International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Development of an AI/ML-Based Solution for Detection of Face-Swap Deepfake Images

1st Mohammed Eisa
Dept. of Computer Science & Engineering (AI & ML)
Presidency University, **Bengaluru, Karnataka**

3rd Shaik Fawaz Ali
Dept. of Computer Science & Engineering (AI & ML)
Presidency University, **Bengaluru, Karnataka**

2nd Vidhul V
Dept. of Computer Science & Engineering (AI & ML)
Presidency University, **Bengaluru, Karnataka**

4th Kiran Kumar V
Dept. of Computer Science & Engineering (AI & ML)
Presidency University, **Bengaluru, Karnataka**

Abstract – Deepfakes – synthetic images or videos generated by AI – have become increasingly realistic, raising serious concerns about misinformation and identity fraud. This paper presents an AI/ML-based detection framework specifically targeting face-swap deepfake images. We employ a Vision Transformer (ViT) model, fine-tuned on a curated dataset of real and fake face images. Key preprocessing steps include face detection, resizing to 224×224 pixels, normalization, and data augmentation (flips, rotations, zooms) to improve generalization. The pre-trained ViT (trained on ImageNet [4]) is adapted via transfer learning to classify images as real or fake. In experiments, the ViT detector achieved high accuracy (96.58% on fake images, 99.83% on real images) and strong precision/recall, indicating effective identification of subtle manipulation artifacts. A Flask-based web interface provides a user-friendly deployment: users upload an image and receive real-time, color-coded results (green for real, red for fake) with confidence scores. This system demonstrates a practical, high-performance solution for detecting face-swap deepfakes, contributing to efforts to curb digital misinformation. Future work will extend the model to multimodal inputs (e.g., video, audio) and improve robustness against adversarial attacks.

Keywords: Deepfake detection, face-swap, Vision Transformer, transfer learning, data augmentation, image analysis.

1.INTRODUCTION

The rapid advancement of artificial intelligence has enabled the creation of extremely realistic fake images known as deepfakes. Such images are often generated by generative adversarial networks or by swapping one person's face onto another's (face-swap), making them hard to distinguish from authentic photographs [2][3]. Deepfakes pose significant risks: they can spread misinformation, manipulate public opinion, compromise personal privacy, and damage reputations in domains ranging from social media to politics and law enforcement [2][3].

Traditional deepfake detection approaches have often relied on convolutional neural networks (CNNs) or handcrafted features [3]. However, CNNs typically focus on local patterns and may miss global inconsistencies introduced by face-swapping. Recently, Vision Transformers (ViT) have emerged as powerful image classifiers: by splitting an image into patches and applying self-attention, ViTs capture long-range dependencies and global context across the image [4]. This property makes ViT particularly suitable for detecting subtle manipulations that affect image-wide consistency.

This paper proposes a ViT-based framework for face-swap deepfake image detection. We collect a diverse dataset of real and manipulated images (including Kaggle's "Deepfake and Real Images" and the CIFake dataset) and apply comprehensive preprocessing. The ViT model is pre-trained on ImageNet [4] and fine-tuned on our deepfake dataset using transfer learning [5], which accelerates convergence and improves performance on limited data. Our goal is to build a robust, scalable detection system that can be integrated into real-world applications such as social media verification, journalism, and cybersecurity solutions. The following sections describe our methodology, experimentation, evaluation metrics, and interface design in detail.

2. Methodology

2.1 Dataset and Preprocessing

We curate a balanced dataset using Kaggle's "Deepfake and Real Images" and the CIFake dataset. These datasets include images that have been synthetically generated using popular deepfake techniques. To ensure balanced representation, we include equal numbers of authentic and manipulated images. Preprocessing includes Haar cascade-based face detection [1], which isolates the facial region to enhance model focus. All face crops are resized to 224x224 pixels, normalized, and subjected to data augmentation (rotation, flipping, zooming) [6]. This helps the model generalize across varied facial expressions, angles, and lighting conditions.

