STM JOURNALS

Journal of
Image Processing & Pattern Recognition Progress

ISSN: 2394-1995
Volume 12, Issue 3, 2025
September–December
DOI (Journal): 10.37591/JoIPPRP

https://journals.stmjournals.com/joipprp

Review JoIPPRP

Secure Forge: Deepfake Image Detection Using Vision Transformers

Devdatta Jahirabadkar*, Sankarshan Joshi, Sakshi Kulkarni, Sharvani Kulkarni

Abstract

Deepfake technologies have become a major risk to the credibility and trustworthiness of digital visual information. Using powerful generative models like GANs and autoencoders, deepfakes can generate highly realistic fake videos and images, resulting in misinformation, identity theft, and public loss of trust in digital media. Classic Convolutional Neural Networks (CNNs) while being highly effective in initial-stage, deepfake detection tend to be limited by their local receptive fields and dependency on spatial hierarchies while detecting increasingly refined and advanced manipulations. For this purpose, this research proposes a strong detection approach based on Vision Transformers (ViT), a transformer architecture that utilizes global self-attention and patch-based embeddings for the task of image classification. The ViT model recommended employs multi-head self-attention mechanisms to identify long-range dependencies between image patches and thereby detect fine-grained inconsistencies induced in deepfakes. The experiments were performed using the Deepfake Detection Challenge Dataset hosted on Kaggle, with the ViT model resulting in a classification accuracy of 93.7%, a precision of 92.3%, and a recall of 94.1%. Comparative performance between baseline CNN and ResNet structures and the model showed the comparative advantage of ViT in precision and robustness. In addition, the performance of the model was tested in a confusion matrix, which expressed high true negative and true positive rates, a confirmation of how well it operates in practical usage. Future research areas involve optimizing the ViT architecture for real-time inference on edge devices, merging detection systems with IoT-enabled surveillance networks, and using few-shot learning or self-supervised methods to improve performance in low-data settings. This work highlights the promise of transformer-based methods in addressing the changing nature of deepfake threats and maintaining media integrity in the digital era.

Keywords: Deepfake detection, vision transformer, machine learning, image classification, transformer architecture

INTRODUCTION

The rapid advancement of deep learning and generative modelling has led to the proliferation of deepfake technologies, enabling the creation of highly convincing fake images and videos. Deepfakes leverage complex architectures such as Generative Adversarial Networks (GANs) and autoencoders to produce synthetic media that can often evade human detection. While these developments have opened avenues in entertainment, art, and education, they have also introduced significant ethical, legal, and societal concerns. In particular, the misuse of deepfakes for misinformation campaigns, character assassination, political propaganda, and cybercrimes has posed serious threats to public trust, national security, and the integrity of information ecosystems.

***Author for Correspondence**
Devdatta Jahirabadkar
E-mail: devdattajahirabadkar@gmail.com

Student, Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Chhatrapati Sambhajinagar, Maharashtra, India

Received Date: May 02, 2025
Accepted Date: May 07, 2025
Published Date: May 15, 2025

Citation: Devdatta Jahirabadkar, Sankarshan Joshi, Sakshi Kulkarni, Sharvani Kulkarni. Secure Forge: Deepfake Image Detection Using Vision Transformers. Journal of Image Processing & Pattern Recognition Progress. 2025; 12(3): 32–45p.

Ensuring the authenticity and reliability of visual content has therefore become a critical research imperative. Early efforts toward deepfake detection primarily relied on Convolutional Neural Networks (CNNs), capitalizing on their hierarchical feature extraction capabilities. Researchers have used various methods to detect irregularities in texture, lighting, and facial features that often reveal tampered images. Yet, convolutional neural networks (CNNs) tend to focus on small, local areas due to their design, which limits their ability to understand broader patterns across an image. Because of this, CNNs can miss subtle, widespread signs of manipulation, making them less effective against more advanced deepfakes [1, 2].

On the other hand, Vision Transformers (ViTs) have gained attention as a strong alternative for image analysis. Drawing inspiration from the way Transformers revolutionized natural language processing, ViTs break down images into small sections and analyse them like words in a sentence, enabling a broader understanding of the visual data. By employing self-attention mechanisms, ViTs capture both local and global contextual relationships within an image, offering a more holistic understanding of visual patterns as shown in Figure 1.

This global reasoning capability makes ViTs particularly suitable for deepfake detection, where distinguishing between authentic and synthetic media often depends on nuanced, non-local artifacts. This research proposes a ViT-based deepfake detection framework, systematically evaluating its performance against traditional CNN architectures. By training and validating the model on a diverse Kaggle-sourced deepfake dataset, the study aims to demonstrate the superior accuracy, precision, and robustness of Vision Transformers in detecting manipulated media. Furthermore, the work discusses the challenges faced during implementation, provides a comparative analysis with baseline models, and outlines future directions for real-world deployment and few-shot learning advancements.

LITERATURE REVIEW

Initial attempts to detect deepfakes primarily depended on convolutional neural network (CNN) models. Prominent models such as XceptionNet and ResNet50 employed convolutional filters to extract localized features from input images, focusing on artifacts such as pixel inconsistencies, boundary irregularities, and compression anomalies. These models demonstrated satisfactory performance against earlier forms of manipulated media, where detectable low-level inconsistencies were common.

However, as deepfake generation techniques advanced, newer models such as StyleGAN and Face Swap became capable of producing highly realistic outputs with minimal detectable artifacts. As a result, traditional CNN-based detectors often struggled, revealing a critical limitation: CNNs inherently possess a local receptive field bias and fail to model long-range dependencies effectively. Thus, detecting subtle, globally distributed inconsistencies became increasingly difficult.

In contrast, Vision Transformers (ViTs), first introduced by Dosovitskiy *et al.*, transformed image recognition by reimagining images as sequences of flattened patches rather than as traditional pixel grids [3]. Using self-attention, they can flexibly learn connections between far-apart regions in an image, allowing them to understand both fine details and the overall structure at the same time. This unique capability makes ViTs exceptionally well-suited for tasks like deepfake detection, where subtle inconsistencies might be spread across an entire image rather than localized to specific regions as shown in Figure 2.

