International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Deepfake Video Detection using CNN based Architectures and Vision Transformer Model

Sumedha Arya

aryasumedha06@gmail.com

Abstract

Deepfake videos are becoming more advanced and realistic, that their detection is getting more challenging. It brings a deep concern for privacy and security on digital platforms. Current techniques perform well in closed environment but faces issues in real world scenario for deepfake detection. Therefore, in this work, we focus on detecting deepfake videos by first breaking each video into individual image frames and then analyzing those frames using deep learning models. We applied transfer learning using various CNN based pretrained models such as VGG, ResNet, DenseNet, MobileNet, EfficientNet along with Vision Transformer (ViT). Our results show that all models perform well, with accuracy values between 92.56% and 97.16% for CNN-based models. The Vision Transformer performs the best overall, achieving 99.00% accuracy by capturing global patterns and small manipulation details in deepfake videos. Overall, the study proves that transfer learning, is a strong and reliable approach for detecting deepfake videos even when the size of the dataset is less.

Keywords: Deepfake Detection, CNN, Vision Transformer, Transfer Learning, Video Frame Analysis

1. Introduction

Nowadays, lots of deepfake content is circulating over internet. Such content looks so real that its detection, classification and removal is difficult. This is because, deepfake technology uses deep learning algorithms which creates highly realistic fake videos. Not only it threatens the reliability of online information and security systems, but also it can easily perform manipulations in faces and voices [1, 2]. Deepfakes such as voice cloning, text-to-speech, voice conversion, sound editing, language translation, and segmented images can be created easily using low-cost technologies and mobile apps [3, 4]. Some of them are FaceApp [6], advanced frameworks like dynamic identity recognition networks [7] and vocoder-based voice synthesis systems [8]. As a result, deepfakes have now become a mode of spreading fake news and misinformation by different groups, including automated accounts, political actors, and online manipulators [9, 10].

Although there are many deepfake detection techniques, but still, they struggle to identify such content in real-time. The problem they face is about lack of adaptability, and protection against new types of attacks [5]. Therefore, now it has become a necessity for detecting deepfakes to prevent the misuse of artificial intelligence in data manipulation [11]. Some detection approaches, based on graph neural networks, have

been explored [12], but still challenges prevail. In various studies, it has been found that, detection systems perform well in controlled environments, but they struggle in real-world situations [13].

Deepfake generation techniques are evolving very quickly making it difficult to improve detection tools [15]. Such as early detection models could only identify basic fake content [14], but modern deepfakes now look so real that it requires more advanced detection techniques [16]. This is because deepfake tools can precisely manipulate facial expressions, speech, and identity features [17].

Deepfake detection methods are generally grouped into three types. These are feature-based, deep learning-based and hybrid. Feature-based methods focus on manually designed features [18] while deep learning-based methods focus on learning complex patterns from large datasets [19] and hybrid methods that combine both approaches for better performance [20].

In real life, videos shared on social media are often compressed with low resolution, created from multiple low-quality sources, which makes detection more difficult [21, 22]. Recent research emphasizes the need to develop more robust detection techniques that can perform well on different types of deepfakes and poor-quality videos [23].

This research aims to perform deepfake video detection by implementing and comparing a variety of deep learning models. This study evaluates how deep learning pretrained models, both CNN and Transformer based can enhance the ability of deepfake detection systems. The paper is divided into following sections; literature review, research methodology, results analysis, conclusion and references.

2. Literature Review

Nowadys, many people are active on social media platforms. Lots of content in form of images, and videos continue to circulate on them. Such data, can be utilized for training advanced deep learning architectures to create deepfake. This can lead to circulation of fake images and videos quickly among large audiences within minutes. It creates fear, confusion, and mistrust among the public [24]. AI-based tools such as DeepFaceLab [25] and Face Swap [26] make it easy to change facial features, expressions, age, and appearance, even for non-experts. In this literature review, we focussed on recent research for deepfake video detection by covering modern detection techniques, their performance comparison, and key challenges.

Authors [27] in their research, reviewed many deepfake detection techniques. They identified the current research trends, and suggested future directions in this field. Their study mainly focused on video and image deepfakes, with less work on audio. They examined deep learning models such as CNNs, DNNs, LSTM, and GANs. Their analysis showed that deep learning-based image methods perform better than traditional feature-based techniques for detection [27].

A new method that uses Particle Swarm Optimization (PSO) to improve deepfake detection was proposed by the authors [28]. They made a hybrid model with EfficientNet along with GRU for video classification and applied PSO to automatically tune its parameters. Their experiments reveal that the proposed approach applied on large datasets (DFDC) outperformed earlier methods in detecting fake videos [28].

UFCC, a Unified Forensic Scheme by Content Consistency, was proposed by the authors [29]. It detects deepfakes by checking inconsistencies between edited and unedited regions in images and videos. The model compares facial similarities across video frames using neural networks and Siamese networks. This method achieved better performance than existing techniques and worked well for both image and video forgeries [29].

Authors [30] proposed a deepfake detection approach based on 3D motion analysis. Their method tracks facial expressions movement in both 2D and 3D frames and separates them from head movement. This improves detection performance, especially under difficult conditions such as low lighting, large head movements, and video compression. Their experiments confirmed that the method works well in real-world scenarios [30].