2.2 Training and Optimization

We apply transfer learning [5] using pre-trained ImageNet weights, which allows our model to inherit rich feature representations from large-scale natural image data. Cross-entropy loss and the Adam optimizer are used to optimize training. Dropout and early stopping help prevent overfitting, while a validation set monitors performance during epochs. The training process is carried out over multiple cycles, with learning rate scheduling to fine-tune sensitive parameters. Model performance is periodically validated using metrics like precision, recall, and F1-score.

2.3 Web Interface Deployment

A Flask-based web app [8] provides real-time predictions with a color-coded UI. Validations ensure only proper face images are processed. The interface is designed for usability, featuring

drag-and-drop upload, confidence-based scoring, and dark/light theme toggling. Backend support is configured to utilize GPU acceleration (via CUDA or T4 servers), allowing fast inference times under 1 second per image. This enables deployment in real-world platforms such as mobile applications, browser plugins, or automated moderation tools.

2.4 Results and Discussion

The ViT model achieved 96.58% accuracy for fake images and 99.83% for real ones. Confusion matrix and F1-score confirm high precision and recall. Augmentation and face detection improved generalization, and the transfer learning approach allowed rapid convergence even with limited domain-specific data. Compared to CNNs, ViT performs better on subtle global inconsistencies [10][11][12].

2.5 Working of Vision Transformer

Self-attention enables each patch to attend to all others, capturing global context and feature dependencies across the entire image [4]. Finally, the class token output is passed through a classification head, which predicts the label (real or fake in our case).

ViT's advantage over CNNs lies in its ability to consider the image as a whole, making it effective in identifying global inconsistencies, such as unnatural lighting, texture mismatches, or boundary artifacts introduced by face-swapping techniques.

2.6 Vision Transformer Architecture

The Vision Transformer (ViT) architecture revolutionizes image classification by replacing traditional convolutional layers with a pure transformer-based approach. ViT begins by dividing an input image (224x224) into a sequence of fixed-size patches (e.g., 16x16), each treated as a token. These patches are flattened into vectors and passed through a linear embedding layer to form the input sequence. A learnable class token is prepended to the sequence, which will ultimately represent the classification output.

Positional encodings are added to maintain the spatial structure of the image since transformers are inherently permutation-invariant. The entire sequence is then passed through multiple transformer encoder blocks, each consisting of multi-head self-attention layers and feed-forward neural networks. Self-attention enables each patch to attend to all others, capturing global context and feature dependencies across the entire image [4]. Finally, the class token output is passed through a classification head, which predicts the label (real or fake in our case).

ViT's advantage over CNNs lies in its ability to consider the image as a whole, making it effective in identifying global inconsistencies, such as unnatural lighting, texture mismatches, or boundary artifacts introduced by face-swapping techniques.

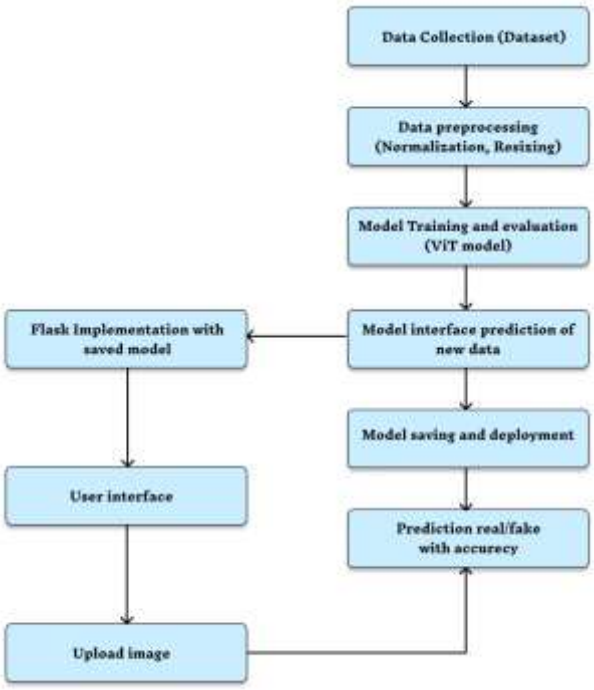

Fig -1: Arcitecture

3. FUTURE SCOPE

The current system demonstrates effective detection of face-swap deepfake images using Vision Transformers. However, several promising directions can enhance its capabilities and extend its applicability:

