko, M. Nagel, and T. Blankevoort, "Understanding and Overcoming the Challenges of Efficient Transformer Quantization," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 7947–7969, 2021, doi: 10.18653/v1/2021.emnlp-main.627.
[6] L. Ma, W. Zhang, R. Sun, and T. Liu, "A compare aggregate transformer for understanding document-grounded dialogue," *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pp. 1358–1367, 2020, doi: 10.18653/v1/2020.findings-emnlp.122.
[7] D. Yunia and M. I. Ibrahim, "Memprediksi Arus Kas Dengan Laba Bersih Dan Total Pendapatan Komprehensif Lain," *Monex Journal Research Accounting Politeknik Tegal*, vol. 10, no. 1, pp. 64–72, 2021, doi: 10.30591/monex.v10i1.2207.
[8] D. Szelogowski, "Deep Learning for Protein Structure Prediction: Advancements in Structural Bioinformatics," *Bioinformatics*, vol. 2023, pp. 1–8, 2023.
[9] M. Farhan Naeem *et al.*, "A novel method for life estimation of power transformers using fuzzy logic systems: An intelligent predictive maintenance approach," *Front Energy Res*, vol. 10, no. September, pp. 1–20, 2022, doi: 10.3389/fenrg.2022.977665.
[10] R. Anggrainingsih, G. M. Hassan, and A. Datta, *Transformer-based models for combating rumours on microblogging platforms: a review*, vol. 57, no. 8. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10837-9.
[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," 2020.
[12] B. Yang, B. Zhang, Y. Han, B. Liu, J. Hu, and Y. Jin, "Vision transformer-based visual language understanding of the construction process," *Alexandria Engineering Journal*, vol. 99, no. May, pp. 242–256, 2024, doi: 10.1016/j.aej.2024.05.015.
[13] M. ELAffendi and K. Alrajhi, "Beyond the Transformer: A Novel Polynomial Inherent Attention (PIA) Model and Its Great Impact on Neural Machine Translation," *Comput Intell Neurosci*, vol. 2022, pp. 1–14, Sep. 2022, doi: 10.1155/2022/1912750.
[14] S. Ren and X. Li, "HResFormer: Hybrid Residual Transformer for Volumetric Medical Image Segmentation," vol. 14, no. 8, pp. 1–10, 2024.
[15] T. Liu *et al.*, "The Role of Transformer Models in Advancing Blockchain Technology: A Systematic Survey," pp. 1–37, 2024.
[16] S. Liu, Y. Hou, Z. Xiong, Y. Fang, and L. Tong, "Study on Impact Response Characteristics of Capacitive Voltage Transformer," *J Phys Conf Ser*, vol. 1486, no. 6, 2020, doi: 10.1088/1742-6596/1486/6/062017.
[17] S. Jamil, M. Jalil Piran, and O. J. Kwon, "A Comprehensive Survey of Transformers for Computer Vision," *Drones*, vol. 7, no. 5, pp. 1–27, 2023, doi: 10.3390/drones7050287.
[18] O. Hourrane and E. H. Benlahmar, "Topic-Transformer for Document-Level Language Understanding," *Journal of Computer Science*, vol. 18, no. 1, pp. 18–25, 2022, doi: 10.3844/jcssp.2022.18.25.
[19] C. Chen *et al.*, "Understanding the brain with attention: A survey of transformers in brain sciences," *Brain-X*, vol. 1, no. 3, 2023, doi: 10.1002/brx2.29.
[20] C. Sanford *et al.*, "Understanding Transformer Reasoning Capabilities via Graph Algorithms," 2024.
[21] P. C. Chen, H. Tsai, S. Bhojanapalli, H. W. Chung, Y. W. Chang, and C. S. Ferng, "A Simple and Effective Positional Encoding for Transformers," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 2974–2988, 2021, doi: 10.18653/v1/2021.emnlp-main.236.
[22] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and Improving Relative Position Encoding for Vision Transformer," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10013–10021, 2021, doi: 10.1109/ICCV48922.2021.00988.
[23] K. M. Choromanski *et al.*, "Learning a Fourier Transform for Linear Relative Positional Encodings in Transformers," *Proc Mach Learn Res*, vol. 238, pp. 2278–2286, 2024.
[24] N. Tyagi and B. Bhushan, "Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions," *Wirel Pers Commun*, vol. 130, no. 2, pp. 857–908, 2023, doi: 10.1007/s11277-023-10312-8.

[25] Zulkarnain and T. D. Putri, "Intelligent transportation systems (ITS): A systematic review using a Natural Language Processing (NLP) approach," *Heliyon*, vol. 7, no. 12, p. e08615, 2021, doi: 10.1016/j.heliyon.2021.e08615.