A new model called ReLAF-Net was proposed by the authors [31] to improve deepfake detection using a special attention mechanism. This model focuses on both local and global features present in the images by analyzing high-frequency patterns to detect complex forgery elements. When tested on the FaceForensics++ dataset, the proposed model achieved an accuracy of 97.92%, outperforming many existing techniques [31]. Although ReLAF-Net performs very well in most cases, it may still face difficulties when detecting extremely high-quality deepfakes created using advanced technologies [31].

Another method proposed by authors [32] combines Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) with deep learning. This approach extracts both spatial and temporal features from videos and trains a deep learning classifier to distinguish between real and fake videos. The method achieved an accuracy of 98.91% and showed strong performance even with smaller datasets, making it reliable for practical use [32].

Authors [33] in their research highlighted the limitations of CNN-based models that only analyze individual frames and ignore motion information between frames. Therefore, to overcome this issue, they proposed a model using optical flow to capture temporal changes across video frames. Their proposed model in combination with CNNs, RNNs, and optical flow, results in better understanding of video dynamics for improved deepfake detection [33].

A systematic review of deepfake detection methods was done by the authors [34] highlighting the strengths and weaknesses of existing deep learning techniques. They found that many studies used CNNs and RNNs for deepfake detection focusing mainly on accuracy. However, they ignored the importance of security, robustness, and real-world deployment. Authors also pointed the issue of lack of access to multilingual datasets and insufficient technical details in some studies. Their findings show path for building more practical and reliable detection systems [34].

Deepfake technology, uses GANs and VAEs deep learning and fake content generation algorithms to create highly realistic fake images and videos [35]. Many detection systems consider deepfake identification as a binary classification problem using CNN-based pretrained models such as Xception showing good results. However, their performance often decreases due to video compression and noise. To overcome this issue, a new model called MSIDSnet was proposed by the authors [36] which uses dual-stream learning, multi-scale feature extraction, and Vision Transformers (ViT). It combines spatial and frequency-based information to improve detection accuracy and efficiency [36].

In short, recent research has improved deepfake detection using diverse techniques. However, there is still challenges exists with the advent of new technologies for deepfake video generation.

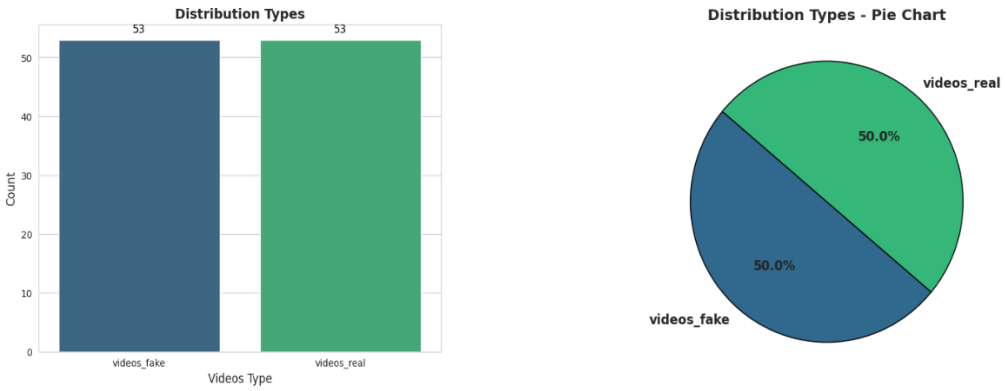
3. Research Methodology

This section describes a systematic methodology used for deepfake detection using deep learning techniques. It includes data acquisition, exploratory data analysis, frame extraction, data preprocessing, modeling, and performance evaluation. All experiments were conducted using python programming in a Kaggle environment with GPUs.

3.1.Dataset Acquisition and Exploratory Analysis

The dataset is sourced from Kaggle. It is a Small-scale Deepfake Forgery Video Dataset (SDFVD) consisting of 106 video samples, with 53 deepfake and 53 real videos. This maintained the equal class distribution among them.

Figure 1: Count plot and Pie plot for Video Type

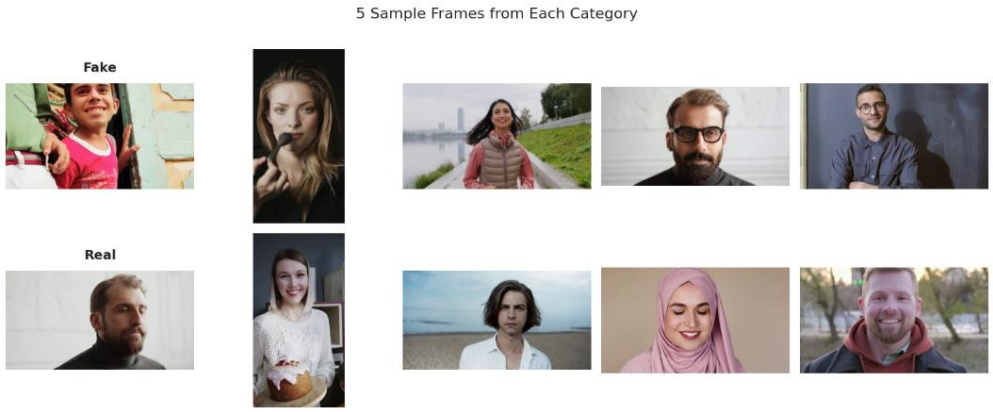


3.2.Frame Extraction and Dataset Augmentation

Since convolutional neural networks (CNNs) works on image data, therefore the video samples were converted into image frames. To construct a sufficiently large and balanced dataset, a target of approximately 4000 frames per class was defined. The number of frames extracted from each video was computed based on the total number of videos per category. Frames were sampled uniformly across each video by selecting frames at fixed intervals to avoid redundancy and ensure temporal diversity. Extracted frames were stored in JPEG format.