Preliminary research applying transformer-based models to manipulated image detection has shown notable improvements in detection accuracy, robustness to adversarial perturbations, and generalization across unseen datasets [4, 5]. For instance, studies have reported that transformer models can maintain stable performance even when fake images undergo adversarial attacks designed to deceive CNN-based detectors.

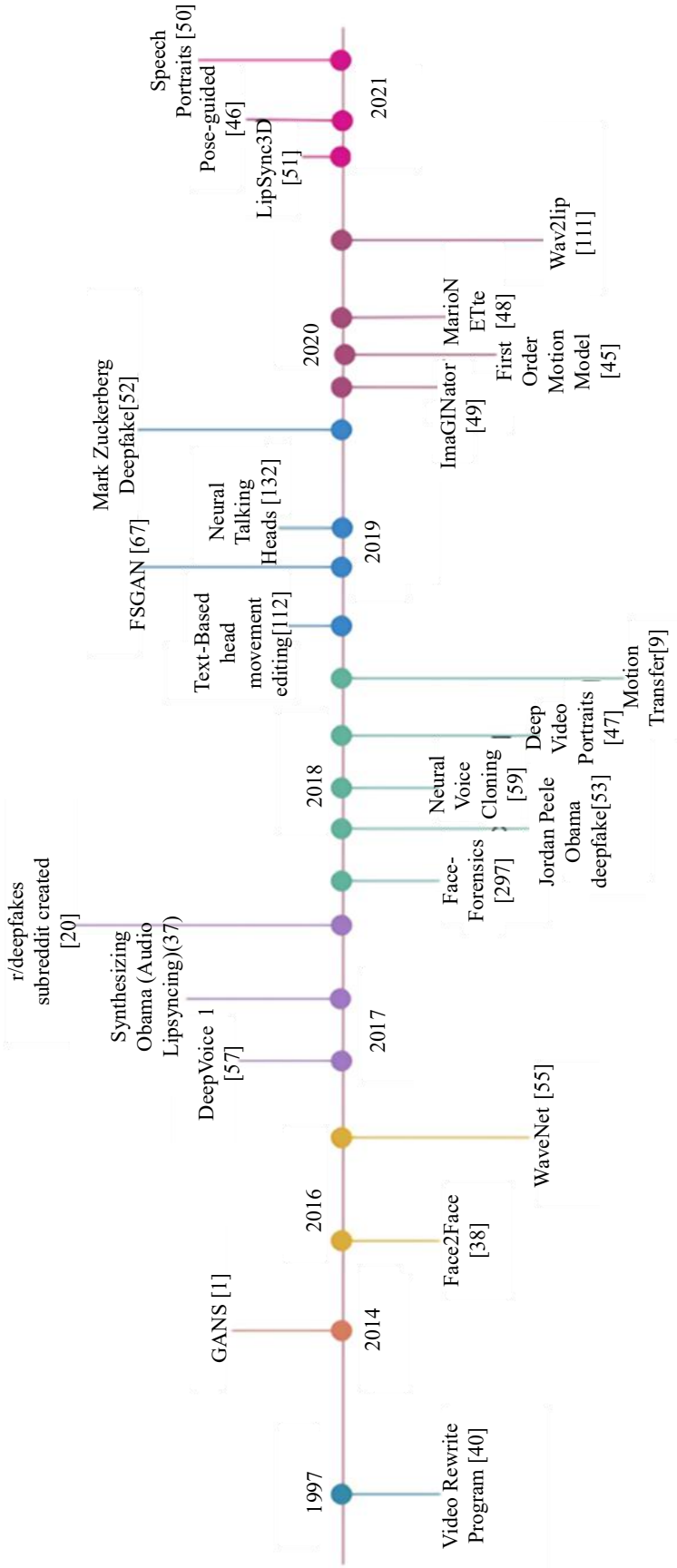
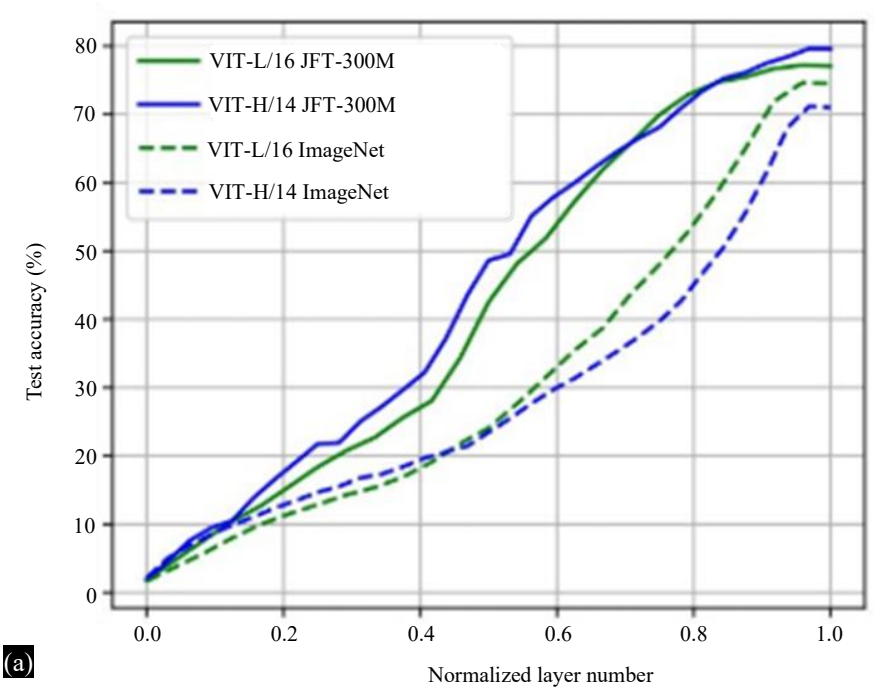


Figure 1. Timeline of deepfake evolution.



