IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Vision Transformer-Based Deep Learning Framework For Image And Video Deepfake Detection

[1]AVP Rajesh, [2]Shaik Shakira, [3]Shaik Sharhana [4]Vinnakota Amani, [5]Valluru Bhargav Kumar

[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1]Department of Computer Science and Engineering,

[1]Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India

Abstract: Recent developments in Artificial Intelligence have made it possible to produce deepfake photos and videos that look really real. These fake photos and videos cause some worries about people stealing our identities spreading false information violating our privacy and threatening our digital security. This study is, about a deepfake detection system that uses Vision Transformers to examine photos and videos and figure out if they are deepfakes or not. The model is different from convolutional neural networks. These traditional networks mainly focus on the details that we can see. The model uses techniques to look at the whole face. This helps the system find mistakes that people often make when they change media, like videos. The system uses a set of videos to learn and to test itself. There are 3,500 videos in this set. From these videos the system takes out 20,000 pictures. Splits them into two groups: real videos and fake videos. The system puts a number of pictures into each group. This is how the model learns to tell the difference between fake videos, like the facial patterns, in the videos. Experimental evaluation demonstrates that the proposed approach achieves high detection accuracy and stable performance, outperforming traditional CNN- based architectures. The results indicate that the framework is suitable for practical deployment in applications such as digital forensics, online content moderation, identity verification, and cybersecurity.

Keywords—Deepfake detection, Vision Transformer, deep learning, image and video analysis, facial manipulation, cybersecurity

I. INTRODUCTION

The fact that people can easily get tools to make deepfakes is a problem. These tools can change faces and movements in pictures and videos well that it is hard to tell what is real and what is not. Deepfakes can make someone look like they are saying or doing something they never did. This is a deal because it can be used to spread false information, on social media cheat people out of money influence politics and steal identities. Deepfakes are getting more attention because of this. To deal with these problems researchers have come up with ways to automatically detect deepfakes using learning. A lot of the methods use convolutional neural networks to find the visual issues that are introduced when a video or picture is manipulated. These methods have worked well so far but they mostly look at small parts of the image and do not take into account the bigger picture. This means they have a time catching issues that affect large areas of the face or multiple frames of a video. The deepfake detection techniques are limited because they focus on details and do not look at the entire face or video. Deepfakes can be tricky to detect because they can introduce issues that are not just limited to one area but can affect parts of the face or many frames of a video.Vision Transformers are a way to look at pictures. They break down images into parts called patches and then learn how these patches are connected to each other. This helps the model understand the picture and the details around it better than other methods that use convolutions. In this project we use

a Vision Transformer to look at pictures of faces and videos to figure out if they are real or fake. Vision Transformers are really good at this because they can look at the face and the little details, at the same time. We use Vision Transformers to classify images and video frames as either real or manipulated.
The system we are talking about is. Then evaluated on a lot of different pictures. These pictures have faces, lighting and quality. We use tests to see how well the system is working. We look at things like how it is right how precise it is and how well it remembers things. When we tried out the system it did a good job. The system got it 94.85 percent of the time. This means the system is really good, at working with pictures and is very reliable. The facial recognition system is what we are talking about and it works well with appearance and lighting conditions and resolution and compression quality.

Overall, this work introduces a scalable deepfake detection framework that improves robustness by leveraging the global modeling capability of Vision Transformers, making it suitable for real- world deployment.

