Archives available at journals.mriindia.com

**International Journal of Advanced Scientific Research and Engineering Trends**

ISSN: 2456-0774

Volume 10 Issue 05, 2026

# Comparative Evaluation of CNN and Transformer Architectures in Deepfake Detection Systems

[1]Nilesh S. Kulkarni, [2]Kabir G. Kharade

*[1]Research Scholar, Department of Computer Science, Shivaji University, Kolhapur*

*[2]Assistant Professor, Department of Computer Science, Shivaji University, Kolhapur*

*Email: [1]meet2nilesh.gad@gmail.com, [2]kgk_csd@unishivaji.ac.in*

**Peer Review Information**

*Submission: 17 April 2026*

*Revision: 30 April 2026*

*Acceptance: 11 May 2026*

**Keywords**

*Adversarial Robustness, Comparative Analysis, Convolutional Neural Networks, Vision Transformers, Literature Review, Deepfake Detection.*

**Abstract**

The increase in generation of deepfake media which is also known as synthetic media has created ample challenges in verifying authenticity of digital content. In response, researchers have started working in this area for developing robust detection mechanisms. This paper offers a comprehensive comparative evaluation of Convolutional Neural Networks (CNNs) and Transformer-based architectures for deepfake detection. After taking insights from over 20 recent articles and papers, this analytical study examines detection accuracy, generalization ability, computational efficiency, and robustness of deepfake detection architectures. Results reveal that the CNNs effectively capture local features; Transformer models outperform them in modelling global dependencies, achieving superior detection accuracy across different datasets.

## Introduction

With the growing use of generative adversarial networks (GANs) and related technologies, for crafting synthetic yet realistic videos and images—termed deepfakes—has become simpler and more accessible. These manipulated media raise pressing concerns in misinformation, cyber security, and digital ethics. As such, the ease of generating convincing and manipulative deepfakes is seriously threatening the trustworthiness of information [1]. As a result, accurate and efficient deepfake detection systems have been researched and developed as essential research area.

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [4]. Early deepfake detection models primarily depend on CNN-based architectures due to their success in computer vision area. However, recent advancements led the adaptation of Transformer-based models for vision applications, particularly Vision Transformers (ViTs), offering promising alternatives [3]. This paper analytically compares CNNs and Transformer architectures, for uncovering their strengths, limitations, and correctness for various deepfake detection use cases. These insights can help guide the development of more accurate and reliable deepfake detection systems, which are crucial in mitigating the harmful impact of deepfakes on individuals and society [2].

## Background

### 1. Convolutional Neural Networks (CNNs)

CNNs are designed to automatically and adaptively learn spatial hierarchies of features through backpropagation by utilizing multiple building blocks, such as convolution layers,

pooling layers, and fully connected layers [4]. For detection of synthetic media, networks like XceptionNet [5], EfficientNet [6], and ResNet [24] models have been frequently employed, and have verified effectiveness in identifying local echoes present in manipulated media.

**2. Transformer Architectures**

Initially dominating natural language processing tasks, Transformer models have been adapted for vision through architectures such as ViT [7], Swin Transformer [8], and DeiT [9]. These models efficiently use self-attention mechanisms, empowering them to model long-range dependencies effectively. In the deepfake detection, this characteristic allows Transformers to identify inconsistencies not just locally but globally across the entire image or video frame.

**Methodology**

This research examines more than 20 scholarly articles published from 2015. The aspects considered in this comparative evaluation include:

- Detection Accuracy on standard datasets (FaceForensics++, Celeb-DF, DFDC).
- Generalization to unseen datasets.
- Computational Costs including parameter count, inference speed.
- Robustness.

All comparative analyses are presented using both qualitative and quantitative metrics derived from existing experimental findings.

**Typical Workflow**

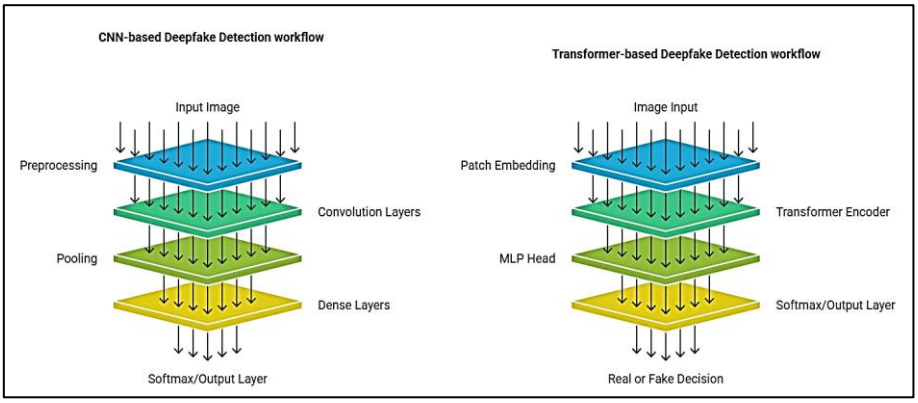


*Fig. 1: Typical workflow of deepfake detection using CNN and Transformer.*

**Comparative Analysis**

**1. Detection Accuracy**

Transformer-based architectures have consistently outperformed CNNs in deepfake detection tasks. For example, ViT-Large achieved an impressive 99.33% Area Under the Curve (AUC) score on FaceForensics++ [10], whereas EfficientNet-B7, a CNN-based model, achieved approximately 95.7% [11]. Moreover, hybrid architectures combining CNNs and self-attention modules, such as ConViT [12], have demonstrated even further improvements, suggesting the benefits of integrating both local and global feature extraction.