As a result, a total of 7,974 frames were generated, consisting of 3,987 fake and 3,987 real frames. This process transformed the original video classification problem into an image-based classification task.

Figure 2: Fake vs Real Sample Images



3.3.Data Preprocessing and Dataset Splitting

Class labels were numerically encoded to Real as 0 and Fake as 1. This is done to ensure compatibility with deep learning models training. The dataset was then divided into training, validation, and testing sets using stratified sampling to preserve class balance. The split ratio was 70% for training, 15% for validation, and 15% for testing.

Image preprocessing and augmentation were applied. For training data, transformations included resizing, random cropping, horizontal and vertical flipping, rotation, tensor conversion, and normalization using ImageNet mean and standard deviation values. Validation and test datasets were processed using resizing, center cropping, tensor conversion, and normalization only.

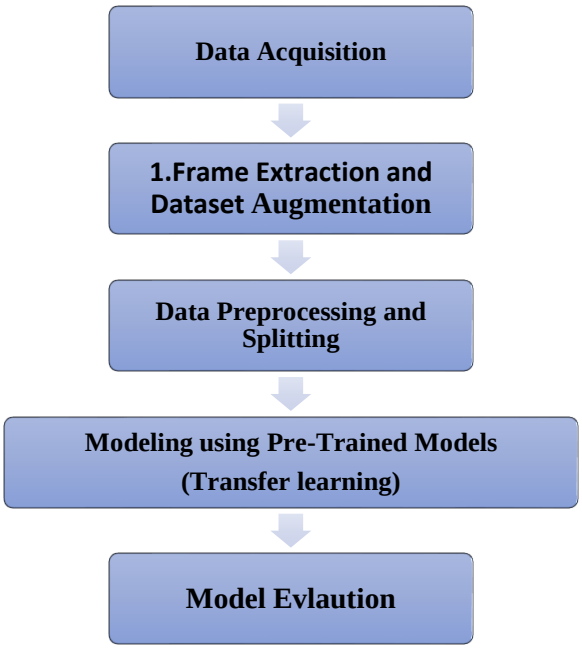
A custom dataset class was implemented to load images, apply transformations, and return image–label pairs. DataLoader objects were created with a batch size of 32, training data was shuffled, and multiple worker threads were applied for efficient loading.

3.4.Model Selection and Architecture

Seven pre-trained convolutional neural network architectures were selected for transfer learning:

- VGG16: A deep CNN that used 3x3 sized filters for hierarchal feature extraction.
- VGG19: An extended version of VGG16 with increased depth of 19 layers, enabling improved feature extraction.
- ResNet18: A residual CNN that uses skip connections to overcome the problem of vanishing gradient descent.
- DenseNet121: A densely CNN where each layer receives inputs from all preceding layers, promoting feature reuse and reducing parameter redundancy.
- MobileNetV2: A lightweight CNN based architecture designed for efficient computation using inverted residual bottleneck and depthwise separable convolutions.
- EfficientNetB0: A compound-scaled CNN that balances depth, width, and resolution to achieve high accuracy with fewer parameters.
- Vision Transformer (ViT): A transformer-based vision model that divides images into fixed-size patches and processes them using attention mechanism to capture long range relationships in images.

Figure 3: Model Architecture



3.5.Training Procedure

Model training was performed using GPUs. The cross-entropy loss function was used for binary classification, with Adam optimizer having a learning rate of 0.0001. Each model was trained for five epochs.

Training and validation losses and accuracies were recorded at each epoch. Validation was performed without gradient computation to reduce memory usage.

3.6.Evaluation and Performance Analysis

After training, model performance was evaluated on the independent test dataset. Predictions were obtained in evaluation mode, and standard classification metrics, including accuracy, precision, recall, and F1-score, were computed. Both macro and weighted averages were reported to ensure balanced evaluation. Confusion matrices were generated to analyze class-wise prediction performance and visualized using heatmaps. Additionally, training and validation loss curves, as well as accuracy curves, were plotted to assess convergence behavior and detect overfitting.

4. Results And Discussion

This section presents the experimental results obtained from the evaluation of six pre-trained CNN models and one transformer-based vision model. The performance was assessed using standard classification metrics. All evaluations were conducted on an independent test set comprising 1,197 frames (598 Real and 599 Fake).

Table 1 summarizes the performance of the evaluated models. The results indicate that all models achieved high classification accuracy, ranging from 92.56% to 99%, demonstrating the effectiveness of transfer learning for deepfake detection.