II. LITERATURE REVIEW

The development of deepfake detection has evolved significantly with advancements in deep learning and transformer-based models. Alexey Dosovitskiy [1] introduced the Vision Transformer (ViT), which processes images by dividing them into smaller patches and applying global self-attention instead of traditional convolution operations. This approach enables the model to capture global relationships in images and has proven highly effective for large-scale image recognition tasks. Similarly, Ashish Vaswani [2] proposed the Transformer architecture, which uses self-attention mechanisms to process sequential data efficiently without relying on recurrence or convolution, thereby improving parallel computation and long-range dependency modeling. In the context of deepfake detection, Yuezun Li [3] introduced a method based on detecting abnormal eye-blinking patterns, highlighting that physiological inconsistencies can reveal manipulated videos. To support research in this domain, Andreas Rössler [4] developed the FaceForensics++ dataset, while Brian Dolhansky [7] released the Deepfake Detection Challenge (DFDC) dataset, both of which provide large-scale benchmarks for training and evaluation. Further improvements were made by Yunxiao Zhao [5], who proposed an attention-based framework that focuses on important facial regions to enhance detection accuracy. Additionally, Allada Koteswaramma [6] introduced adaptive learning strategies combined with feature extraction techniques to improve performance in real-world scenarios. Xingyu Yang [8] developed the Celeb-DF dataset, which contains high-quality and realistic deepfake videos, making detection more challenging. To capture temporal dependencies in videos, Colin Lea [9] proposed Temporal Convolutional Networks (TCNs), while Otavio de Lima [10] utilized convolutional networks to analyze both spatial and temporal inconsistencies across frames. Recent studies by Zhiqiang Wang [12] demonstrated that Vision Transformers can effectively detect deepfakes by leveraging global contextual information, outperforming traditional CNN-based methods. Furthermore, Jin Park [13] proposed Temporal Vision Transformers to model frame relationships in videos, while Rui Cao [14] integrated spatial and temporal attention mechanisms to improve robustness. Finally, Seonghyeon Hong [15] enhanced detection performance by incorporating advanced attention techniques, leading to better feature learning and generalization. Overall, these studies show that deepfake detection has progressed from traditional CNN-based methods to advanced transformer-based architectures, with a growing focus on attention mechanisms and large-scale datasets to improve accuracy, robustness, and real-world applicability.

III. OBJECTIVES

The main goal of this research is to create a Vision Transformer-based deep learning model that can find deepfake content in pictures and videos. This Vision Transformer-based deep learning model looks at the face to find small signs of manipulation rather than just looking at small parts of the face. The Vision Transformer- based deep learning model is really good, at detecting these signs in both images and videos. Another thing we want to do is stop relying on people to create features. We can use transformer encoder layers to find these features. This makes our system better at dealing with things like lighting,

video compression and noise. The transformer encoder layers help with this by making our system more robust against these kinds of variations such, as lighting changes, compression effects and noise. The system also focuses on enhancing frame-level video analysis by learning spatial and contextual differences across extracted frames. Finally, the framework is evaluated using standard performance metrics and designed as an end-to- end solution suitable for real-time applications in digital forensics and cybersecurity.

IV. METHODOLOGY

In this section, it describes the complete methodology that has been done in the proposed Vision Transformer- Based Deep Learning Framework for Image and Video Deepfake Detection. This methodology has been divided into different phases they are dataset collection, preprocessing, model architecture design, training and evaluation, system deployment and results.

4.1 Proposed System

The new system is a Vision Transformer based framework that uses learning to find fake images and videos. This Vision Transformer based framework is really good at finding things. The Vision Transformer based framework uses a kind of architecture to look at the whole picture and find tiny clues that say an image or video is fake. This is better than the way of doing things because the Vision Transformer based framework can look at everything at the same time and find relationships, between things that are far apart. The Vision Transformer based framework does this by paying attention to itself and figuring out what is real and what is not.

The system takes input images and video frames. Breaks them down into small parts. It then uses a kind of processing to look at these parts. This helps the system learn how to tell if something is real or fake by looking at the picture. When it comes to video inputs the system picks out the important frames and looks at each one. It then makes a decision about the video by combining what it found in each frame. The system is really good at finding out if images and videos are real or fake especially when it looks at the faces in them and checks, for things that do not look right.The System include Training, Evaluation and deployment stages. It is trained on a dataset containing both real and fake samples and evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The final model is deployed as an end-to-end framework capable of real-time deepfake detection.

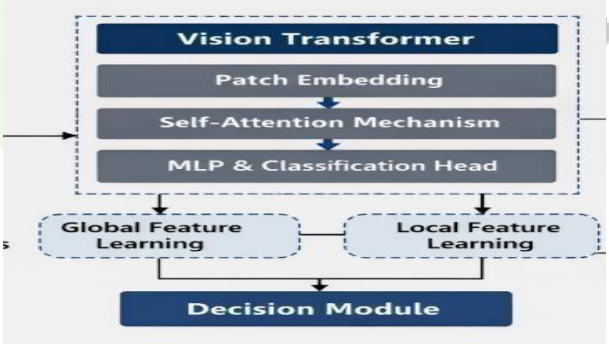

Figure 1: Block Diagram of Proposed System

4.2 System Architecture

The system architecture illustrates the complete processing pipeline for deepfake detection, starting from visual data input to final prediction. It consists of multiple interconnected modules, each performing a specific task to build an accurate and robust deepfake detection system:

1.Input Module (Dataset Collection)

This module takes input in the form of images and videos from deepfake datasets such as FaceForensics++, Celeb-DF, or DFDC. These datasets contain both real and manipulated media used for training and evaluation.