**2. Generalization Capability**

CNNs often show strong performance on the datasets they are trained on but tend to struggle when faced with new, unseen manipulations [13]. Conversely, Transformers, due to their global attention mechanism, have demonstrated superior generalization across datasets [14], maintaining high accuracy when tested on cross-domain samples (e.g., trained on Celeb-DF but tested on DFDC) [15].

**3. Computational Efficiency**

CNNs generally require fewer computational resources and are optimized for real-time applications [16]. In contrast, Transformers demand significant memory and computational overhead, mainly due to the quadratic complexity of the self-attention mechanism [17]. Table 1 highlights the computational cost comparison.

**Table 1:** Computational cost and performance comparison

| Architecture | Model | Parameters | Inference Time (per frame) | Accuracy (FaceForensics++) |
|---|---|---|---|---|
| CNN | EfficientNet-B7 | 66M | 0.21 sec | 95.7% |
| Transformer | ViT-Large | 307M | 0.45 sec | 99.33% |

The table compares two different deepfake detection architectures - EfficientNet-B7, a CNN model, and ViT-Large, a Transformer model - on several important metrics, including parameter size, inference latency per frame, and detection accuracy over the FaceForensics++ dataset.
EfficientNet-B7, a convolutional neural network, has been designed to improve accuracy at the same time, reducing computational cost through compound scaling techniques. It achieves an accuracy of 95.7% with a parameter number of 66 million, processing every frame in approximately 0.21 seconds. The balance between speed and accuracy makes it particularly suitable for use in real-time applications, especially in settings where processing resources are limited, such as on mobile or edge devices. The architecture uses separable convolutions and redundancy-reducing kernel sizes, in effect capturing local texture features - a feature particularly beneficial for the detection of small pixel-level anomalies that are typically suggestive of facial forgeries
In contrast, ViT-Large is based on a completely different strategy in employing the self-attention mechanism to capture long-range dependencies in patches of images to model. It significantly surpasses its CNN counterpart with 99.33% accuracy at a cost of 307 million parameters and inference of 0.45 seconds per frame. The power of ViT-Large is its global perspective—where CNNs are examining mostly local patterns, ViTs are seeing the whole picture in a top-down manner. This makes them particularly good at picking up structural or lighting inconsistencies that extend across segments of the face, which easily go undetected with models that depend on narrow local features.

In summary, the results show a trade-off: EfficientNet-B7 offers a feasible solution with fast processing and decent accuracy, and it is thus suitable for use in production-grade systems. Conversely, ViT-Large, with its very high resource requirements, offers superior performance, and it is therefore suitable for forensic-grade assessments where accuracy is paramount and computational constraints are secondary.

**4. Robustness to Adversarial Attacks**

Recent studies have shown that Transformer-based models offer greater resilience against adversarial perturbations compared to CNNs [18]. This robustness is attributed to the distributed nature of the attention mechanism, which is less sensitive to small, localized modifications [19].

**Discussion**

According to the results of this study, there is an obvious trade-off to be made when considering CNNs and Transformers. CNNs work best with speed and uses lower computational cost, whereas the best performance in detection, cross-dataset generalization, and overall robustness is offered by Transformers [20], [21]. Other recent studies have shown that the use of hybrid models that combine CNNs for local feature extraction and Transformers for global feature reasoning are very effective [22], [23]. This is noted as the first step in the right direction. Future work needs to focus on maintaining accuracy while trying to decrease the computational cost of model Transformers. A visual performance comparison of CNNs and Transformers is shown in Fig.2