Table 1: Models Performance Evaluation

Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score
VGG16	0.9591	0.9592	0.9591	0.9591
VGG19	0.9482	0.9504	0.9482	0.9481
ResNet18	0.9256	0.9332	0.9257	0.9253
DenseNet121	0.9666	0.9684	0.9666	0.9666
MobileNetV2	0.9474	0.9504	0.9474	0.9473
EfficientNetB0	0.9716	0.9725	0.9716	0.9716
Vision Transformer (ViT)	0.9900	0.9900	0.9900	0.9900

Among the evaluated models, the Vision Transformer (ViT) achieved the highest performance, attaining an accuracy of 99.00% and a macro-averaged F1-score of 0.9900, thereby outperforming all CNN based architectures. Among CNN models, EfficientNetB0 emerged as the best-performing architecture, achieving an accuracy of 97.16% and a macro-averaged F1-score of 0.9716, followed closely by DenseNet121 with an accuracy of 96.66%. The VGG-based models exhibited competitive but comparatively lower performance, with VGG16 outperforming VGG19 (95.91% vs. 94.82%). MobileNetV2 achieved an accuracy of 94.74%. In contrast, ResNet18 recorded the lowest accuracy of 92.56%.

Training and validation loss and accuracy curves were analyzed to assess convergence and generalization behavior. The Vision Transformer, EfficientNetB0, and DenseNet121 demonstrated rapid convergence, with validation losses stabilizing within the first three epochs, indicating efficient learning of discriminative features. ResNet18 exhibited slower convergence and higher validation loss, suggesting underfitting. Conversely, VGG19 displayed signs of overfitting by showing a wide gap between training and validation losses. MobileNetV2 showed stable and closely aligned training and validation curves.

