IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

Deepfake Detection using ViT_B_16 model

Dr. Pritesh Patil[1], Chinmay Nakwa[2], Hrishikesh Wadile[3]
Professor, Department of Information Technology [1]
Students, Department of Information Technology[2,3]
AISSMS Institute of Information Technology, Pune, India

Abstract: *The technology of Deepfake has rapidly evolved in today's world, it poses significant challenges to society and individuals as it enables high realistic fake images, audios and videos. There is increase of risks of deception, misinformation, and reputational damage due to these advancements. To counteract this emerging threat, we have explored Vision Transformer (ViT)-based models for deepfake detection, leveraging deep learning techniques. Our study implements ViT models —ViT-B-16 trained on datasets of 5,000 images. A Flutter-based application is developed to classify uploaded images as real or fake, providing a prediction confidence score. Experimental results indicate that the ViT-based models achieve promising detection performance, with the highest accuracy reaching 87.33%. Our research highlights the importance advanced architectures in improving deepfake detection techniques. The study of Vision Transformers, showcase the potential in tackling deepfake challenges. Our research contributes to the ongoing and future efforts to enhance the deepfake detection techniques and mitigate its social, personal and environmental impacts.*

Keywords: Deepfake detection, Vision Transformer(ViT), Deep Learning, Image Classification, Misinformation prevention

I. INTRODUCTION

Deep-learning based approach which can manipulate facial images in videos, which can make a target person appear to be doing or saying things they never did is known as Deepfake. Deepfake technology have potential for legitimate applications but, it is often misused in creation of misleading content, such as spreading misinformation, defaming celebrities, and causing economic chaos.[1] First-ever deepfake videos was emerged in 2017 when a Reddit user swapped superstars' faces by altering videos into inappropriate content.[9] Since then, various such deepfakes are being developed and also detection techniques are being developed side-by-side to them for detection to mitigate the harmful effects the images possess.[8]

As in today's world, Machine Learning(ML), Artificial Intelligence(AI), and Deep Learning(DL) techniques, editing digital content has become more accessible publicly. There is a significant contribution of Generative Adversarial Networks (GANs) in development of deepfakes. There are two competing neural networks in GANs: consisting of a generator and a discriminator. Generator are used to create synthetic images and the discriminator are used to evaluate the authenticity. As these two train together, the generator improves at creating more realistic images, which makes the distinguishing of real from fake even harder.[2]

Significant risks, including security threats to governmental institutions and privacy violations for individuals are posed by deepfakes. Malicious actors use deepfake algorithms to spread illegal content, including misinformation, digital kidnapping, and cyber fraud. Various deepfake generation techniques, such as FaceSwap, are exploited to bypass authentication systems, making detection crucial. Several machine learning, artificial intelligence and deep learning-based methods have been developed to detect fake images, with Vision Transformer (ViT) models show promising results due to their self-attention mechanism.[6]

To detect deepfake content, researchers have explored various types of convolutional neural network (CNN) architectures, such as InceptionV3, MesoInception4, ResNet50, XceptionNet, Meso4, GoogLeNet, and FWA-based Dual Spatial Pyramid, VGG19-based CapsuleNet. Many of these models have been trained on huge datasets, including

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

Celeb-DF, to evaluate how effective they are. Additionally, CNNs and BlazeFace techniques are used to face region extraction using multitask have been employed to improve detection accuracy. [10]

II. RELATED WORKS

The technology of Deepfakes has advanced significantly, making fake videos, nearly indistinguishable to the human eye. Due to these concerns are growing in media forensics, prompting extensive research into Deepfake detection techniques. Convolutional Neural Networks (CNNs) are widely used for extracting frame features, while Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs) analyze temporal sequences. More such advanced approaches, such as Face X-ray, estimated heart rate detection, and Vision Transformers (ViTs), have also been explored.[6][10]

In Deepfake detection Vision Transformers show promising results due to their attention mechanism, which preserves high-level information which is often lost in CNNs. LSTMs with InceptionV3 to analyse sequential frames were used by researchers like Guera & Delp, while a two-branch Gaussian Laplacian (LoG) model to enhance forgery detection by suppressing facial content and amplifying multi-band frequencies was employed by Masi et al. Spatial-Multiple Instance Learning (S-MIL) to detect inconsistencies in partial faces, improving overall detection accuracy were introduced by Li et al.[11]