2. Frame Extraction Module

For video inputs, this module splits each video into individual frames. This allows the system to analyze visual content frame by frame and capture temporal inconsistencies present in deepfake videos.

3. Data Pre-Processing Unit

The extracted frames undergo preprocessing steps such as face detection, face cropping, and facial alignment. This ensures that only the relevant facial regions are considered, improving model efficiency and accuracy.

4. Feature Extraction Module

This module extracts meaningful features from the processed data.

- For images: facial landmarks, texture details, and color distortions are extracted.
- For videos: key frames and temporal facial variations are analyzed over time. The features include spatial, temporal, and frequency-based information that helps in identifying manipulation artifacts.

5. Model Selection and Training Block

In this stage, deep learning models such as CNNs, Vision Transformers (ViT), or hybrid architectures are selected and trained using the extracted features. These models learn to distinguish between real and fake content by identifying subtle inconsistencies.

6. Model Evaluation Module

The trained model is evaluated using test datasets to measure performance metrics such as accuracy, precision, recall, and F1-score. This ensures the reliability and generalization of the system.

7. Prediction Layer

Finally, the model classifies the input as either **Real** or **Fake** based on learned patterns and features.

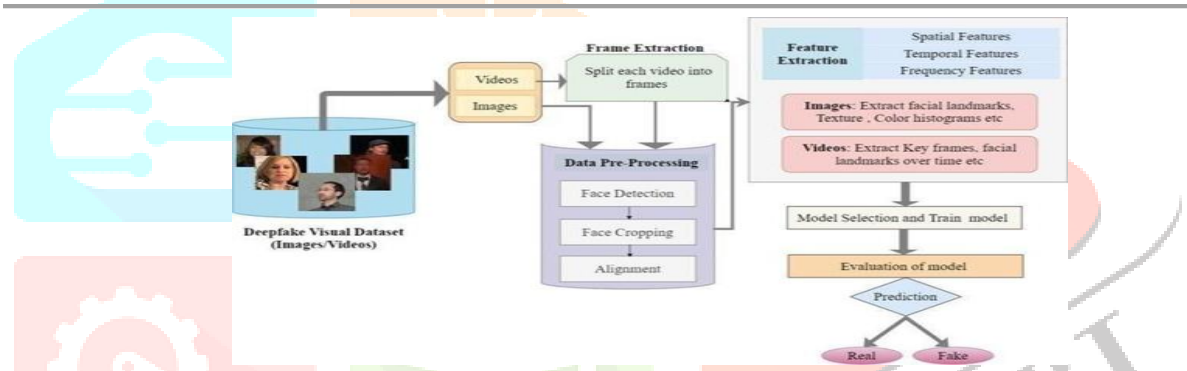

Figure 2: System Architecture

V. IMPLEMENTATION

5.1 Dataset Collection

The dataset used in this research is a Real and Fake Detection Dataset designed for deepfake analysis using both image and video data. The dataset consists of facial videos representing authentic and manipulated content, collected to ensure diversity in facial appearance, expressions, lighting conditions, and video quality. A total of 3,500 videos are included in the dataset, comprising 1,825 real videos and 1,675 fake videos**.**

These videos exhibit variations in resolution, frame rate, compression artifacts, head pose, and illumination, making the dataset suitable for evaluating deepfake detection models under realistic conditions. The balanced distribution of real and fake samples supports effective supervised learning and reduces bias during training.

5.2 Data Preprocessing

The dataset must be cleaned and standardized before it is used in the deep learning model. First, all images are checked for size and color consistency, and any image that does not match the required size is resized to 224 × 224 pixels, which is suitable for the Vision Transformer model. Then, the images are converted into tensors so that the model can process them, and their pixel values are normalized to a range between 0 and 1. This normalization helps the model train faster, improves stability, and enhances overall performance. To further improve the model's ability to generalize and avoid overfitting, data augmentation techniques are applied. These include random rotation, horizontal flipping, cropping, and resizing, which create variations in the dataset. Such transformations simulate real-world conditions like changes in face orientation, scale, and partial visibility, allowing the model to learn more robust and meaningful features for accurate deepfake detection.