Figure 4: VGG16 Loss and Accuracy Plots

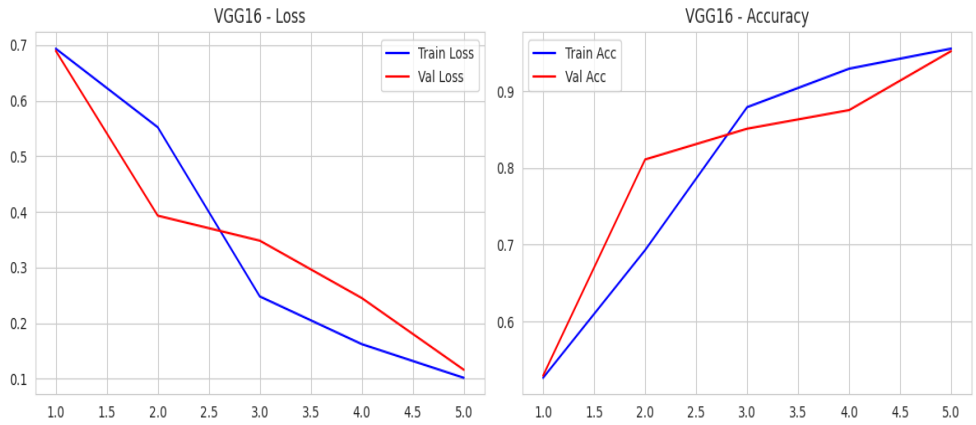


Figure 5: VGG19 Loss and Accuracy Plots

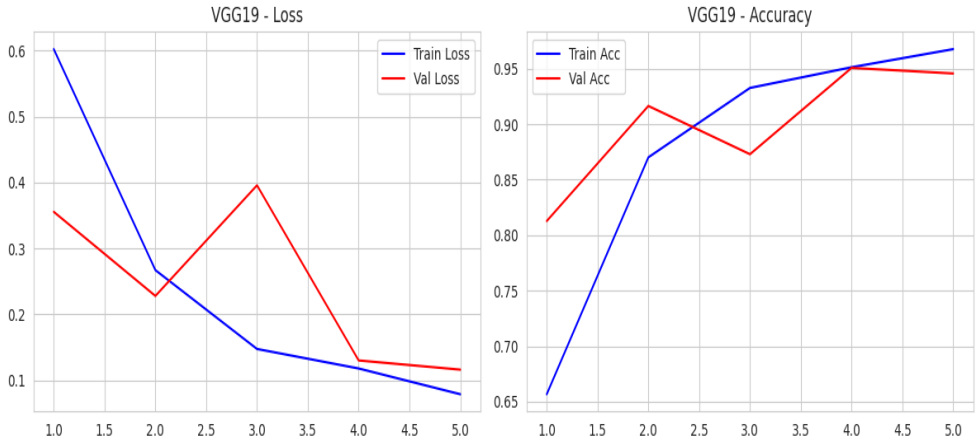


Figure 6: Resnet-18 Loss and Accuracy Plots

Figure 7: DenseNet121 Loss and Accuracy Plots

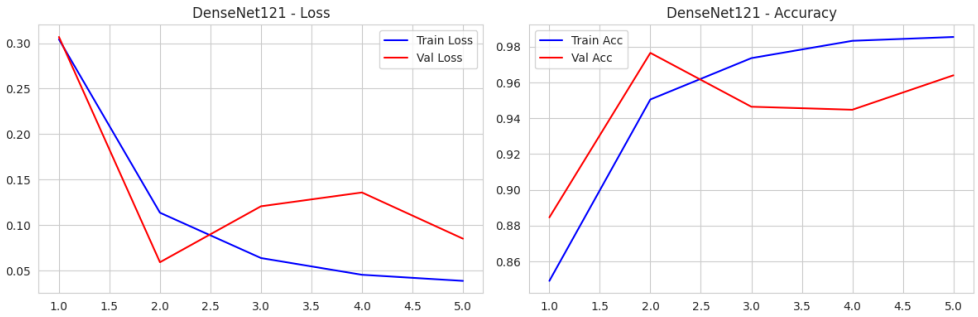


Figure 8: MobileNetv2 Loss and Accuracy Plots

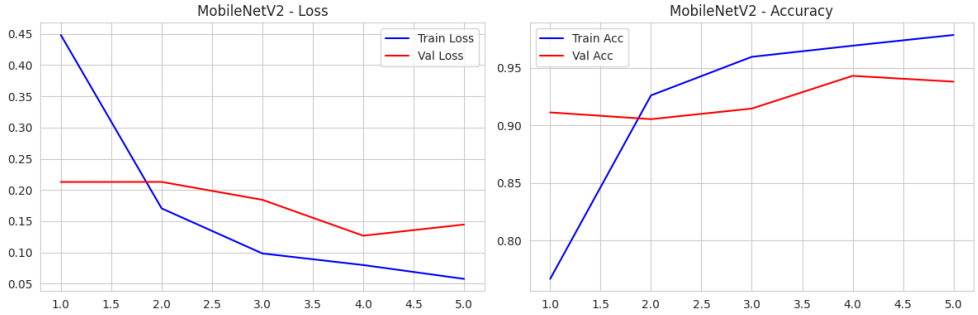


Figure 9: EfficinetNetB0 Loss and Accuracy Plots

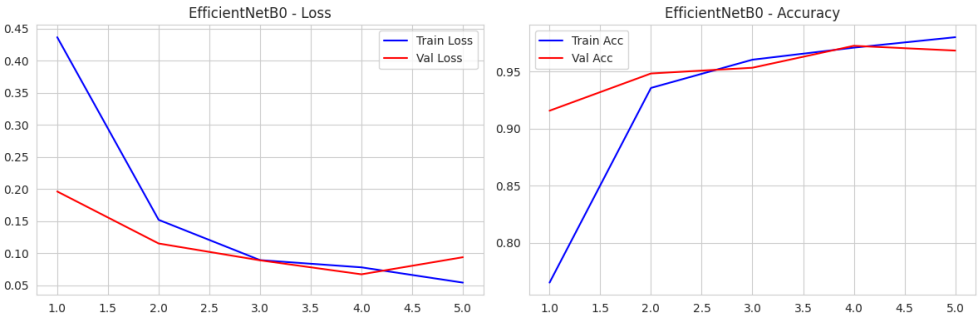
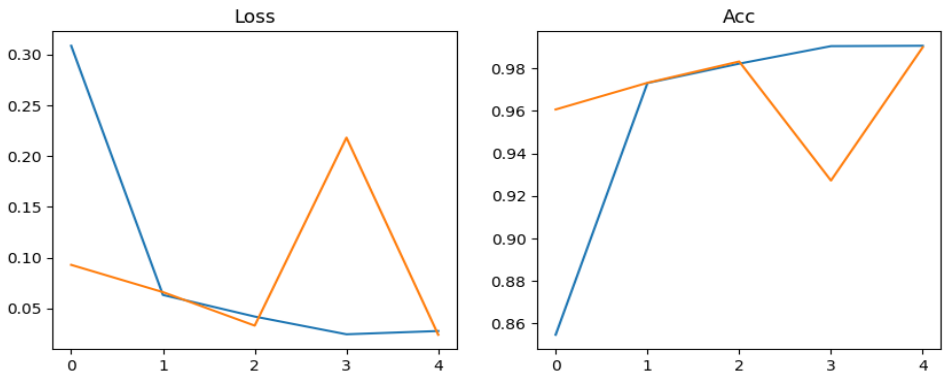


Figure 10: ViT Loss and Accuracy Plots



Confusion matrix analysis provided detailed insights into classification behavior. High-performing models, particularly the Vision Transformer, EfficientNetB0, and DenseNet121, exhibited strong diagonal

dominance, indicating a high proportion of correct predictions, with ViT showing fewer misclassifications across the test set. Off-diagonal elements were minimal for these models, while ResNet18 showed a higher incidence of false negatives for fake frames, suggesting difficulty in identifying subtle manipulations such as facial blending artifacts. Overall, the balanced distribution of true positives and true negatives across models validates the effectiveness of data augmentation strategies and stratified dataset splitting.