Other methods include assembling CNN models like EfficientNetB4 with attention layers, YOLO-CRNNs for facial region detection, and MesoNet-based classification focusing on facial texture analysis. Some studies utilize biological signals, such as teeth and mouth movements, to improve Deepfake detection. Vision Transformer models combined with CNNs have also demonstrated strong results. However, a key challenge remains: generalizing detection models across different datasets. Researchers continue refining deep learning techniques to enhance accuracy, detect forged elements, and improve robustness against sophisticated Deepfakes.[4][5]

III. LITERATURE REVIEW

TABLE I Literature Review

Paper Number	Year	Key Focus	Methodology/ Model Used	Key Findings	Summary
[1]	2020	Overview of Deepfake technology	Literature review of deepfake detection and generation techniques	Discusses the evolution, applications, and challenges of Deepfake detection	Provides a broad review of Deepfake technology, highlighting its impact and potential threats.
[2]	2017	Introduction to GANs and their applications	Overview of different GAN architectures	Highlights GANs' potential in image synthesis and anomaly detection	Provides a foundational understanding of how GANs work and their future potential.
[3]	2024	Monitoring vehicle loads using computer vision	Deep learning-based image analysis	Vehicle loads real-time identification for bridge health monitoring	Demonstrates the effectiveness of AI in bridge safety assessment.
[4]	2022	Deepfake detection	Convolutional Neural Networks (CNN) and Recurrent Convolutional Neural Networks	Achieved high accuracy in detecting Deepfake videos	Proposes a hybrid CNN-RCNN model for effective Deepfake detection.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

			(RCNN)		
[5]	2023	Image-based Deepfake detection	Customized Convolutional Neural Network (CNN)	Improves Deepfake detection through CNN-based image classification	Introduces an optimized CNN for identifying Deepfake images.
[6]	2022	Weld pool image classification	Vision Transformer (ViT)	Achieved high accuracy in predicting penetration state using ViTs	Demonstrates the application of ViT in industrial welding inspection.
[7]	2022	Brain tumor classification	Vision Transformer (ViT) ensemble model	ViTs outperform traditional CNNs in medical imaging tasks	Highlights the advantages of ViTs for medical image analysis.
[8]	2021	Deepfake detection in social media	Machine Learning (ML) with keyframe extraction	Keyframe-based analysis improves Deepfake detection efficiency	Suggests a novel ML-based approach for detecting Deepfakes in social media videos.
[9]	2022	Survey of methods of Deepfake detection	Review of existing Deepfake detection models	Summarizes advancements and challenges in Deepfake detection	Provides a comprehensive analysis of Deepfake detection strategies.
[10]	2022	Video detection of Deepfakes	Convolutional Neural Network (CNN)	CNNs effectively identify manipulated videos	Proposes an optimized CNN architecture for Deepfake detection.
[11]	2021	Deepfake detection	CNN + RecurrentNeural Network(RNN) hybrid model	Combining CNN and RNN improves Deepfake detection performance	Introduces a hybrid model leveraging both CNN and RNN for enhanced detection.