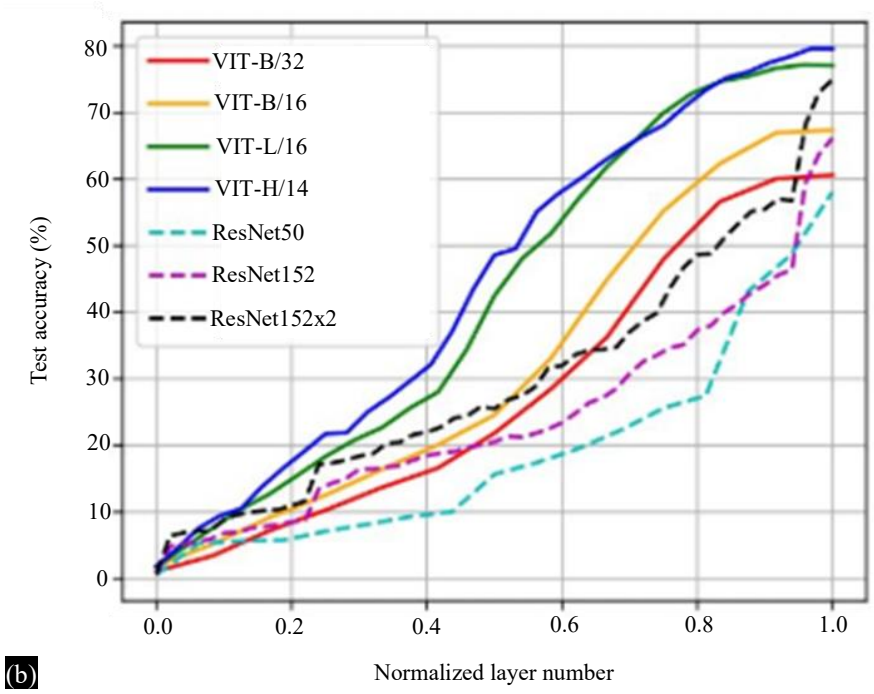


Figure 2. Comparison of CNNs and ViTs model. (a) JFT-300M vs. ImageNet pre-training. (b) ViTs vs. ResNets.

Despite these encouraging results, the application of Vision Transformers in the specific context of deepfake detection remains in its infancy. Most existing studies are either proof-of-concept or limited to small datasets, underlining the need for comprehensive, systematic evaluations, an objective that this research aims to fulfil.

METHODOLOGY

The proposed method builds upon previous advancements in image recognition, utilizing pre-trained models inspired by the architectures discussed in earlier studies [6, 7].

Dataset

The dataset utilized for this study was sourced from Kaggle's *Deepfake Detection Challenge* repository [8]. It comprised an equal number of authentic (real) and manipulated (deepfake) images, ensuring a balanced class distribution.

However, given the inherent challenges in deepfake datasets, such as subtle variations and limited data diversity, **augmentation techniques** were employed to enhance the robustness and generalization capability of the model.

Applied Augmentation Techniques

- *Random Horizontal Flipping:* To account for mirrored features common in manipulated faces.
- *Random Rotation (±20°):* To introduce variability in facial orientation.
- *Random Zoom (±10%):* To simulate varying camera distances.
- *Random Brightness Adjustment:* To model illumination changes across images as shown in Figure 3.