Figure 11: Confusion Matrix of VGG16

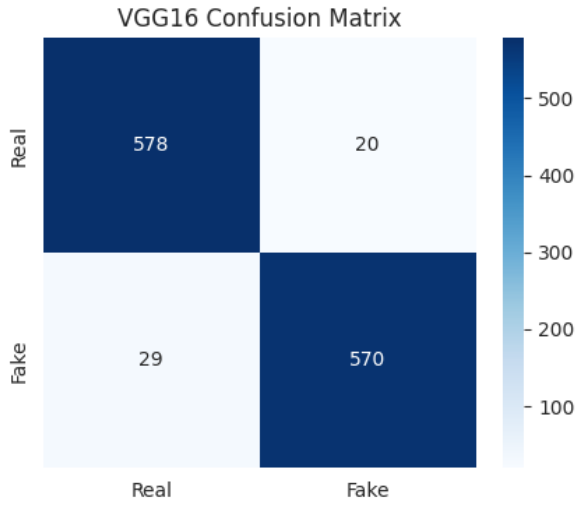


Figure 12: Confusion Matrix of VGG19

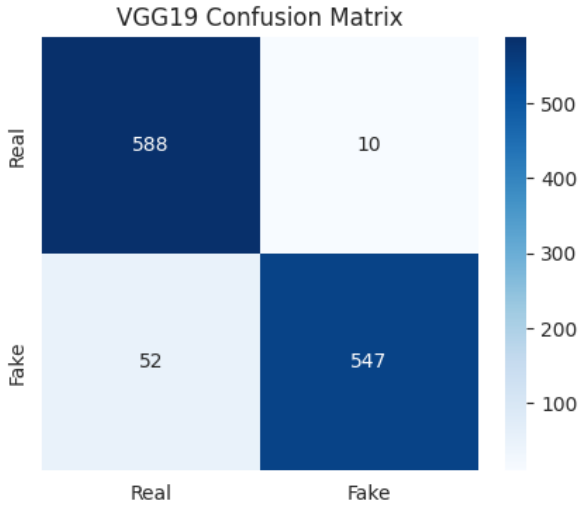


Figure 13: Confusion Matrix of ResNet18

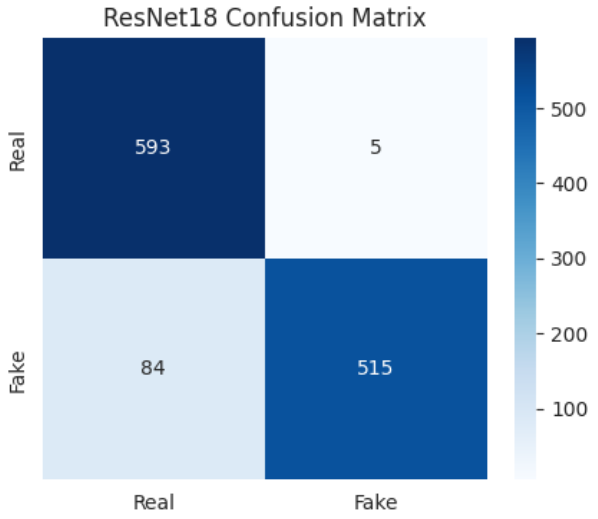


Figure 14: Confusion Matrix of DenseNet121

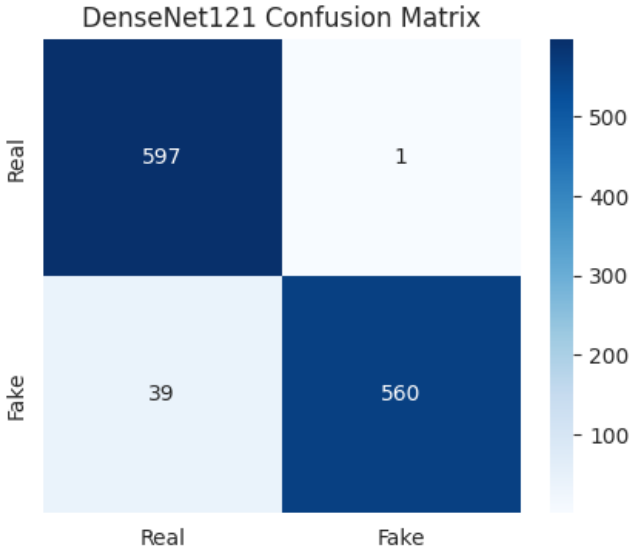


Figure 15: Confusion Matrix of MobileNetV2

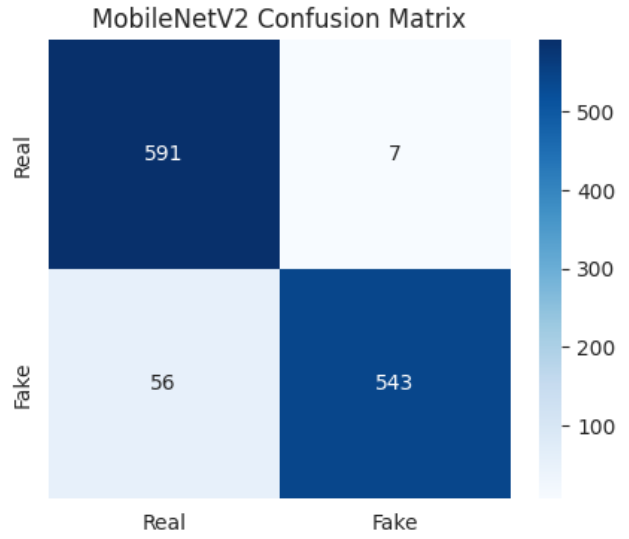


Figure 16: Confusion Matrix of EfficinetNetB0

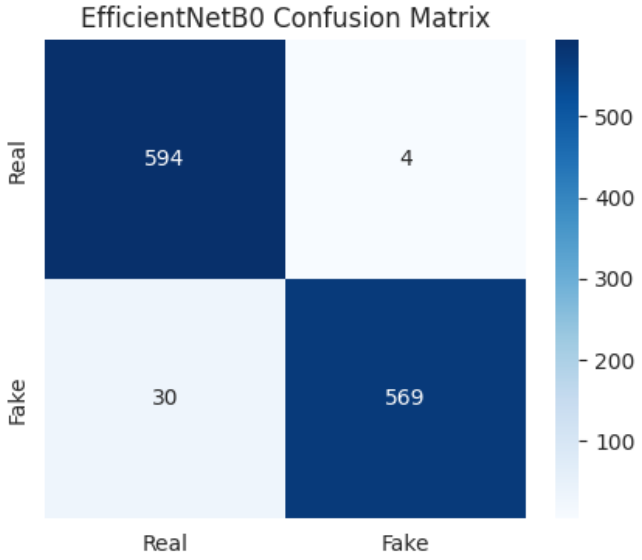
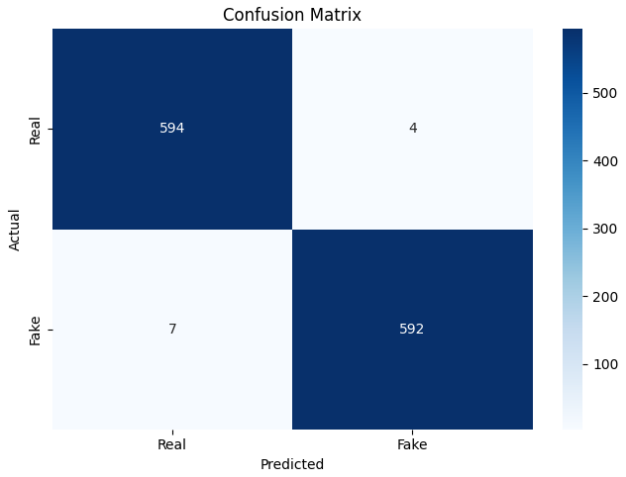


Figure 17: Confusion Matrix of Vision Transformer (ViT)



5. Conclusion

This research successfully demonstrated a deepfake detection system using both CNNs and a ViT model. Instead of analyzing videos directly, they were converted into image frames, making detection task simple and efficient for deep learning models.

The results show that all CNN-based models achieved strong accuracy, ranging from 92.56% to 97.16%. The Vision Transformer (ViT) achieved the highest performance with 99.00% accuracy and perfectly balanced metrics. This demonstrates the advantage of transformer-based models in capturing global image relationships through self-attention. Among CNN models, EfficientNetB0 performed the best, followed closely by DenseNet121, while older architectures such as VGG16 and ResNet18 showed comparatively lower performance. These findings confirm that transfer learning is highly effective for deepfake detection, even when trained on a relatively small dataset.

In conclusion, this work provides a strong foundation for deepfake detection using both CNNs and transformer-based models. As deepfake generation techniques continue to advance, developing detection systems that are accurate, efficient, explainable, and robust across domains are essential.

References

1. Gong L.Y., Li X.J., “A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges”, Electronics, 2024, 13, 585.
2. Ahmed S.R., Sonuç E., Ahmed M.R., Duru A.D., “Analysis Survey on Deepfake Detection and Recognition with Convolutional Neural Networks”, Proceedings of the International Congress on Human–Computer Interaction, Optimization and Robotic Applications (HORA), 2022, pp. 1–7.
3. Zobaed S., Rabby F., Hossain I., Hossain E., Hasan S., Karim A., Hasib K.M., “Deepfakes: Detecting Forged and Synthetic Media Content Using Machine Learning”, Artificial Intelligence and Cyber Security: Impact, Implications, Security Challenges, Technical and Ethical Issues, Forensic Investigation Challenges, 2021, pp. 177–201.
4. Suratkar S., Bhiungade S., Pitale J., Soni K., Badgujar T., Kazi F., “Deep-Fake Video Detection Approaches Using Convolutional–Recurrent Neural Networks”, Journal of Control and Decision, 2023, 10, pp. 198–214.