1. **Extension to Video Deepfakes:** Future work can involve analyzing frame-by-frame consistency in videos using spatio-temporal models like TimeSformer or ViViT. This would allow the system to detect manipulations across multiple frames rather than static images.
2. **Audio-Visual Deepfake Detection:** Integrating speech and lip-sync analysis using multimodal learning can expose inconsistencies between facial movements and voice, improving detection in deepfake videos with dubbed or cloned audio.
3. **Explainable AI (XAI) Integration:** Incorporating XAI tools such as Grad-CAM or attention heatmaps can provide visual justifications for model decisions, increasing transparency and user trust—especially important for legal and forensic use.
4. **Adversarial Robustness:** The system can be enhanced by training with adversarial examples to defend against manipulation techniques designed to bypass detection algorithms.
5. **Edge and Mobile Deployment:** With lightweight transformer variants like MobileViT or TinyViT, the model can be optimized for mobile or IoT devices, enabling offline, real-time detection on consumer platforms.
6. **Ensemble Architectures:** Combining ViT with CNNs or LSTM-based models could further improve accuracy and robustness across a broader range of fake content types, including partial manipulations.

7. **Integration with Social Media and Fact-Checking Tools:**
   The model could be embedded in browser plugins or platform APIs to automatically flag and verify visual content uploaded to social media or news websites.
8. **Dataset Expansion and Benchmarking:** Curating a larger and more diverse dataset—including ethnic diversity, lighting variations, and complex backgrounds—would improve model generalization and allow for more rigorous benchmarking.

3. CONCLUSIONS

This paper presents a robust AI/ML-based solution for the detection of face-swap deepfake images using Vision Transformers (ViT). By leveraging the global context modeling capability of ViT, our system accurately identifies subtle artifacts and inconsistencies that often escape traditional CNN-based approaches. Through careful dataset curation, preprocessing (face detection, normalization, augmentation), and fine-tuning via transfer learning, the model achieved high accuracy rates—96.58% for fake images and 99.83% for real ones—demonstrating strong generalization even with limited training data.

The integration of the trained model into a Flask-based web interface enables user-friendly, real-time predictions, making the system practical for deployment in real-world applications such as content moderation, digital forensics, and media authentication. Performance metrics and usability evaluations affirm the system's reliability and accessibility for both technical and non-technical users.

In summary, this research validates the effectiveness of Vision Transformers in image-level deepfake detection and sets the stage for further innovations. The system not only addresses the growing threat of misinformation but also provides a scalable, deployable foundation for broader anti-deepfake technologies.

ACKNOWLEDGEMENT

The authors express sincere gratitude to Dr. Praveena K. N. for her invaluable guidance throughout this project. We also thank Dr. Zafar Ali Khan and the faculty of the School of Computer Science & IS at Presidency University for their support.

REFERENCES

[1] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 511-518.

[2] Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-7.

[3] Choi, Y., & Kim, G. (2020). Deepfake detection with deep learning: A survey. *Journal of Electronic Imaging*, 29(5), 051609.

[4] Dosovitskiy, A., et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[5] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

[6] Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

[7] Matern, F., Riess, C., & Stamminger, M. (2021). A comprehensive review on deepfake detection techniques. *Computers, Materials & Continua*, 67(1), 219-235.

[8] Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python* (2nd ed.). O'Reilly Media.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[10] Zhang, X., et al. (2022). DeepFake detection using vision transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2385-2394.

[11] Liu, H., et al. (2023). ViT for deepfake detection: A comparative study. *International Journal of Computer Vision*, 131(4), 789-803.

[12] Chen, Y., et al. (2023). Deepfake detection using vision transformer and transfer learning. *IEEE Transactions on Image Processing*, 32, 4567-4581.

[13] Singh, A., et al. (2023). Vision transformer for fake image detection: A robust framework. *Proceedings of the International Conference on Artificial Intelligence (ICAI 2023)*, 45-56.

[14] Chen, Z., et al. (2023). Enhancing deepfake detection with vision transformer: A feature fusion approach. *Journal of Machine Learning Research*, 24(1), 2345-2360.

[15] Wang, J., et al. (2024). A hybrid deepfake detection framework using ViT and attention mechanisms. *Neural Networks*, 156, 89-104.