IV. VIT BASED DEEPFAKE DETECTION APPLICATION

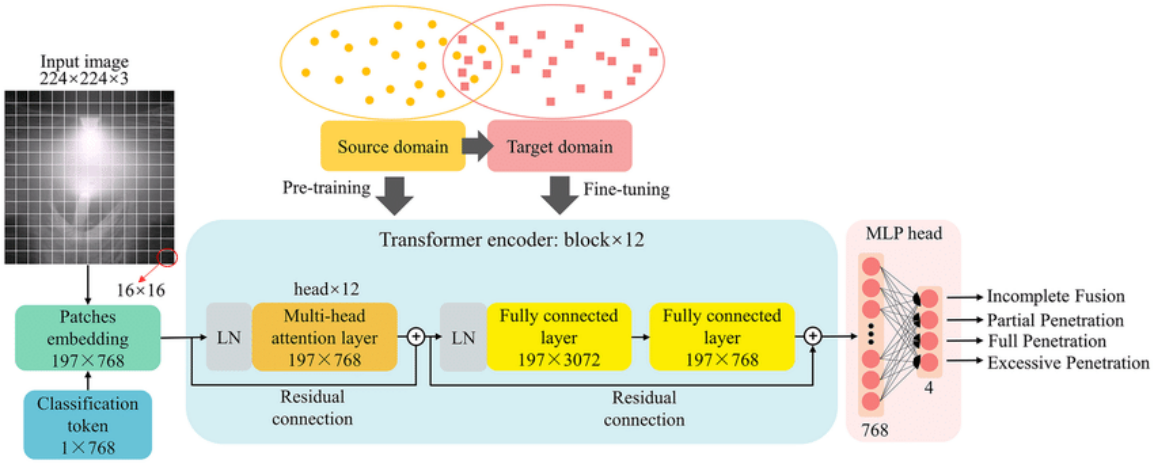

Figure 1 The schematic diagram of transfer learning and the overall architecture of the pretrained ViT-B/16 for penetration recognition

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

ViT_B_16 model[6][7]

The Vision Transformer, ViT_B_16 is a variant of the Vision Transformer architecture, which applies the transformer model it is originally designed for Natural Language Processing (NLP) for computer vision tasks. The "B_16" denotes a Base-sized model with 16x16 patch embeddings, creating a balanced choice between computational efficiency and performance

Architecture Breakdown:

ViT_B_16 divides an input image (e.g., 224x224 pixels) into fixed-size 16x16 non-overlapping patches, flattening them into vectors. Each flattened patch (768 dimensions for RGB images) is projected to the model's hidden dimension (typically 768 for ViT-B) via a trainable linear layer, creating a sequence of patch tokens.

Since transformers lack inherent spatial understanding, learnable positional embeddings are added to the patch tokens to preserve spatial information. The architecture consists of multiple transformer encoder layers (12 in ViT-B), each containing Multi-Head Self-Attention (MSA) and MLP blocks. Layer Normalization (LN) and residual connections are employed to stabilize training.

A special classification token [CLS] is prepended to the sequence of patch tokens, which aggregates global information through the transformer layers and is used by an MLP head for final classification.

Key Features:

A balance between model size (~86M params) and accuracy, outperforming CNNs like ResNet when pre-trained on large datasets (e.g., ImageNet-21k) is striked by ViT_B_16. Confined receptive fields, self-attention captures long-range interdependence which are in contrast to CNNs. Finer patches, 16x16 patches shorten the sequence length (e.g., 196 for 224x224 photos), which lowers the computing cost when compared.

Limitations:

Needs extensive pre-training, in contrast to CNNs, which perform well when generalized from smaller datasets. The input size must be the same as it was before training (e.g., 224x224), though flexible or hybrid ViTs help to mitigate this.

Application

The application uses Vision Transformer (ViT) models to identify deepfake images. It is composed of a FastAPI backend for model inference and a Flutter-based frontend for user interaction. The user may use the program to obtain a prediction that indicates if the image is authentic or not, along with a confidence prediction score, after selecting a ViT model and uploading an image.

Components:

Home Page (home_page.dart): This page allows the user to select the image to test for authenticity or falsity and select the ViT model using a dropdown menu.

Result Page (result_page.dart): Prediction results based on the image and model chosen by the user is shown in this page such that the image is real or fake with a confidence prediction score.

Deepfake detection based on ViT model is implemented by the Deepfake Detection Model (deepfake.py). According to the selected model the image is processed, and the outputs are given to the app for display.

Workflow:

When the app is opened user can see the home page of the app

User can choose an image from gallery to check for deepfake

The user may select the desired model using the dropdown menu.

The image is submitted when the submit button is clicked.

The image and chosen model are sent to deepfake.py over a web socket connection for processing.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

The backend loads the selected ViT model and carries out image processing after receiving the picture and model path.
The model predicts whether or not the image is legitimate and calculates the confidence score.
The frontend receives results from the backend.
The frontend receives the predicted results and forwards them to the result page.
Submitted image is shown on the result page and below that the confidence score (e.g., "Probability: 86.9%") and the prediction (e.g., "Fake" or "Real") is shown.