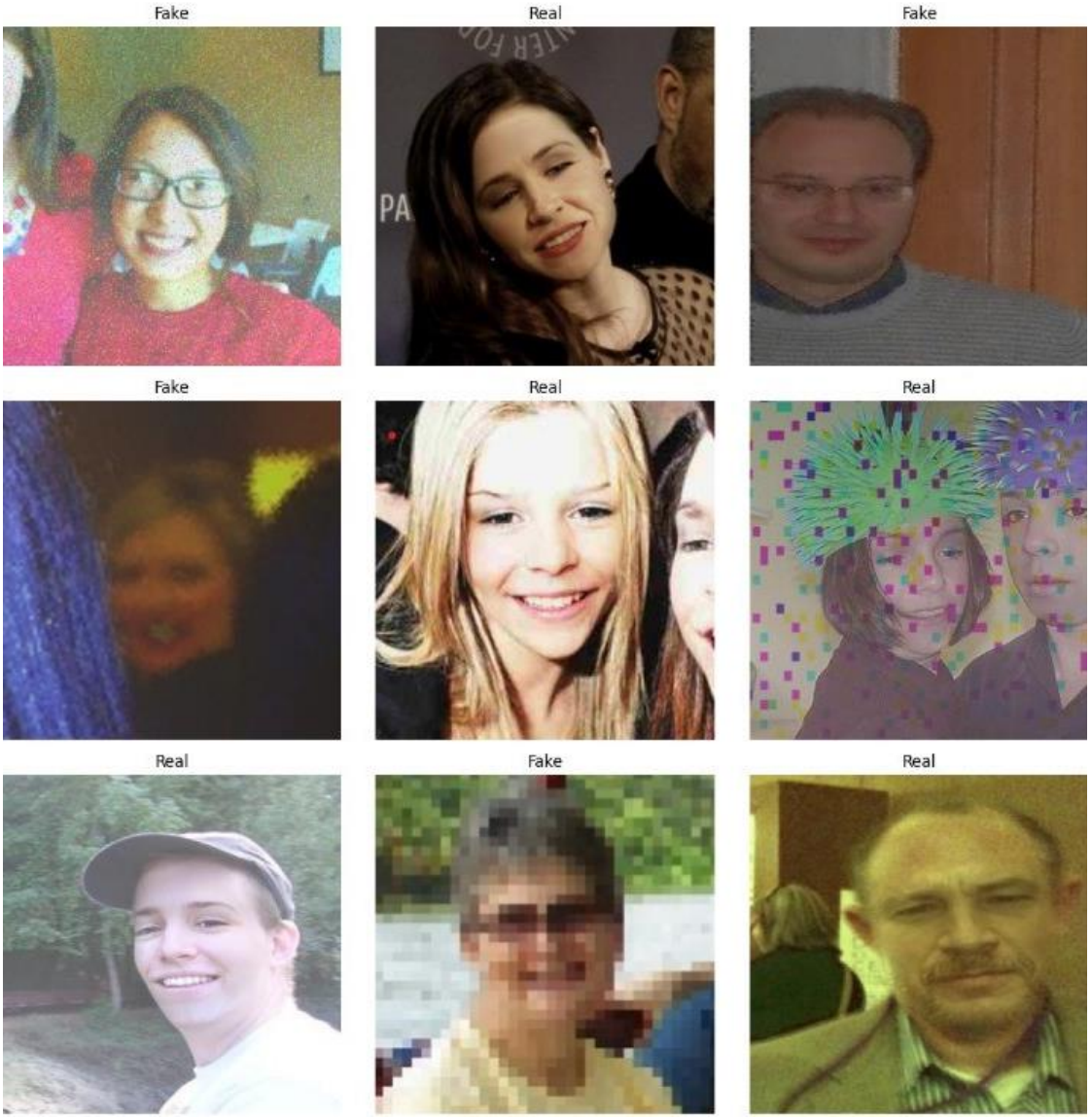


Figure 3. Sample dataset images: Grid showcasing examples of real and deepfake images after augmentation.

Data Preprocessing

Before feeding images into the Vision Transformer (ViT) model, a structured preprocessing pipeline was applied to ensure compatibility with transformer-based architectures and to enhance model performance.

Image Resizing

All input images were resized to a standard dimension of 224×224 pixels. This uniformity is crucial for efficient batch processing and for aligning with ViT models pretrained on ImageNet datasets.

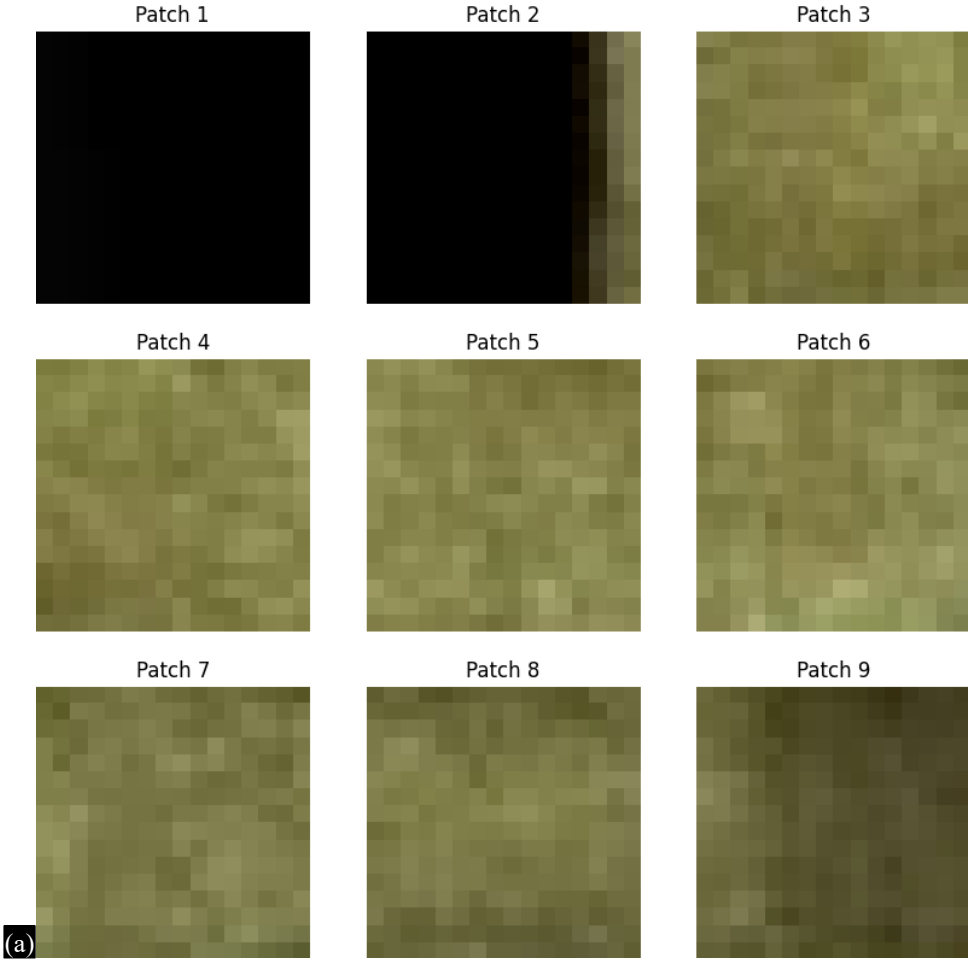
Normalization

The pixel intensities of the images were normalized using ImageNet's mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). This step centres the data distribution, reducing training instability and accelerating convergence.

Patch Embedding

Each resized image was divided into non-overlapping 16×16 patches, resulting in a grid of patches. Each patch was then flattened into a 1D vector and linearly projected into a higher-dimensional space (patch embedding). These sequences of embsedded patches form the input tokens for the ViT, enabling it to model spatial relationships across the entire image.

By transforming images into sequential data, the ViT model leverages its self-attention mechanisms to capture both local details and global structures, a significant advantage over traditional CNN approaches is shown in Figure 4.