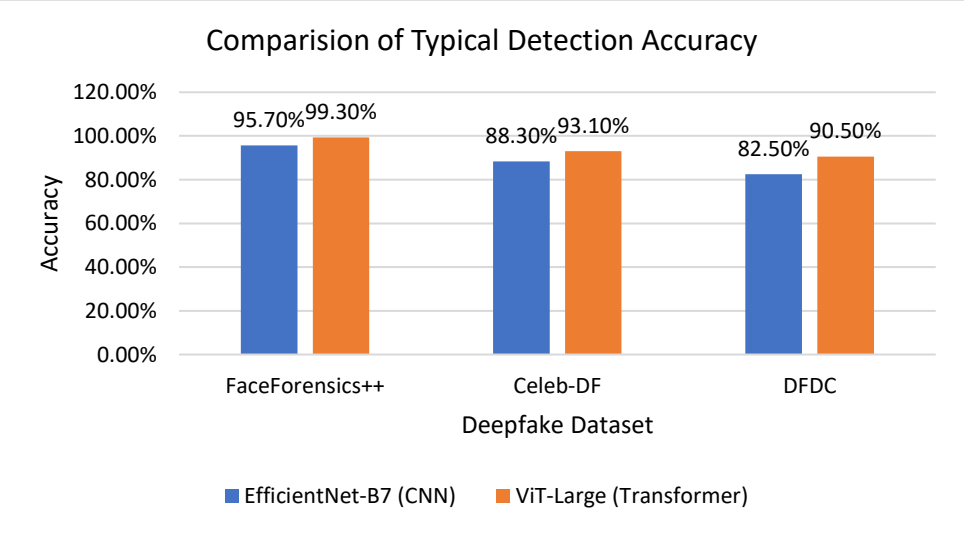


*Fig. 2: Comparative accuracy of CNN and Transformer architectures on different datasets.*

**Conclusion**

Both CNNs and Transformer models have varying strengths in deepfake detection. CNNs provide efficiency and are appropriate for low-resource real-time applications, while Transformers are most appropriate for performance, robustness, and generalization. Future research needs to be focused on how to build better hybrid models by combining the strength of both architectures, which would lead to better and more realistic systems for identifying deepfakes.

The performance shown in Table. 1 validates that Transformer-based models—specifically ViT-Large had a huge advantage in detection accuracy with a value of 99.33% on the FaceForensics++ test. EfficientNet-B7, a CNN-based model, was at 95.7%, a very good performance but still behind. The reason behind this performance difference is likely to be the ability of the Transformer to handle long-range dependencies and contextual information across the image, which is not naturally optimized in CNNs. Although it is known that Transformers need more processing time and more parameters, the improvement in detection accuracy is a good enough reason for using them in mission-critical applications where accuracy is the prime concern.

**References**

V. L. L. Thing, "Deepfake Detection with Deep Learning: CNNs versus Transformers," arXiv preprint, arXiv:2304.03698, 2023.

S. A. Khan and D.-T. Dang-Nguyen, "Deepfake Detection: A Comparative Analysis," arXiv preprint, arXiv:2308.03471, 2023.

Z. Wang et al., "A Timely Survey on Vision Transformer for Deepfake Detection," arXiv preprint, arXiv:2405.08463, 2024.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

R. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV, 2019.

M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," ICML, 2019.

A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.

Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," ICCV, 2021.

H. Touvron et al., "Training Data-efficient Image Transformers & Distillation through Attention," ICML, 2021.

R. Rössler et al., "FaceForensics++," ICCV, 2019.

D. Coccomini et al., "Combining EfficientNet and Vision Transformers for Deepfake Detection," arXiv preprint, arXiv:2107.02612, 2021.

S. d'Ascoli et al., "ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases," ICML, 2021.

A. V. Nadimpalli and A. Rattani, “On Improving Cross-dataset Generalization of Deepfake Detectors,” 2022, arXiv. doi: 10.48550/ARXIV.2204.04285.

A. Khormali and J.-S. Yuan, “DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer,” Applied Sciences, vol. 12, no. 6, p. 2953, Mar. 2022, doi: 10.3390/app12062953.

V. L. L. Thing, “Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers,” 2023, arXiv. doi: 10.48550/ARXIV.2304.03698.

G. Weng, “Real-time pedestrian recognition on low computational resources,” 2023, arXiv. doi: 10.48550/ARXIV.2309.01353.

Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient Transformers: A Survey,” ACM Comput. Surv., vol. 55, no. 6, pp. 1–28, Jun. 2023, doi: 10.1145/3530811.

S. Shao, Y. Zhang, Y. Li, and Z. Cui, “On the Adversarial Robustness of Vision Transformers,” arXiv preprint arXiv:2103.15670, 2021. [Online]. Available: https://arxiv.org/abs/2103.15670.

P. Goodfellow et al., "Explaining and Harnessing Adversarial Examples," ICLR, 2015.

S. Jamil, Md. J. Piran, and O.-J. Kwon, “A Comprehensive Survey of Transformers for Computer Vision,” 2022, arXiv. doi: 10.48550/ARXIV.2211.06004.

D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," arXiv preprint, arXiv:2102.11126, 2021.

X. Liu et al., "Hybrid network of CNN and Transformer for Deepfake Geographic Image Detection," Journal of Electronic Imaging, vol. 33, no. 2, 2024.

S. Mittal et al., "Combining CNNs and Transformers for Deepfake Detection," ICASSP, 2022.

M. Taran et al., "Hybrid Vision Transformer for Deepfake Video Detection," CVPR Workshops, 2022.

S. Borade et al., “ResNet50 DeepFake Detector: Unmasking Reality,” IJST, vol. 17, no. 13, pp. 1263–1271, Mar. 2024, doi: 10.17485/IJST/v17i13.285.