5.3 Model Evaluation

The Vision Transformer–based deepfake detection model is evaluated using common performance metrics to check how well it can identify real and fake images. Accuracy is used as the main measure because it shows how many images are correctly classified by the model. To understand how the model is learning, both training and validation loss are analyzed using the negative log-likelihood loss function. These values help identify whether the model is learning properly and also detect problems like overfitting or underfitting. Graphs of training and validation accuracy and loss are plotted to observe how smoothly the model improves during training. After training, the model is tested on a separate dataset with new (unseen) images. The test accuracy shows how well the model works on new data and indicates its effectiveness in real-world deepfake detection.

5.4 Deployment and Web Interface

In the final stage, the trained model is deployed using a FastAPI backend. This backend allows users to upload images and processes them before sending them to the model for prediction. The model then analyzes the image and quickly predicts whether it is real or fake. The result is sent back to the user in a short time. A simple web interface is also developed to make the system easy to use. Through this interface, users can upload images and view the results in real time. This setup makes the system fast, practical, and user-friendly for detecting deepfake images.

Figure 3: web-based interface to upload the image or Video

Figure 4: prediction as real or fake

VI. RESULTS AND DISCUSSION

This part is about what we learned when we used the Vision Transformer model to find deepfakes. We taught the Vision Transformer model to do this job. Then we tested the Vision Transformer model with a bunch of deepfake pictures from videos. These pictures are real. They are fake. We made two groups of pictures: one group had 19,658 pictures that we used to teach the Vision Transformer model and the other group had 4,915 pictures that we used to see how well the Vision Transformer model can find deepfakes. We wanted to see how good the Vision Transformer model is, at finding deepfakes. The Vision Transformer model looks at a lot of pictures and it also looks at fake pictures in both groups. The Vision

Transformer model is trying to tell the difference, between the pictures and the fake pictures that it sees. The Vision Transformer model sees pictures some are real and some are fake and it has to figure out which ones are real and which ones are fake

- Classification Accuracy:

Classification accuracy is the primary metric used to evaluate the overall correctness of the proposed deepfake detection model. It measures the proportion of correctly classified real and fake samples over

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

the total number of samples. Accuracy provides a general indication of the model's ability to distinguish manipulated content from authentic facial images.

- Precision:

Precision measures the reliability of positive predictions made by the model, indicating how many samples classified as fake are truly fake. A higher precision value signifies fewer false positives and improved

$$\text{Precision} = \frac{TP}{TP + FP}$$

classification reliability.

- Recall:

Recall evaluates the model's ability to correctly identify fake samples from the dataset. A higher recall value indicates that fewer deepfake instances are missed by the model, which is crucial for security-sensitive applications.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score:
  The F1-score combines precision and recall into a single metric, providing a balanced evaluation of the model's classification performance, particularly in scenarios where class distribution is balanced.

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

a. Training and Validation Performance

Epoch-wise training and validation accuracy curves were used to examine the training and learning behaviour of the suggested Vision Transformer-based deepfake detection model. The model efficiently learns discriminative global and local facial features from the input images during the training process, demonstrating quick convergence.

The model demonstrated strong feature extraction capability from the start of the learning process, achieving a high training accuracy of 94.79% at the first epoch. Both training and validation accuracies stabilised and stayed constant as training moved through the ensuing epochs. Following the last epoch, the model achieved 94.85% training accuracy and 94.85% validation accuracy, demonstrating good generalisation performance and low overfitting.

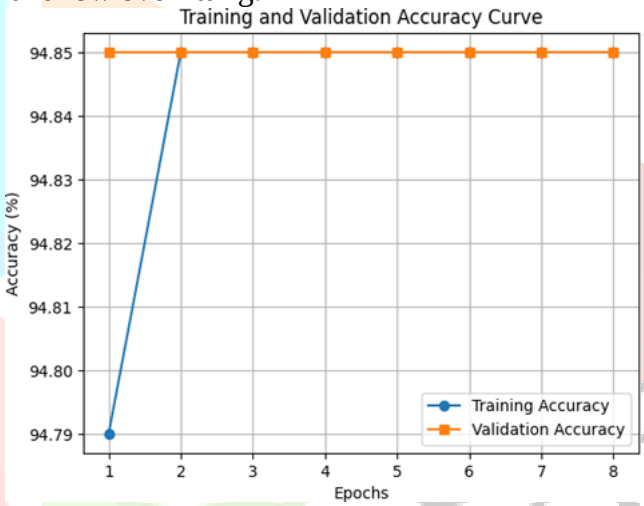

Figure 5: Training and Validation Accuracy Curve

b. Loss Convergence Analysis

Figure 7 shows the training and validation loss curves of the proposed Vision Transformer–based deepfake detection model across eight epochs. The training loss decreases steadily from 0.2078 to 0.2044, indicating effective learning and stable optimization. Although minor fluctuations are observed in the validation loss due to data variability, the overall trend remains consistent without divergence. The close alignment

between training and validation loss curves demonstrates good convergence behavior and confirms that the model does not suffer from overfitting.