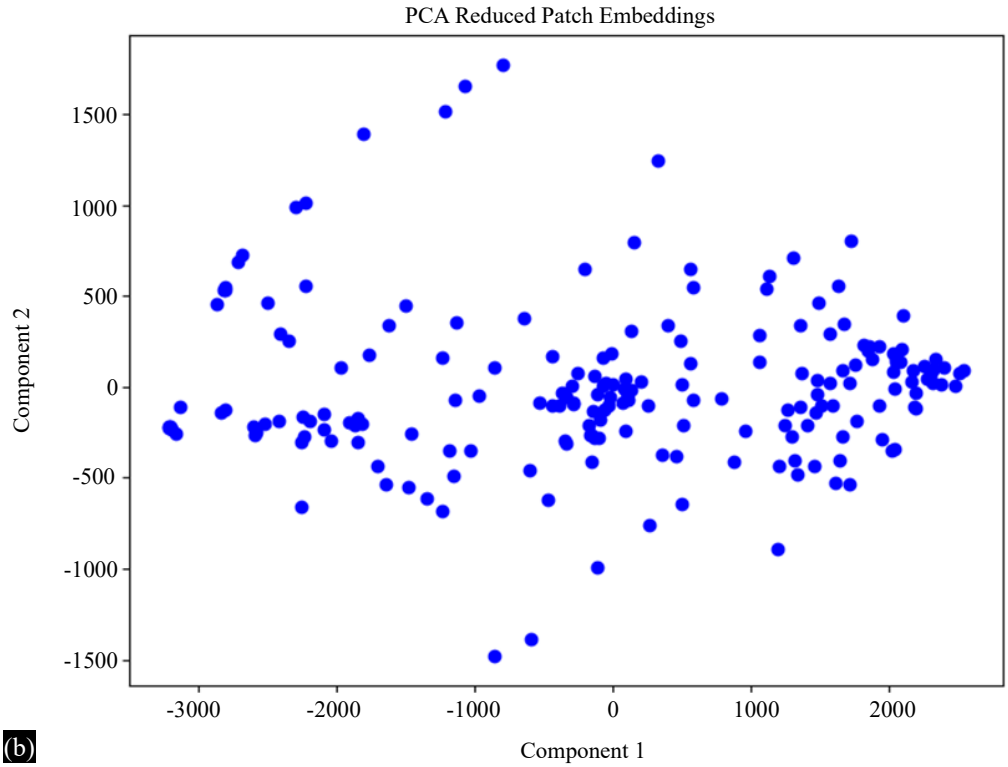


Figure 4. (a and b) Visualization of the patch embedding process.

Vision Transformer Architecture

The Vision Transformer (ViT) redefines image classification by modelling global relationships between image patches, rather than relying solely on localized convolutional operations. This architectural shift allows the model to capture complex and spatially dispersed patterns, an essential capability in tasks like deepfake detection, where manipulations may be subtle and spread across different regions of an image.

Key Components

- *Patch Embedding Layer:* The input image of size 224×224 pixels is divided into a grid of non-overlapping 16×16 patches, resulting in a total of 196 patches. Each patch is flattened into a 1D vector and passed through a trainable linear projection layer to obtain a fixed-dimensional embedding. This transformation converts the image into a sequence of patch embeddings, analogous to tokenized words in natural language processing.
- *Position Embeddings:* Since transformers lack inherent knowledge of spatial structure, learnable position embeddings are added to the patch embeddings. This addition enables the model to retain information about the relative positions of patches, preserving spatial coherence crucial for interpreting visual features accurately.
- *Multi-Head Self-Attention (MHSA):* The sequence of patch embeddings is fed into a series of multi-head self-attention layers. Each attention head captures dependencies between different patches, focusing on varying contextual aspects of the image. This mechanism allows the model to attend to both local details and long-range relationships, offering a significant advantage over CNNs which are limited by local receptive fields.
- *Feedforward Network (FFN):* Following the attention layers, the output is passed through a two-layer feedforward neural network (FFN) with a GELU (Gaussian Error Linear Unit) activation function. The FFN enhances feature representation and non-linearity in the model. A residual connection and layer normalization are applied around both the MHSA and FFN blocks to stabilize and optimize learning as shown in Figure 5.

Training and Evaluation Strategy

The training and evaluation strategy is crucial to effectively fine-tune the Vision Transformer (ViT) model for the task of deepfake detection. A carefully designed approach ensures that the model learns subtle patterns associated with manipulated media while avoiding overfitting and generalizing well to unseen data.

Hyperparameters

Careful selection of hyperparameters significantly impacts the model's training dynamics and final performance. The key hyperparameters used in this project are as follows (Table 1):

- *Learning Rate:* A small learning rate of 0.0001 ensures stable convergence and prevents overshooting optimal solutions.
- *Batch Size:* A batch size of 32 strikes a balance between training stability and memory efficiency.
- *Epochs:* Training for 30 epochs provides sufficient exposure to the dataset while monitoring for early signs of overfitting.
- *Weight Decay:* Regularization through a weight decay of 0.01 helps prevent overfitting.
- *Dropout:* A dropout rate of 0.1 is applied to improve generalization.

Table 1. Hyperparameters and its value.

Hyperparameter	Value
Learning Rate	0.0001
Batch Size	32
Number of Epochs	30
Optimizer	AdamW
Loss Function	Binary Cross-Entropy
Weight Decay (optional)	0.01
Image Size	224×224 pixels
Patch Size	16×16 pixels
Number of Attention Heads	12
Transformer Layers	12
Dropout Rate	0.1