[26] C. Yang and C. Huang, "Natural Language Processing (NLP) in Aviation Safety: Systematic Review of Research and Outlook into the Future," *Aerospace*, vol. 10, no. 7, pp. 1–20, 2023, doi: 10.3390/aerospace10070600.

[27] H. Haidir, T. Muhamad, R. Roviati, E. Evi, and D. Deka, "Penerapan Chat GPT dalam Pembelajaran Biologi," *Jurnal Sosial Teknologi*, vol. 4, no. 3, pp. 182–189, 2024, doi: 10.59188/jurnalsostech.v4i3.1064.

[28] S. Hadi and F. A. Diantoro, "Peluang dan Ancaman: Penggunaan Chat GPT (Generative Pre-Trained Transformer) Terhadap Praktik Akuntansi," *Jurnal Ekonomi dan Bisnis Islam (JEBI)*, vol. 4, no. 1, pp. 13–28, 2024, doi: 10.56013/jebi.v4i1.2711.

[29] E. W. Ambarsari *et al.*, "Pemanfaatan AI-Language Model Tools untuk Menunjang Copywriting Skill Jurnalis Media Have Fun," *Prioritas: Jurnal Pengabdian Kepada Masyarakat*, vol. 6, no. 01, pp. 20–28, 2024.

[30] A. El-Komy, O. R. Shahin, R. M. Abd El-Aziz, and A. I. Taloba, "Integration of Computer Vision and Natural Language Processing in Multimedia Robotics Application," *Information Sciences Letters*, vol. 11, no. 3, pp. 765–775, 2022, doi: 10.18576/isl/110309.

[31] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[32] A. Mandal, S. Little, and S. Leavy, "Multimodal Bias: Assessing Gender Bias in Computer Vision Models with NLP Techniques," *ACM International Conference Proceeding Series*, pp. 416–424, 2023, doi: 10.1145/3577190.3614156.

[33] S. Khanna, "Identifying Privacy Vulnerabilities in Key Stages of Computer Vision, Natural Language Processing, and Voice Processing Systems," *International Journal of Business Intelligence and Big Data Analytics (IJBIBDA)*, vol. 4, no. 1, 2021.

[34] Q. Pu, Z. Xi, S. Yin, Z. Zhao, and L. Zhao, "Advantages of transformer and its application for medical image segmentation: a survey," *Biomed Eng Online*, vol. 23, no. 1, pp. 1–22, 2024, doi: 10.1186/s12938-024-01212-4.

[35] K. He *et al.*, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023, doi: 10.1016/j.imed.2022.07.002.

[36] J. Liao, C. Li, and Z. Huang, "A Lightweight Swin Transformer-Based Pipeline for Optical Coherence Tomography Image Denoising in Skin Application," *Photonics*, vol. 10, no. 4, 2023, doi: 10.3390/photonics10040468.

[37] M. Gwak, J. Cha, H. Yoon, D. Kang, and D. An, "Lightweight Transformer Model for Mobile Application Classification," *Sensors*, vol. 24, no. 2, pp. 1–14, 2024, doi: 10.3390/s24020564.

[38] A. Sharma, "Solid State Transformer: An Overview of Application and Advantages," *Int J Res Appl Sci Eng Technol*, vol. 12, no. 7, pp. 335–337, 2024, doi: 10.22214/ijraset.2024.63557.

[39] B. Pacewska and I. Wilińska, "Usage of supplementary cementitious materials: advantages and limitations: Part I. C–S–H, C–A–S–H and other products formed in different binding mixtures," *J Therm Anal Calorim*, vol. 142, no. 1, pp. 371–393, 2020, doi: 10.1007/s10973-020-09907-1.

[40] D. Bischof *et al.*, "Advantages, Challenges and Limitations of Audit Experiments with Constituents," *Political Studies Review*, vol. 20, no. 2, pp. 192–200, 2022, doi: 10.1177/14789299211037865.

[41] G. P. Tsafaras, P. Ntontsi, and G. Xanthou, "Advantages and Limitations of the Neonatal Immune System," *Front Pediatr*, vol. 8, no. January, pp. 1–10, 2020, doi: 10.3389/fped.2020.00005.

[42] J. Y. Hong *et al.*, "Animal Models of Intervertebral Disc Diseases: Advantages, Limitations, and Future Directions," *Neurol Int*, vol. 16, no. 6, pp. 1788–1818, 2024, doi: 10.3390/neurolint16060129.

[43] W. Hariri, "Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing," 2023.