V. GRAPHS AND OBSERVATIONS

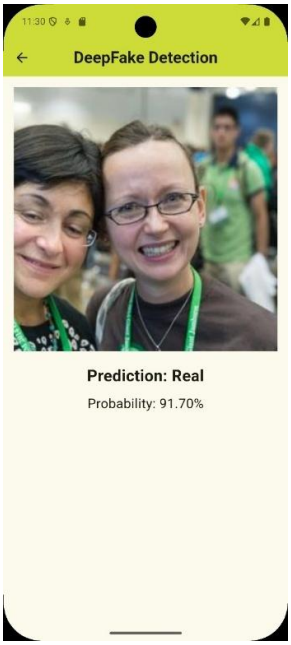

Figure 2 Predicted real image by the ViT_B_16 model

Figure 3 Predicted Fake image by the ViT_B_16 model

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

Outcomes of the detection of deepfake is displayed by the given test images. First image is real is predicted with a probability of 91.70%. This suggests that the image possesses natural facial features without noticeable manipulations. However, in the probability score we can see a little level of uncertainty, implying minor elements that could resemble deepfake characteristics but were not strong enough to alter the classification.

Second image says that 94.20% the image is a fake, which implies that the model has recognised traits like uneven facial features, uneven lighting, or abnormal skin textures that are frequently linked to deepfake creation. Second image says that 94.20%, the image is a fake, which implies that the model has recognized traits like uneven facial features, uneven lighting, or abnormal skin textures that are frequently linked to deepfake creation. Though there is uncertainty for the rest of the percentage that the image can be real.

The model has moderate level of confidence and can verify whether an image is real or fake with that; however, dependability on the model could be increased with more accuracy improvements.

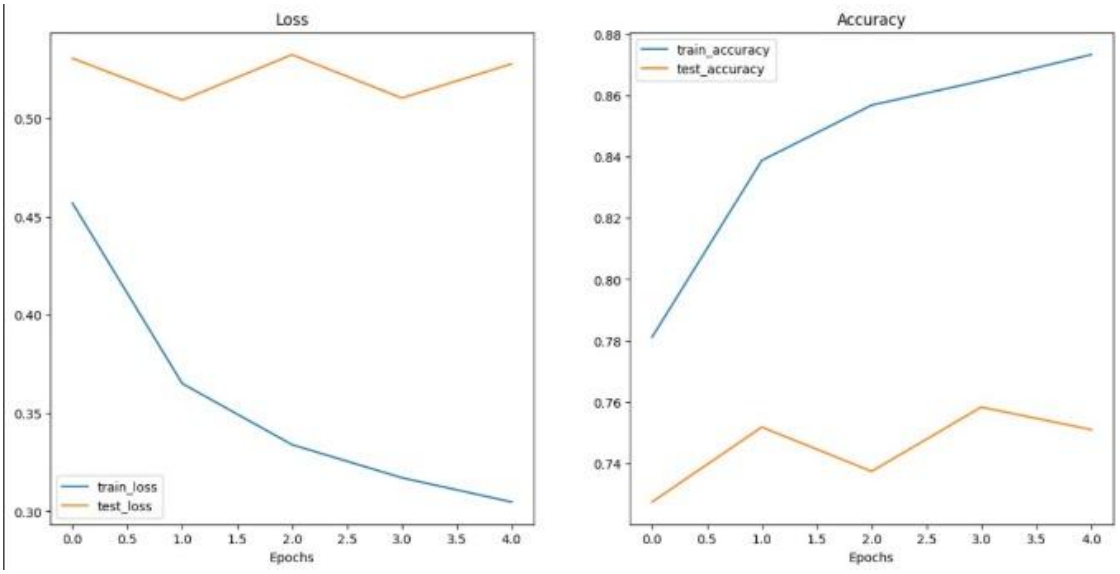

Figure 4 Training vs Testing Performance Metrics

The model is learning from the training dataset when the training loss, which is shown by the blue line, progressively drops. The test loss, shown by the orange line, which is rather constant with just slight fluctuations, suggests that the model may have trouble generalizing the test data. By the increase in the difference between training and test loss we can clearly observe that the model may be overfitting, where it learns patterns from the training data but performs badly on unknown data.

The training accuracy, shown by the blue line, shows a consistent upward trend, which indicates the model's ability to learn and adapt to the training dataset. The concern over insufficient generalization is further supported by the fact that the test accuracy, represented by the orange line, fluctuates and does not appreciably increase across the epochs. The increasing disparity between test and training accuracy draws attention to the overfitting issue. By this we can observe that the model is struggling with test samples but performing considerably better on the training data.