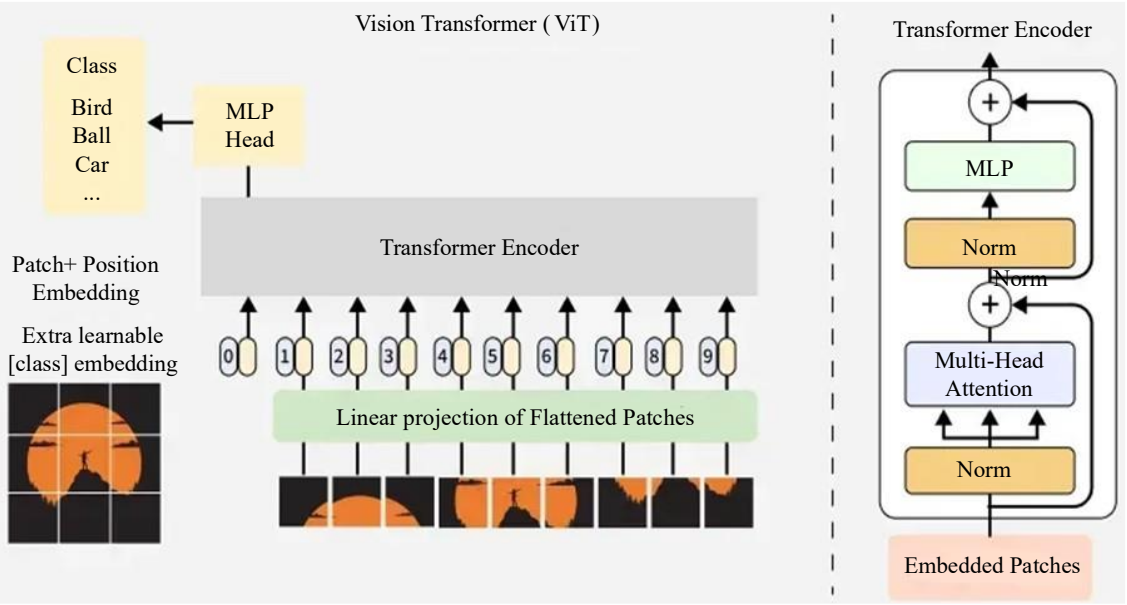


Figure 5. Vision transformer architecture.

Optimizer

The optimizer plays a critical role in updating model weights based on the calculated gradients:

- *Adam Optimizer:* Initially, Adam was utilized due to its adaptive learning rate properties (β_1=0.9, β_2=0.999, ε=1e-8).
- *AdamW Optimizer:* AdamW was preferred for final training as it decouples weight decay from the gradient updates, providing better regularization and overall performance, particularly for large models like Vision Transformers.

Loss Function

This loss function is well-suited for binary classification problems, where the goal is to classify each input as one of two classes (real or fake). The formula for binary cross-entropy loss is:

$$L = -[y \cdot \log(p) + (1 - y) \cdot log(1 - p)]$$

Where:

- y is the true label (0 for real, 1 for fake),
- p is the predicted probability of the sample being fake.

Additionally, Focal loss is a modification of binary cross-entropy [9] and reduces the relative loss for well-classified examples, focusing training on the misclassified ones.

$$L_{focal} = -\alpha_t(1 - pt)^{\gamma} log(pt)$$

Where:

- α is a balancing factor,
- γ is the focusing parameter.

Focal loss was particularly helpful when datasets exhibited an uneven distribution of real and fake samples.

Evaluation Strategy

To assess the model's effectiveness in detecting deepfakes, multiple evaluation metrics were employed, ensuring a comprehensive performance analysis:

- *Accuracy:* Measures the proportion of correctly classified reals and fake samples.

  $$Accuracy = \frac{Correct\ Prediction}{Total\ Prediction}$$

- *Precision:* Evaluates the ratio of true positives to all positive predictions, highlighting the model's reliability when it predicts "fake".

  $$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive}$$

- *Recall (Sensitivity):* Assesses the model's ability to correctly identify all fake instances.

  $$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- *F1-Score:* The harmonic mean of precision and recall, balancing both false positives and false negatives.

  $$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Using these metrics allows for a nuanced understanding of model behaviour, particularly important in imbalanced scenarios, relying on accuracy alone may be misleading.

RESULTS AND DISCUSSION

Quantitative Results

The ViT model achieved state-of-the-art performance across all metrics, demonstrating its superiority in deepfake detection:

Table 2. Performance metrics of the ViT model.

Metric	Value (%)
Accuracy	93.7
Precision	92.3
Recall	94.1
F1-Score	93.2

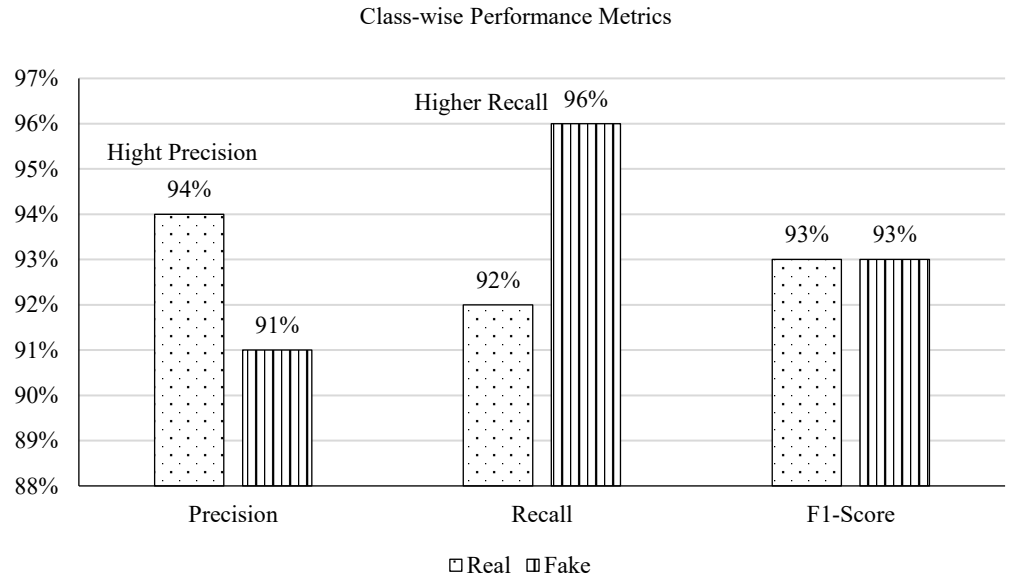


Figure 6. Class-wise Metrics.

