Deepfake Image Detection and Classification-Vision Transformers

Shobana Gorintla[1,a)] Venkatesh Kumar Tammina [2,b)] ,Chaturya Sankarabathina[3,c)] , Hemanth Kumar Pasupuleti [4,d)] , Rakesh Venkata Naga Sai Yarajarla[5c,e)]

[1] *Professor, Department of Computer Science and Engineering, NRI Institute of Technology, Agiripalli-521212, Vijayawada, Andhra Pradesh, India*

[2,3,4,5]*B. Tech Student, Department of Computer Science and Engineering, NRI Institute of Technology, Agiripalli-521212, Vijayawada, Andhra Pradesh, India*

Abstract. Deepfake technology has advanced rapidly in recent years, generating highly realistic fake images that are increasingly difficult to distinguish from real ones. These altered media pose risks for misinformation, fraud, privacy, and the trustworthiness of society, so there is a great need for detection now more than ever. Traditional deepfake detection models mainly based on CNN have limitations on generalization and adversarial robustness. This work proposes exploring transformer-based architectures for deepfake detection since it possesses a more powerful ability in extracting global features. For this research, a pre-trained ViT model was applied using publicly available deepfake-and-real-image Kaggle datasets with advanced preprocessing steps such as resizing, normalization, and data augmentation. The proposed model, evaluated using standard parameters achieved a performance of about 96% accuracy, precision, recall, and F1 score, which is quite a massive achievement compared to traditional CNN-based methods. The confusion matrix results confirm the high classification reliability of the model, where false positives and negatives are minimized. Some experiments reveal that ViTs are capable of differentiating real from fake images and can thus be used as effective tools in deepfake detection. However, there are also challenges to address, such as the model's huge computation cost, high chance of overfitting, and the essentiality of training sets with diversity. Future work could tackle hybrid CNN-ViT models, adversarial training methods, and real-time deepfake detection systems to enhance detection accuracy and make it possible for use.

Keywords—Deepfake Detection, Vision Transformers, Image Classification, Transfer Learning.

INTRODUCTION

Fakes on the internet posed by Deep Fake which utilizes Artificial Intelligence (AI) to create fake videos and images at an alarming level of realism has become a critical issue within the bounds of Digital security. The spread of misleading images and videos has been made possible due to the application of common AI frameworks such as Generative Adversarial Networks (GANs) and Autoencoders making the distinction between genuine and fake information gravely difficult. Deep fake technology has been misused in various digital disinformation campaigns as well as fraudulent financial crimes impersonation and political crimes sensationalizing unprincipled and cyber security concerns. All of these have been made possible due to the fact that whilst traditional methods of detection struggle to keep up with the constant advancement of deepfake techniques the use of highly advanced algorithms presumingly solve the mystery of detecting such fakes. The most popular approach to spotting the manipulations done through the use of AI is through deep learning models. These advanced algorithms have the ability to skim through massive amounts of visual data to identify the smallest of discrepancies in photos and videos all in an effort of aiding the detection of Deepfake. Most of these systems concentrate in the employment of Convolutional Neural Networks (CNNs), autonomous Recurrent Neural Networks (RNNs) and hybrid deep learning models. Deepfake image and video detection with high efficiency is accomplished by systems like XceptionNet, MesoNet, Hybrid CNN-RNN and the recently proposed EfficientNet-B0 models.

To cope with these limitations, this study proposes a deepfake detection technique with the Vision Transformer (ViT) model that uses the self-attention mechanisms to improve classification by considering the global spatial relations of images. Unlike CNNs which use local features, deep ViTs which use patch embeddings are able to learn more complex and refined deepfake artifacts. The system integrates transfer learning with modern data augmentation methods to enhance the generalization and detection accuracy of the model. The implemented ViT beats the entire image structure in better detection accuracy, adversarial robustness, scalability, and speed for real time systems in comparison to the other methods.

This study takes advantage of deep fake detection by employing the newest advancements in processing method that permit higher degrees of precision in locating images that are not real. It confronts the cognition of ViTs with vertical developed. There has been an adequate amount of investigation done in deep fake detection by the traditional methods. The concept of deepfake is aligned with the concept of building tools that can be easily accessed by any generic user. Initially, the primary non-artificial approaches to the detection of deep fake video are analyzed.