VI. Results and Discussions

TABLE II: Training and Testing Performance Summary

Training Loss	Training Accuracy	Test Loss	Test Accuracy
0.3047	0.8733	0.5278	0.7510

Our trained model showed a high learning ability and it has a training accuracy of 87.33% and a training loss of 0.3047after a total of 10 epochs. Although the model does well on training data, it need improvement in terms of generalization for test data that is not visible.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, April 2025 Impact Factor: 7.67

To efficiently categorize images as real or false, the deepfake detection program uses the ViT_B_16 model. Based on the Flutter and Python code, the application successfully detects user-uploaded photographs and predicts authenticity using a confidence likelihood score. While the backend model (deepfake.py) handles the inference, the Flutter UI (main.dart, home_page.dart, result_page.dart) makes interaction fluid. However, some uncertainty might still exist due to changes in likelihood. For more accurate real-world deepfake detection, our future developments might combine post-processing, optimize inference time, and improve generalization.

VII. CONCLUSION

By using their attention mechanisms to detect small distortions in modified photos, Vision Transformers showed great promise in deepfake identification. The achieved accuracy of 87.33% validates their suitability for this task. The model's performance is dataset-dependent; larger and more diverse datasets could improve generalization was a huge challenge. High-quality deepfakes (e.g., those generated by advanced GANs) remain challenging to detect even now. In the future, we can use multimodal inputs (such audio and video) to increase detection robustness. Experiment with bigger ViT designs (such as the ViT-L-16) and hybrid CNN-ViT models. Expand the dataset's range of deepfake methods (such as FaceSwap and Neural Textures). This work contributes to lowering the risks associated with deepfakes by providing individuals and organizations with a scalable, user-friendly method of verifying the authenticity of digital material (via the Flutter app).

REFERENCES

[1]. Mahmud, Bahar & Sharmin, Afsana. (2020). Deep Insights of Deepfake Technology : A Review.

[2]. Gou, Chao & Duan, Yanjie & Yilun, Lin & Zheng, Xinhu. (2017). Generative Adversarial Networks: Introduction and Outlook. 4. 588-598. 10.1109/JAS.2017.7510583.

[3]. Jiaxin Yang, Yan Bao, Zhe Sun, Xiaolin Meng, Computer Vision-Based Real-Time Identification of Vehicle Loads for Structural Health Monitoring of Bridges, Sustainability, 10.3390/su16031081, 16, 3, (1081), (2024).

[4]. Rahman, Ashifur & Siddique, Nipo & Moon, Mohasina & Islam, Md. Mazharul & Tasnim, Tahera. (2022). Deepfake Video Detection Using CNN and RCNN This project "Deepfake Video Detection Using CNN and RCNN" report submitted by Ashifur. 10.13140/RG.2.2.16605.69607.

[5]. Usha Kosarkar, Gopal Sarkarkar, Shilpa Gedam, Revealing and Classification of Deepfakes Video's Images using a Customize Convolution Neural Network Model, Procedia Computer Science, Volume 218, 2023.

[6]. Wang, Zhenmin & Chen, Haoyu & Zhong, Qiming & Lin, Sanbao & Wu, Jianwen & Xu, Mengjia & Zhang, Qin. (2022). Recognition of penetration state in GTAW based on vision transformer using weld pool image. The International Journal of Advanced Manufacturing Technology. 119. 1-14. 10.1007/s00170-021-08538-6.

[7]. Tummala, Sudhakar & Kadry, Seifedine & Bukhari, Syed & Rauf, Hafiz Tayyab. (2022). Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling. Current Oncology. 29. 7498-7511. 10.3390/curroncol29100590.

[8]. Mitra, Alakananda & Mohanty, Saraju & Corcoran, Peter & Kougianos, Elias. (2021). A Machine Learning Based Approach for Deepfake Detection in Social Media Through Key Video Frame Extraction. SN Computer Science. 2. 10.1007/s42979-021-00495-x.

[9]. M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[10]. V. Jolly, M. Telrandhe, A. Kasat, A. Shitole and K. Gawande, "CNN based Deep Learning model for Deepfake Detection," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-5, doi: 10.1109/ASIANCON55314.2022.9908862.

[11]. Y. Al-Dhabi and S. Zhang, "Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)," 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), SC, USA, 2021, pp. 236-241, doi: 10.1109/CSAIEE54046.2021.9543264.