IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A HYBRID CNN–VISION TRANSFORMER FRAMEWORK FOR ROBUST DEEPFAKE VIDEO DETECTION

[1]Shivani,[2]Sakshi, [3]Sneha Shalgar, [4]Rekha , [5]Asst.Prof.Swati V P

[1-4]Students, [5]Assistant Professor

[1-5]Computer Science & Engineering,

[1-5]Sharnbasva University,Kalaburagi,Karnataka,India

Abstract: Deepfake technology uses advanced artificial intelligence and deep learning techniques to generate highly realistic manipulated videos that imitate human appearance, facial expressions, and behavior. The rapid growth of deepfake generation methods has created significant challenges related to misinformation, identity fraud, cybersecurity, and digital media authenticity. Detecting manipulated content has become increasingly difficult due to continuous improvements in visual quality and synthesis techniques. This study presents an effective deepfake video detection framework based on EfficientNet-B0 and Vision Transformer architectures. EfficientNet-B0 serves as the primary feature extraction model, capturing subtle facial artifacts such as texture inconsistencies, blending errors, and unnatural boundaries, while the Vision Transformer analyzes global contextual relationships through an attention mechanism. A structured preprocessing pipeline involving frame extraction, face detection, cropping, normalization, and data augmentation is employed to enhance model robustness and generalization. The framework is trained and evaluated using benchmark deepfake datasets containing diverse manipulation techniques. A Flask based web application is integrated for video upload and prediction. Experimental results demonstrate high detection accuracy, reliability, and effectiveness in distinguishing real and manipulated videos.

Index Terms – Deepfake Detection, EfficientNet-B0, Vision Transformer, Deep Learning, Computer Vision, Media Forensics, Video Classification, Flask Application.

I. INTRODUCTION

The rapid advancement of artificial intelligence and multimedia technologies has transformed the way digital content is created, shared, and consumed across online platforms. Images and videos have become powerful communication mediums in education, entertainment, journalism, social networking, and business applications. Along with these developments, sophisticated content generation techniques have emerged that can automatically synthesize highly realistic visual media. While such innovations offer numerous benefits for creative industries, they also introduce serious concerns regarding authenticity, trustworthiness, and information security in digital environments.Among these emerging technologies, deepfake generation has gained significant attention due to its ability to manipulate facial appearances, expressions, and identities with remarkable realism. Deepfake videos are commonly created using deep learning models such as Generative Adversarial Networks (GANs), autoencoders, and face reenactment techniques that can produce convincing synthetic content. The increasing accessibility of these tools has raised concerns related to misinformation, identity theft, political manipulation, financial fraud, and reputational damage. As deepfake generation methods continue to improve, distinguishing manipulated

content from authentic videos has become increasingly challenging.Early deepfake detection approaches focused on identifying visible artifacts, facial inconsistencies, and physiological irregularities such as abnormal blinking patterns and unnatural head movements. Although these methods achieved promising results, modern generation techniques have significantly reduced such detectable cues, limiting the effectiveness of traditional detection systems. Convolutional Neural Networks (CNNs) have demonstrated strong capabilities in extracting discriminative spatial features and detecting subtle manipulation artifacts present in forged videos. However, CNN-based approaches primarily focus on local feature learning and may not effectively capture long-range dependencies and global contextual relationships across facial regions.Recent advances in Vision Transformers have introduced attention-based mechanisms capable of modeling global feature interactions and complex spatial dependencies. These architectures provide enhanced representation learning for computer vision tasks and have shown considerable potential in multimedia forensics applications. Combining convolutional feature extraction with transformer-based attention enables effective analysis of both local artifacts and global contextual inconsistencies present in manipulated videos.This study presents a deepfake video detection framework based on EfficientNet-B0 and Vision Transformer architectures. EfficientNet-B0 performs robust facial feature extraction, while the Vision Transformer enhances classification through global attention mechanisms. A comprehensive preprocessing pipeline, benchmark dataset evaluation, and Flask-based deployment are incorporated to develop an accurate, scalable, and practical solution for deepfake video detection and media authentication applications.