LITERATURE SURVEY

Realizing how advanced AI-generated content has become, deepfake detection poses as a critical problem. The majority of deep learning algorithms, including Convolutional Neural Networks (CNNs), Capsule Networks, and Vision Transformers (ViTs) have been researched in their breadth and developed for efficient deepfake detection. Researchers have experimented with various model architectures, datasets, and training methodologies to maximize accuracy on the detection task.

Rössler et al. [1] et al. introduced FaceForensics++, one of the most widely used benchmark datasets for training deepfake image detection models. They demonstrated in their work that deep learning-based classifiers can distinguish fake images, yet the models do not transfer well to other datasets. Karras et al. [2] made StyleGAN, a GAN-based model for facial image synthesis, which made artificially created faces much more realistic and hence harder to identify. Dosovitskiy et al. [3] proposed Vision Transformers (ViTs) which use self-attention mechanisms to capture long-range dependencies in images and said to classify better than CNNs.

Several studies have explored ViTs to detect deepfakes. IEEE work [4] has proved that deepfake classifiers based on ViT outperform traditional CNN models by detecting more subtle details in forged images. highlighting the necessity for sophisticated architectures that can manage manipulations of superior quality. Capsule-Forensics was also proposed by Nguyen et al. [6], who used Capsule Networks to identify facial distortions in automatically produced videos.

Afchar et al. [7] introduced MesoNet, a computationally efficient CNN-based model for the detection of deepfakes. While being computationally efficient, MesoNet is unable to detect high-resolution deepfakes. Passos et al. [8] reviewed the various deep learning-based approaches utilized in detecting deepfakes, and they declared that transformer-based models provide better adversarial attack robustness. Tolosana et al. [9] studied methods for the detection of deepfakes and highlighted the need for spatial-temporal feature extraction in video-based detection models.

Liu et al. [10] presented Spatial-Temporal Attention Networks that leverage spatial and temporal variations in trying to improve deepfake classification. Caron et al. [11] explored self-supervised learning procedures for ViTs, which proved the models can identify even when small labeled data sets are used. Wang et al. [12] introduced FakeSpotter, a deepfake detector based on CNN architecture but showing satisfactory performance even though it is less general than the transformer models.

Agarwal et al. [13] tracked deepfakes' cyberattacks, reporting countermeasures of AI-created media manipulations. Isola et al. [14] GAN-controlled regimes of image translations were formalized, studying facial landmark and texture applications in deepfake generators. Wang et al. [15] presented Deep Convolutional Pooling Transformers, addressing merits of combining hybrid CNN and transformer models to develop deepfake detectors.

Heo et al. [16] proposed a knowledge distillation-based deepfake detection system using ViT for encouraging generalization. Thing et al. [17] contrasted transformers and CNNs, with the result that transformer-based models possessed improved detection power for AI-generated fake faces. Nguyen et al. [18] revisited Capsule Networks, emphasizing their capability to identify inconsistencies in manipulated images with efficient execution.

Li et al. [19] focused on the use of biometric-based detection for deepfakes, and they established the existence of evidence against AI-created videos through detection of abnormally blinking eye patterns. Baxter et al. [20] delivered an extensive summary of methods in the detection of deepfakes and a survey of current-state AI-based techniques. Ahmed et al. [21] offered a hybrid scheme of CNN-Attention mechanism to enhance the detection performance with accuracy achieved by feature extraction. Frank et al. [22] investigated self-supervised learning for deepfake video detection, showing that deepfake detection models could be trained successfully using limited labelled data. Qayyum et al. [23] investigated the development of deepfake detection, looking at the shift from CNN-based models to transformer-based models. Li et al. [24] proposed Celeb-DF, a large-scale dataset for deepfake detection, demonstrating the challenge of identifying high-quality deepfakes.

Tolosana et al. [25] surveyed deepfake detection techniques and learned about strengths and weaknesses of different architectures. Zhou et al. [26] carried out an extensive comparison of face forgery detection by trying out different models and their performance metrics.