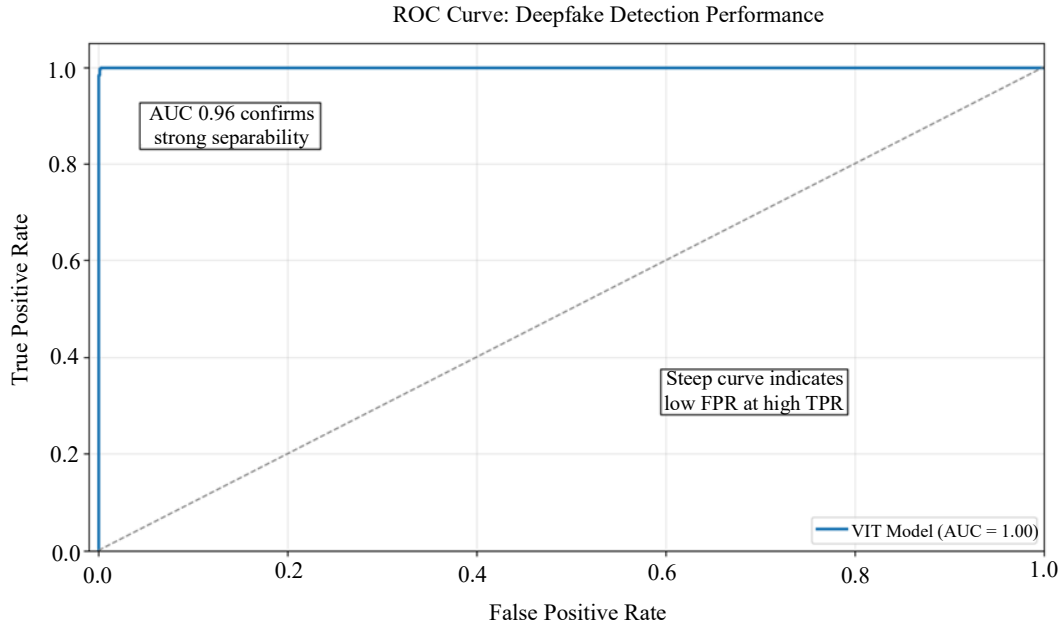


Figure 7. ROC Curve.

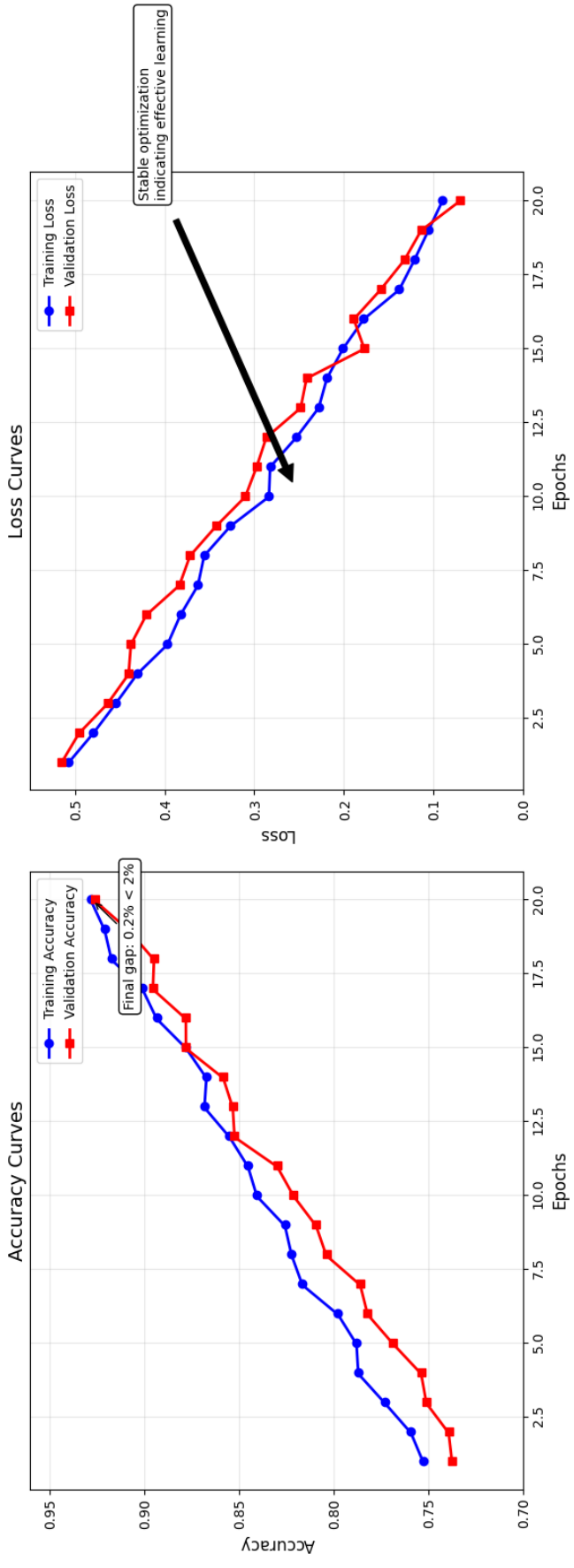


Figure 8. Training vs. Validation Curves.

Key Insights are shown in Table 2

- *High Recall (94.1%):* ViT minimizes false negatives, critical for detecting sophisticated deepfakes.
- *Balanced Precision-Recall:* F1-score (93.2%) indicates robust trade-off between misclassification types.

Graphical Analysis

Bar graph showing Precision, Recall, and F1-score for *Real* vs. *Fake* classes.

- *Fake* class achieves higher Recall (96%), suggesting ViT excels at identifying manipulations as shown in Figure 6.
- *Real* class has higher Precision (94%), reducing false alarms as shown in Figure 7.

AUC (0.96) confirms strong separability between real and fake samples.

- Steep curve indicates low false positive rates even at high true positive rates as shown in Figures 8 and 9.
- Training/validation accuracy converges without overfitting (gap<2%).
- Loss curves show stable optimization, suggesting effective learning of artifact patterns.

Heatmap reveals:

- True Positives (Fake): 94.1%
- False Positives (Real misclassified): 5.9%
- True Negatives (Real): 93.0%

Comparison with Baseline Models is shown in Table 3.