5. Sontakke N., Utekar S., Rastogi S., Sonawane S., “Comparative Analysis of Deep-Fake Algorithms”, arXiv Preprint, 2023, arXiv:2309.03295.
6. FaceApp Technology Limited, “FaceApp”, Available online: https://www.faceapp.com/ (Accessed on 16 December 2023).
7. Xu Z., Hong Z., Ding C., Zhu Z., Han J., Liu J., Ding E., “MobileFaceSwap: A Lightweight Framework for Video Face Swapping”, Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36, pp. 2973–2981.
8. Jia Y., Zhang Y., Weiss R., Wang Q., Shen J., Ren F., Nguyen P., Pang R., Moreno I.L., Wu Y., “Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis”, Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2018, 31, pp. 1–11.
9. Khatri N., Borar V., Garg R., “A Comparative Study: Deepfake Detection Using Deep Learning”, Proceedings of the 13th International Conference on Cloud Computing, Data Science and Engineering, 2023, pp. 1–5.
10. Integrity PT, “Voice Deepfake: Real Life Examples”, Available online: https://www.integrity.pt/real-life/voice-deepfake.html.
11. BR S.R., Pareek P.K., Bharathi S., Geetha G., “Deepfake Video Detection System Using Deep Neural Networks”, Proceedings of the IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), 2023, pp. 1–6.
12. She H., Hu Y., Liu B., Li J., Li C.T., “Using Graph Neural Networks to Improve Generalization Capability of Models for Deepfake Detection”, IEEE Transactions on Information Forensics and Security, 2024.
13. Khan S.A., Artusi A., Dai H., “Adversarial Robust Deep Media Detection Using Fused Convolutional Neural Network Prediction”, arXiv Preprint, 2021, arXiv:2102.05950.
14. Silva V.E.S., Lacerda T.B., Miranda P., Câmara A., Chagas A.R.C., Furtado A.P.C., “Federated Learning and Mel-Spectrograms for Physical Violence Detection in Audio”, Brazilian Conference on Intelligent Systems, Springer Nature, 2023, pp. 379–393.
15. Cheng H., Guo Y., Wang T., Nie L., Kankanhalli M., “Towards Generalized Deepfake Detection via Primary Region Regularization”, arXiv Preprint, 2023, arXiv:2307.12534.
16. Bhat M., Agrawal P., Gupta C., “DFDA: Analysis of Deep Learning Models for Detecting Deepfake Videos”, 2024.
17. Nadimpalli A.V., Rattani A., “ProActive Deepfake Detection Using GAN-Based Visible Watermarking”, ACM Transactions on Multimedia Computing, Communications, and Applications, 2023.
18. Kaddar B., Fezza S.A., Hamidouche W., Akhtar Z., Hadid A., “On the Effectiveness of Handcrafted Features for Deepfake Video Detection”, Journal of Electronic Imaging, 2023, 32(5), 053033.
19. Shah Y., Shah P., Patel M., Khamkar C., Kanani P., “Deep Learning Model-Based Multimedia Forgery Detection”, Proceedings of the Fourth International Conference on I-SMAC, IEEE, 2020, pp. 564–572.
20. Rajesh N., Prajwala M.S., Kumari N., Rayyan M., Ramachandra A.C., “Hybrid Model for Deepfake Detection”, Proceedings of the 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication (MARC), Springer Nature, 2022, pp. 639–649.

21. Zi B., Chang M., Chen J., Ma X., Jiang Y.G., “WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection”, Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2382–2390.
22. Rahman A., Siddique N., Moon M.J., Tasnim T., Islam M., Shahiduzzaman M., Ahmed S., “Short and Low-Resolution Deepfake Video Detection Using CNN”, IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2022, pp. 259–264.
23. Narayan K., Agarwal H., Thakral K., Mittal S., Vatsa M., Singh R., “DF-Platter: Multi-Face Heterogeneous Deepfake Dataset”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9739–9748.
24. Mukta M.S.H., Ahmad J., Raiaan M.A.K., Islam S., Azam S., Ali M.E., Jonkman M., “An Investigation of the Effectiveness of Deepfake Models and Tools”, Journal of Sensor and Actuator Networks, 2023, 12(4), 61.
25. Perov I., Gao D., Chervoniy N., Liu K., Marangonda S., Ume C., Jiang J., Zhang S., “DeepFaceLab: An Integrated, Flexible and Extensible Face-Swapping Framework”, arXiv Preprint, 2020, arXiv:2005.05535.
26. Kowalski M., “FaceSwap”, GitHub Repository, 2021, Available online: https://github.com/MarekKowalski/FaceSwap (Accessed on 6 November 2022).
27. Stroebel L., Llewellyn M., Hartley T., Ip T.S., Ahmed M., “A Systematic Literature Review on the Effectiveness of Deepfake Detection Techniques”, Journal of Cyber Security Technology, 2023, 7(2), pp. 83–113.
28. Cunha L., Zhang L., Sowan B., Lim C.P., Kong Y., “Video Deepfake Detection Using Particle Swarm Optimization Improved Deep Neural Networks”, Neural Computing and Applications, 2024, 36(15), pp. 8417–8453.
29. Su P.C., Huang B.H., Kuo T.Y., “UFCC: A Unified Forensic Approach to Locating Tampered Areas in Still Images and Detecting Deepfake Videos by Evaluating Content Consistency”, Electronics, 2024, 13(4), 804.
30. Chen Z., Liao X., Wu X., Chen Y., “Compressed Deepfake Video Detection Based on 3D Spatiotemporal Trajectories”, arXiv Preprint, 2024, arXiv:2404.18149.
31. Gao Y., Zhang Y., Zeng P., Ma Y., “Refining Localized Attention Features with Multi-Scale Relationships for Enhanced Deepfake Detection in Spatial-Frequency Domain”, Electronics, 2024, 13(9), 1749.
32. Alhaji H.S., Celik Y., Goel S., “An Approach to Deepfake Video Detection Based on ACO-PSO Features and Deep Learning”, Electronics, 2024, 13(12), 2398.
33. Reddy E.M.S., Kumar A.P., Swetha P., “Deepfake Video Detection Using CNN and RNN with Optical Flow Features”, IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), 2024, pp. 1–7.
34. Heidari A., Jafari Navimipour N., Dag H., Unal M., “Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2024, 14(2), e1520.
35. Berrahal M., Boukabous M., Idrissi I., “A Comparative Analysis of Fake Image Detection in Generative Adversarial Networks and Variational Autoencoders”, International Conference on Decision Aid Sciences and Applications (DASA), IEEE, 2023, pp. 223–230.

36. Cheng Z., Wang Y., Wan Y., Jiang C., "Deepfake Detection Method Based on Multi-Scale Interactive Dual-Stream Network", Journal of Visual Communication and Image Representation, 2024, 104, 104263.