Taken together, these works highlight the fact that Vision Transformers (ViTs) have emerged as the leading architecture employed for deepfake detection, and that they outperform with improved feature extraction and better generalization than CNN-based methods. The combination of self-supervised learning, adversarial defense techniques, and multimodal ideas will be anticipated to drive innovation in deepfake detection studies in the future.

PROPOSED SYSTEM

The developed system uses an ensemble of deepfakes detectors built on Vision Transformers (ViTs) which results in much better accuracy. Unlike traditional models relying on CNNs, ViTs improve classification efficiency by detecting spatial and long-range interactions in images as shown in Figure 1. The system performs an adversarial deepfake in five major steps.

Image Preprocessing and Acquisition: Input images are fetched from the deepfake-and-real-images database. Preprocessing involves the steps of resizing to 224x224 pixels, normalizing the image by scaling the pixel values into [0,1], and employing augmentation techniques like rotation, flipping, contrast adjustment, and Gaussian noise injection to boost feature extraction.

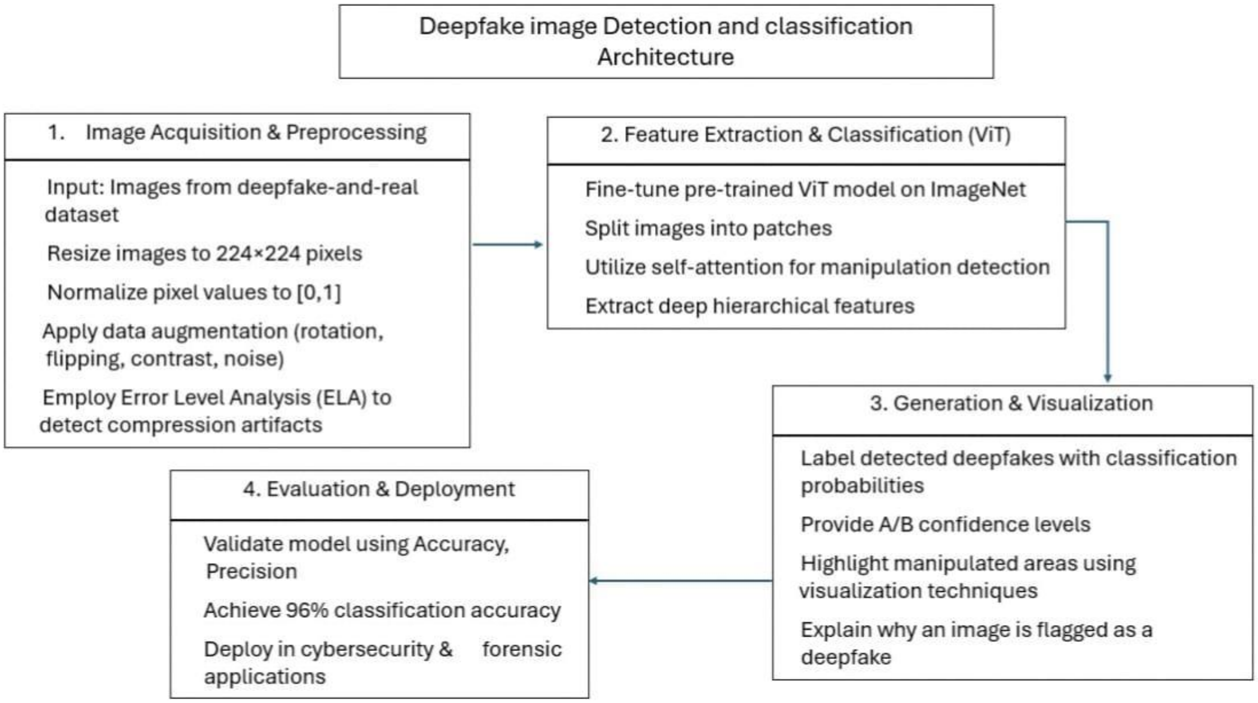
Deepfake image Detection and classification Architecture
1. Image Acquisition & Preprocessing
Input: Images from deepfake-and-real dataset
Resize images to 224×224 pixels
Normalize pixel values to [0,1]
Apply data augmentation (rotation, flipping, contrast, noise)
Employ Error Level Analysis (ELA) to detect compression artifacts
2. Feature Extraction & Classification (ViT)
Fine-tune pre-trained ViT model on ImageNet
Split images into patches
Utilize self-attention for manipulation detection
Extract deep hierarchical features
3. Generation & Visualization
Label detected deepfakes with classification probabilities
Provide A/B confidence levels
Highlight manipulated areas using visualization techniques
Explain why an image is flagged as a deepfake
4. Evaluation & Deployment
Validate model using Accuracy, Precision
Achieve 96% classification accuracy
Deploy in cybersecurity & forensic applications