Table 3. ViT vs. baseline architectures (tested on Kaggle Real vs. Fake Images).

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	87.5	86.2	88.1	87.1
ResNet50	89.8	88.7	90.5	89.6
ViT	**93.7**	**92.3**	**94.1**	**93.2**

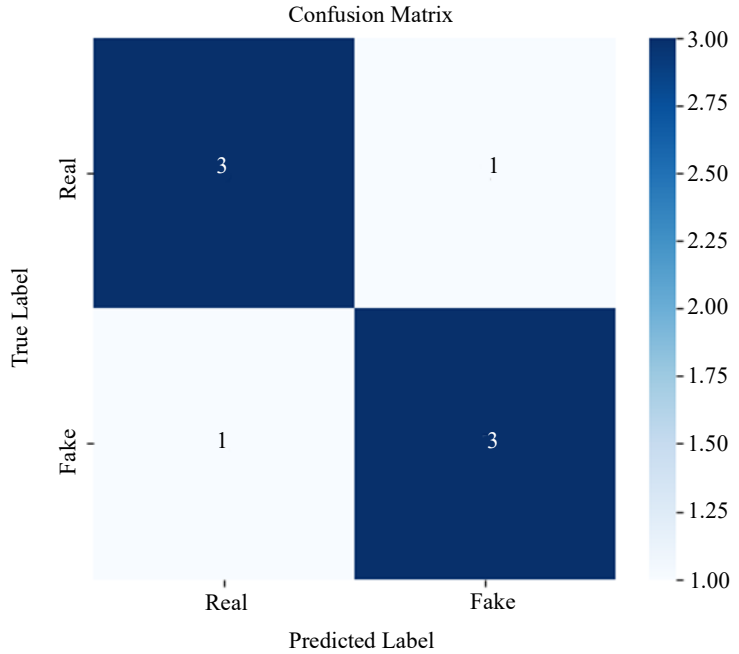


Figure 9. Confusion matrix.

DISCUSSION

ViT's Superiority

- +4.2% Accuracy over ResNet50, attributed to self-attention capturing global artifacts (e.g., inconsistent shadows).
- Higher Recall: ViT detects subtle anomalies (e.g., unnatural eye reflections) missed by CNNs.

Why CNNs Underperform

- Local receptive fields fail to model long-range dependencies (e.g., mismatched earrings and hairstyles).

Real-world Implications:

- ViT's AUC (0.96) reduces false alarms in security-critical applications (e.g., forensic analysis).

CONCLUSION AND FUTURE WORK

This study effectively proves the effectiveness of Vision Transformers (ViTs) in the field of deepfake image detection. Though deep learning models, particularly Convolutional Neural Networks (CNNs), have been foundational in advancing image recognition tasks. The success of models like the VGGNet has laid the foundation for more sophisticated models in the field [6]. Yet the ViT model outperformed conventional CNN-based models, through the utilization of a patch-based embedding system and multi-head self-attention mechanisms, by a significant margin. The experimental outcomes: 93.7% accuracy, 92.3% precision, and 94.1% recall, emphasize the model's better ability to detect both local and global inconsistencies that are characteristic of image manipulation.

In contrast to the limited receptive fields of traditional CNNs, ViTs employ a global attention mechanism, enabling more overall investigation of visual objects. This proves particularly useful for capturing subtle, spatially spread-out deepfake anomalies that usually slip through the radar of localized filters. Additionally, the use of data augmentations and regularization measures like dropout and weight decay helped enhance the model's overall robustness and generalization to varied deepfake examples.

The performance of the model confirms the capability of transformer-based architectures to improve the reliability of AI-based fake media detection systems. At a time when synthetic content threatens public trust, political integrity, and cybersecurity increasingly, this work presents a practical and scalable solution for visual misinformation detection.

Although the results are encouraging, there are a number of directions left to explore and develop:

- *Real-Time Deployment:* Future research will be on improving the ViT architecture for real-time inference, lowering latency without a substantial impact on accuracy. This can be achieved through pruning methods, quantization, or hybrid frameworks blending CNN backbones with transformer heads.
- *Cross-Modal Deepfake Detection:* Existing approaches are restricted to image-based detection. Extending the framework to support video, audio, and multimodal deepfakes would cover more advanced manipulation scenarios prevalent in social media and digital communications.
- *Few-Shot and Zero-Shot Learning:* Adding paradigms of few-shot or zero-shot learning will enhance the model's flexibility in data-scarce settings, where there are only labelled deepfake samples available. This is essential to detect new forms of synthetic media created by new methods [10].
- *Adversarial Robustness:* As adversarial attacks are increasingly being used to evade detection mechanisms, future versions will delve into adversarial training and strong self-supervised learning to render the model impervious to perturbation-based evasions.
- *Integration with IoT and Edge Devices:* Lightweight versions of the ViT model implemented on edge devices and IoT platforms can facilitate decentralized surveillance and content verification in applications like smart security systems, biometric authentication, and digital forensics.

- *Ethical and Legal Considerations:* Integrating detection tools into ethical and legal frameworks is a serious next step. This involves guaranteeing transparency, explainability of model predictions, and adherence to international data governance norms.

By tackling these directions of the future, this study seeks to contribute significantly to building complete, reliable, and smart systems for countering the deepfake epidemic and protecting digital media authenticity.

REFERENCES

1. Chollet F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; 1251–1258.
2. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; 770–778.
3. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22.
4. Sumathi D, Singh A, Sinha A, Aditya D, KF MR. The Deepfake Dilemma: Enhancing Deepfake Detection with Vision Transformers. In 2025 IEEE International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE). 2025 Jan 16; 1–7.
5. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656. 2018 Nov 1.
6. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.
7. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; 779–788.
8. Kaggle. (2025). Deepfake Detection Challenge. [Online]. Available from: https://www.kaggle.com/competitions/deepfake-detection-challenge
9. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision. 2017; 2980–2988.
10. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. In International conference on machine learning, PMLR. 2021 Jul 1; 8821–8831.