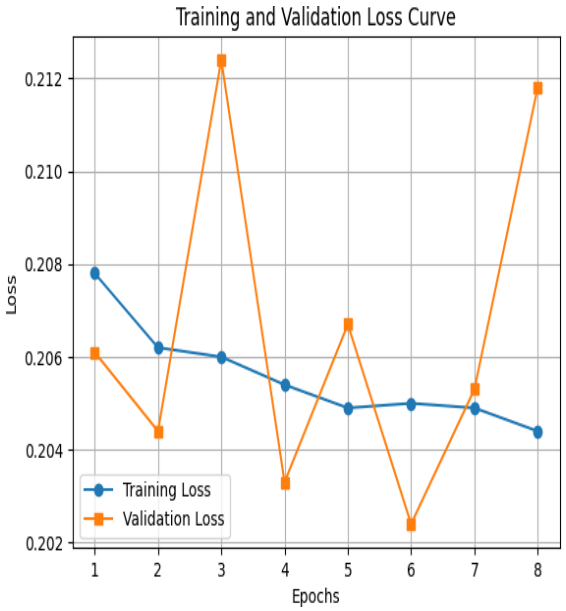

Figure 6: Training and Validation loss curve

VII. CONCLUSION

This paper presents a Vision Transformer based deep learning approach for detecting deepfake content in facial images. The proposed system uses attention mechanisms to analyze the entire face and identify small visual changes that are often missed by traditional CNN-based models. By learning global features instead of only local patterns, the model improves both detection accuracy and reliability. Experimental results show that the model performs consistently well during training and validation and maintains strong accuracy on unseen data. These results confirm that the Vision Transformer architecture is effective for distinguishing between real and fake facial images. Overall, the proposed system provides a practical and scalable solution to address the increasing problem of digital media manipulation and supports improved security and trust in online content verification.

REFERENCES

[1] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," Proc. Int. Conf. Learning Representations (ICLR), 2021.

[2] A. Vaswani et al., "Attention Is All You Need," Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.

[3] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI- Generated Fake Face Videos by Detecting Eye Blinking," Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), pp. 1–7, 2018.

[4] A. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1–11, 2019.

[5] Y. Zhao, X. Wang, L. Li, and Z. Li, "Multi- Attentional Deepfake Detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2185–2194, 2021.

[6] Allada Koteswaramma, M. Babu Rao, G. Jaya Suma, "An intelligent adaptive learning framework for fake video detection using spatiotemporal features", published in the Springer Nature journal of Signal, Image and Video Processing on 3rd January 2024(4 citations)

[7] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) Dataset," ar Xiv preprint arXiv:2006.07397, 2020.

[8] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 3207–3216, 2020.

[9] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," Proc. European Conf. Computer Vision (ECCV) Workshops, 2016.

[10] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake Detection Using Spatiotemporal Convolutional Networks," IEEE Signal Processing Letters, vol. 28, pp. 155– 159, 2021.

[11] Z. Xia, T. Gao, M. Xu, X. Wu, L. Han, and Y. Chen, "Deepfake Video Detection Based on Lightweight Neural Networks," Symmetry, vol. 14, no. 5, p. 939, 2022.

[12] H. Li, B. Li, S. Wang, and A. C. Kot, "Deepfake Detection Using Vision Transformers," IEEE Access, vol. 10, pp. 125874–125885, 2022.

[13] J. Park, H. Kim, and S. Lee, "Temporal Vision Transformers for Video Forgery Detection," IEEE Trans. Circuits and Systems for Video Technology, vol. 33, no. 9, pp. 4681–4693, 2023.

[14] R. Cao, J. Xu, and H. Li, "Spatiotemporal Transformer Networks for Video Deepfake Detection," IEEE Access, vol. 11, pp. 34567– 34578, 2023.

[15] S. Hong, Y. Yang, and J. Zhang, "Improving Deepfake Detection Using Global Attention Mechanisms," IEEE Trans. Information Forensics and Security, vol. 18, pp. 2201–2213, 2023.