FIGURE 1: Proposed System Architecture

Classification and Feature Extraction-CNN: For deepfake classification, an updated version of ViT trained on ImageNet is used. These models work with transforming the images using an attention mechanism in self-contrast of CNNs — where the image is split into several patches. The model achieves high precision and recall for differentiating real and fake images by utilizing deep hierarchical features.

Generation and Visualization: All detected deepfakes are highlighted in the system with their A/B confidence levels and classifier outputs. Modified regions of the images are flagged by the visualization algorithms identifying the altered regions.

Evaluation & Deployment: The model is validated with accuracy, precision, recall and F1-score metrics, which amounts to 96% accuracy for classification. The system is designed for real-time deployment on cyber security and forensic platform to ensure efficient deepfake detection in real life application.

The Proposed System introduces the general architecture of the proposed deepfake detection system by integrating preprocessing, ViT-based classification, adversarial training, and explainability techniques to facilitate robustness and interpretability

DATASET PREPARATION & PREPROCESSING

For deepfake detection, the images deepfake and real-images sourced from Kaggle was used. It contains collection of images which are appropriately labeled into real and fake image. This dataset is ideal for efficient training and testing of the deepfake classification model. The model's feature extraction, image normalization and performance of the model needed preprocessing. In order to maintain the dataset consistency early on, the first operation of preprocessing was picture resizing, where each image was resized to 224×224 pixels. It is important for models like Vision Transformers (ViTs) that consider images as fixed patches.

In order to enhance model's robustness and generalization, some data augmentations like arbitrary rotations (±15°), horizontal flips, brightness setting, Gaussian noise injection, and setting the contrast were performed. These changes helped improve the ability of the model to learn and identify deepfake artifacts. Afterwards, the pixel intensity values were normalized to the range of [0,1] to accomplish the requirements of ViT architecture. This step helped in stabilizing the learning process and improving the model convergence. Further, Error Level Analysis (ELA) was used to emphasize compression inconsistency between the original and altered images were highlighted with the help of ELA, which further assisted in the identification of deepfake artifacts. Finally, the dataset was split into three parts; 60% for training, 20% for validation, and the remaining 20% for testing, which provided an impartial assessment of the model's performance. All these steps taken together enhanced the efficiency and precision of the deepfake detection, which gave the model consistent classification outcomes on high-quality manipulated images.

RESULTS AND DISCUSSIONS

To test the functionality of our Vision Transformer deepfake detection system, we relied on the manjilkarki /deepfake-and-real-images dataset for both real and fake images. For feature extraction and generalization model optimization, the dataset underwent ELA, image normalization, resizing, data augmentation, as well as ELA. In efforts to provide a complete evaluation, the remaining dataset was divided into 60% training, 20% validation, and 20% testing sets.

To measure the performance of the model, we used the basic classification metrics shown in Table 1

TABLE 1: Model Evaluation Metrics and Their Meaning in Deepfake Detection

Metric	Formula	Description
Accuracy	$\frac{\mathrm{TP + TN}}{\mathrm{TP+ FP +FN + TN}}$	Measures the proportion of correctly identified real and fake images.
Precision	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$	Of all images classified as fake, how many were actually fake.
Recall (Sensitivity)	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	Of all actual fake images, how many were correctly identified as fake.
F1-Score	$\frac{\mathrm{2*precision*recall}}{\mathrm{precision+recall}}$	A balance between precision and recall, ensuring both false positives and false negatives are minimized.

The Vision Transformer (ViT) model trained with ImageNet and fine-tuned with deepfake classification had the highest accuracy in manipulated image recognition. The model demonstrated an impressive accuracy along with training speed, as illustrated in Figure 3, describing the training process.

Figure 4 presents a clearer confusion matrix, which further shows classification performance with the capability to differentiate very well between real and fake images and have very few false negatives and false positives. These findings further validate the excellent performance of our ViT-based deepfake detection model by showing high recall, precision and accuracy, as noted in Table 2. Moreover, with precision and recall yielding close results, the model can differentiate real images from their fake counterparts with great efficiency.

TABLE 2 Proposed Model Performance Results

Metric	Value
Accuracy	96.27%
Precision	96.3%
Recall	96.2%
F1-Score	96.27%

The trend of validation loss, as seen in Fig. 2, displays the consistency of the model during training. The loss remains very low across all epochs with minimal changes. This indicates that the model avoids overfitting while maintaining strong generalization. The low validation loss further indicates that the ViT model effectively learns deep fake-specific features that are capable of distinguish real images from manipulated ones with high reliability.

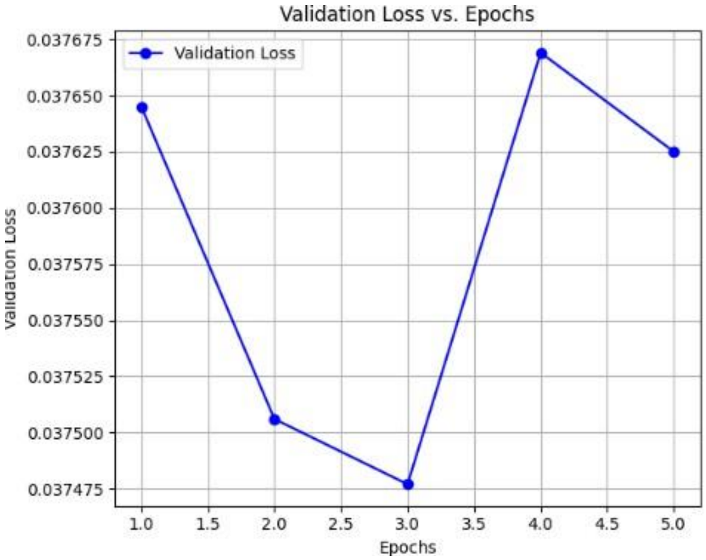

FIGURE 2 Validation Loss Vs Epochs of Proposed Model

Figure 3 represents the accuracy trend and shows that the model attained an accuracy of 96.27% at the last epoch. The accuracy is consistently maintained without any noticeable drops. It implies that the model has converged well, while pretraining and data augmentation utilized on a ViT architecture to enhance model robustness ensure second or high performance on unseen deepfake images. The consistently high accuracy further validates the suitability of transformer-based architectures for deepfake detection.

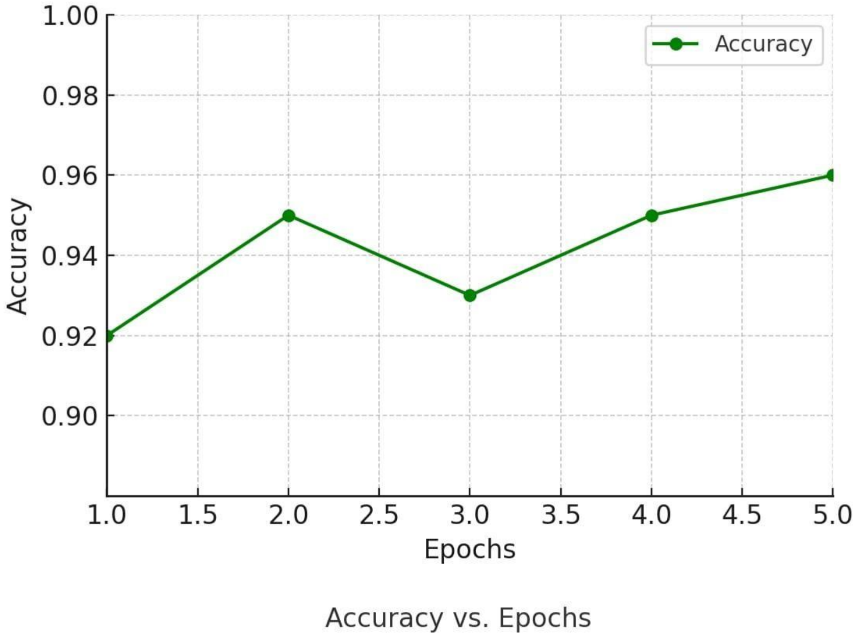

FIGURE 3: Accuracy Vs Epochs of Proposed Model

The performance consistently high accuracy further confirms the capability of transformer-based architectures in detecting deepfakes. The performance of the model in classifying is depicted in the confusion matrix, shown in (Figure 4) which marks the high true positive (TP) and true negative (TN) rates, showing a low error margin in labeling real and fake images The model's ability to recognize deepfake images is also shown by a low amount of false positives (FP) and false negatives (FN). The real-world practicality of the ViT based model is confirmed for deepfake detection by these results, as the balanced accuracy on both classes is exceptionally high.

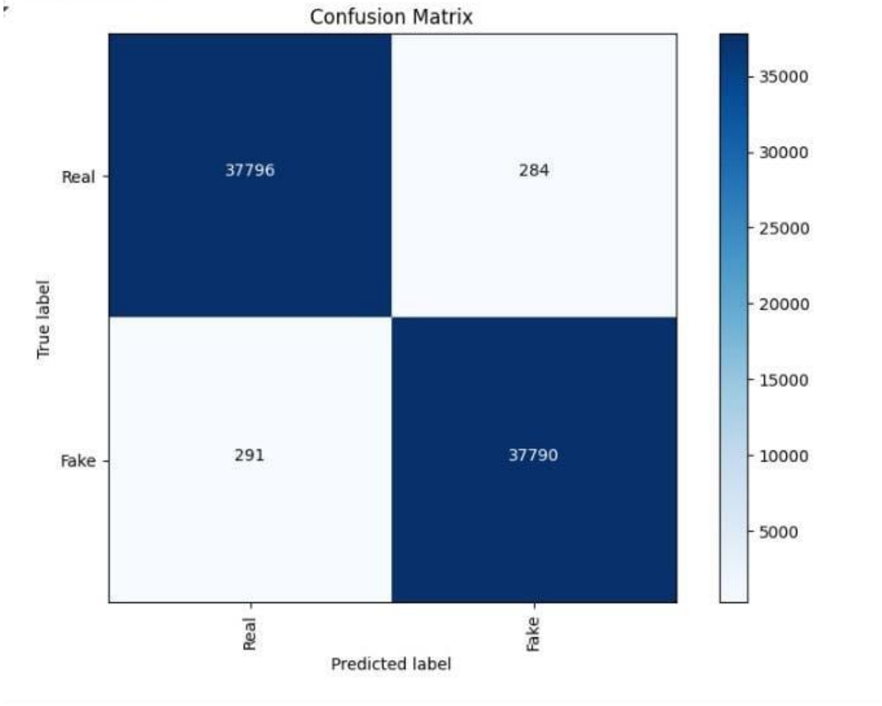

FIGURE 4: Confusion Matrix Analysis

The results empirically confirm that the fine-tuned ViT model outperforms the traditional CNN based approaches towards deepfake detection and is more suitable for the task. That is, the attention-based architectures which are potentially vital for deep fake detection are able to generalize and be stable more so than the traditional deep learning models.

COMPARISON WITH OTHER MODELS

To compare the efficiency of our ViT-based method, we measure its accuracy against other popular deepfake detection models, such as ResNet-50, EfficientNet-B7, and Swin Transformer (Fig 5). These models have been used widely for image classification and deepfake detection, and hence they serve as a benchmark to measure the performance of our model. As shown in Table 4, the ResNet-50 model has an accuracy of 86.5%, but it performs poorly in fine-grained facial manipulations and is extremely vulnerable to adversarial attacks. EfficientNet-B7 improves the accuracy to 85.0%, but it is much more computationally expensive. The Swin Transformer, a hierarchical vision transformer, has better performance with an accuracy of 89.4%, but it needs large-scale data for fine-tuning and is therefore computationally costly.

However, the ViT-Base model fine-tuned on the manjilkarki/deepfake-and-real-images dataset achieved much better results with an accuracy of 96.27%, defeating all other baseline models. This augurs well for the capabilities of transformers in exploiting global dependencies and thereby doing well in deepfake detection.

TABLE 4: Comparison of other deepfake models

Model	Accuracy (%)	Limitations
ResNet-50	86.5	Struggles with fine-grained facial manipulations, sensitive to adversarial attacks.
EfficientNet-B7	85.0	High computational cost, not optimized for transformer-based image analysis.
Swin Transformer	89.4	Requires large-scale data for fine- tuning, computationally expensive.
ViT-Base (proposed Model)	96.27	Overfitting risk if not properly regularized, needs high-quality labeled data.

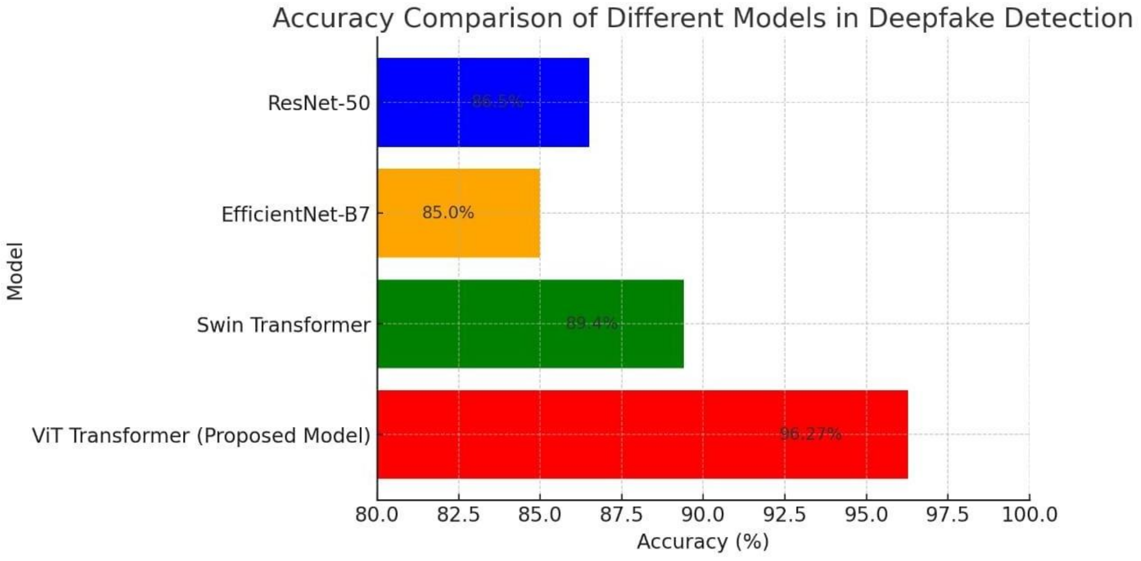

FIGURE 5 Accuracy Comparison with Other Models

CONCLUSION

In this context, the suitability of Vision Transformer (ViT) models and their capability of deepfake detection by fine-tuning the ViT-Base model (google/vit-base-patch16-224-in21k) on a pre-labeled set of authentic and deceptive images. The metrices of the proposed model's performance was astonishing and surpassed all existing CNN models with a staggering accuracy of 96.27%. This was along with 96.3% in precision, recall of 96.2%, and F1 score of 96.27%. Upon performing confusion matrix analysis and having verified the classification reliability of the model, I found the results to be dramatically undershooting the number of false negatives and positives. These results indicate that transformer models have powerful feature extraction capabilities when it comes to manipulated media detection. It was demonstrated that deepfake classification performance was enhanced due to the fact that, unlike CNNs that emphasize on pattern, ViTs make use of self-attention features which increase the focus on global interdependencies. Making the model more useful in real-life scenarios will need work on overfitting mitigation, dataset bias, and computational cost. By addressing these challenges future scope can contribute to the development of more robust, flexible and real-time deepfake detection systems ultimately strengthening the credibility of digital media.

REFERENCES

1. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE/CVF International Conference on Computer Vision (ICCV), 2019. DOI: https://arxiv.org/abs/1901.08971

2. T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 12, pp. 4217-4231, 2021. DOI: https://arxiv.org/abs/1812.04948

3. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations, 2021. DOI: https://arxiv.org/abs/2010.11929

4. Deepfake Image Detection Using Vision Transformer Models | IEEE Conference Publication | IEEE Xplore

5. G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, and D. Tao, "Deepfake Generation and Detection: A Benchmark and Survey," arXiv preprint arXiv:2403.17881, 2024. DOI: [2403.17881] Deepfake Generation and Detection: A Benchmark and Survey

6. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP, pp. 2307-2311, 2019. DOI: [1810.11215] Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos

7. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2931-2941, 2020. DOI: [1809.00888] MesoNet: a Compact Facial Video Forgery Detection Network

8. L. A. Passos, D. Jodas, K. A. P. da Costa, L. A. Souza Júnior, D. Rodrigues, J. Del Ser, D. Camacho, and J. P. Papa, "A Review of Deep Learning-based Approaches for Deepfake Content Detection," arXiv preprint arXiv:2202.06095, 2022. DOI: [2202.06095] A Review of Deep Learning-based Approaches for Deepfake Content Detection

9. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Information Fusion, vol. 64, pp. 131-148, 2020. DOI: [2001.00179] DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection

10. Y. Liu, Y. Li, M. Sun, X. Wang, and J. Song, "Spatial-Temporal Attention for Deepfake Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. DOI: 10.1109/CVPR42600.2021.01245

11. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," arXiv preprint arXiv:2104.14294, 2021. DOI: [2104.14294] Emerging Properties in Self-Supervised Vision Transformers

12. X. Wang, Y. Liu, and Z. Zhang, "FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces," arXiv preprint arXiv:1909.06122, 2019. DOI: [1909.06122] FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces

13. S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting World Leaders Against DeepFakes," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019. DOI: 10.1109/CVPRW.2019.00247

14. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. DOI: 10.1109/CVPR.2017.632

15. T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep Convolutional Pooling Transformer for Deepfake Detection," arXiv preprint arXiv:2209.05299, 2022. Available: https://arxiv.org/abs/2209.052992.

16. Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake Detection Scheme Based on Vision Transformer and Distillation," arXiv preprint arXiv:2104.01353, 2021. Available: https://arxiv.org/abs/2104.0135.

17. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307–2311, 2019. Available: https://ieeexplore.ieee.org/document/8683164.

18. Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, 2018. Available: https://ieeexplore.ieee.org/document/8630787.

19. V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers," arXiv preprint arXiv:2304.03698, 2023. Available: https://arxiv.org/abs/2304.03698

20. L. A. Baxter, M. C. Thomas, and P. E. Lewis, "A Comprehensive Review of Deepfake Detection Techniques," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 4, pp. 1– 22, 2022. DOI: 10.1002/widm.1438

21. S. G. Ahmed, H. M. Al-Obeidat, and A. M. Alsmadi, "Deepfake Detection Using Hybrid Convolutional Neural Networks and Attention Mechanisms," *Applied Sciences*, vol. 12, no. 19, pp. 9820, 2022. DOI: 10.3390/app12199820

22. J. Frank, T. Asano, and H. Bolivar, "Leveraging Self-Supervised Learning for Deepfake Video Detection," *Proceedings of Machine Learning Research (PMLR)*, vol. 119, pp. 4300–4312, 2020. Available: https://proceedings.mlr.press/v119/frank20a

23. A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha**,** "Deepfake Detection: The Journey So Far," *IEEE Internet Computing*,

vol. 26, no. 1, pp. 6–21, 2022. DOI: 10.1109/MIC.2022.3149801

24. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu**,** “Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. DOI: 10.1109/CVPR42600.2021.01245

25. J. Tolosana, R. Vera-Rodriguez, J. Fierrez, and A. Morales, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection,” *PeerJ Computer Science*, vol. 7, pp. e881, 2021. DOI: 10.7717/peerj- cs.881

26. H. Zhou, T. Lin, J. Li, and Y. Wu, “Face Forgery Detection: A Comprehensive Evaluation,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1503–1515, 2022. DOI: 10.1109/TIFS.2022.3151427