II. Related Works

Article [1] "Deepfake Detection with Automatic Face Weighting" by Daniel Mas Montserrat, Hao Hao Nguyen, Junichi Yamagishi, and Isao Echizen in 2020: This paper presents a deepfake detection framework that combines convolutional neural networks with automatic face weighting mechanisms. The method identifies important facial regions and assigns adaptive weights during classification. Spatial and temporal features are extracted from video frames to improve detection accuracy. The framework is evaluated using benchmark deepfake datasets. Experimental results demonstrate improved performance compared with conventional CNN approaches. The automatic weighting strategy helps the model focus on manipulated facial areas. The study highlights the significance of facial attention mechanisms. The proposed approach improves robustness against different manipulation techniques.

Article [2] "DeepfakeStack: A Deep Ensemble-Based Learning Technique for Deepfake Detection" by Md. Shohel Rana and Andrew H. Sung in 2020: This study introduces an ensemble learning framework for deepfake detection. Multiple pretrained CNN architectures are combined using a stacking strategy. The system integrates DenseNet, ResNet, and Xception models to improve classification performance. Feature diversity from different networks enhances robustness. Extensive experiments are conducted on benchmark datasets. Results show higher accuracy compared with individual classifiers. The ensemble approach reduces model bias and variance. The work demonstrates the effectiveness of combining multiple deep learning architectures.

Article [3] "M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection" by Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Ser-Nam Lim, and Yu-Gang Jiang in 2021: This paper proposes a transformer-based deepfake detection model that captures manipulation artifacts at multiple spatial scales. The framework combines RGB information and frequency-domain features. Multi-scale transformers learn local and global inconsistencies simultaneously. Cross-modal fusion improves feature representation quality. The model is evaluated on several benchmark datasets. Experimental results demonstrate superior performance over traditional CNN approaches. The architecture effectively captures subtle forgery traces. The study highlights the advantages of transformer models for deepfake forensics.

Article [4] "Combining EfficientNet and Vision Transformers for Video Deepfake Detection" by Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi in 2021: This work introduces a hybrid architecture that combines EfficientNet and Vision Transformers for video deepfake detection. EfficientNet is used for extracting discriminative facial features from video frames. Vision Transformers model global contextual relationships through attention mechanisms. The framework is evaluated using the DeepFake Detection Challenge dataset. Results indicate strong classification

performance with improved generalization. The architecture balances accuracy and computational efficiency. The study demonstrates the effectiveness of combining CNN and transformer models. This work directly motivates hybrid deepfake detection frameworks.

Article [5] "Vision Transformer for Deepfake Detection" by Jiapeng Li, Xiaoyang Li, and Qiang Ji in 2022: This paper investigates the application of Vision Transformers for detecting manipulated media. The model leverages self-attention mechanisms to capture global image dependencies. Unlike CNN-based approaches, the architecture analyzes entire facial regions simultaneously. Extensive experiments are conducted on deepfake datasets. Results show competitive performance against state-of-the-art methods. The transformer architecture effectively learns manipulation patterns. The study demonstrates the feasibility of transformer-based media forensics. The work provides a foundation for subsequent hybrid models.

Article [6] "DeepFake Detection Algorithm Based on Improved Vision Transformer" by Young Jae Heo, Young Ho Kim, and Kang Ryoung Park in 2023: This study proposes an enhanced Vision Transformer architecture for deepfake detection. The framework extracts both local and global features from manipulated facial images. Additional optimization strategies improve learning efficiency and classification accuracy. Benchmark datasets are used for validation. Experimental results indicate superior performance compared with standard transformer models. The architecture successfully captures subtle forgery artifacts. The study emphasizes the importance of feature interaction mechanisms. The proposed method achieves strong generalization performance.

Article [7] "Deepfake Image Detection Using Vision Transformer Models" by Bogdan Ghita, Ievgeniia Kuzminykh, Abubakar Usama, Taimur Bakhshi, and Jims Marchang in 2024: This IEEE conference paper evaluates Vision Transformer models for deepfake image detection. The study focuses on identifying manipulated facial content through attention-based learning. Multiple transformer configurations are analyzed and compared. The framework is tested on challenging deepfake datasets. Results indicate strong detection capability and improved feature representation. The architecture effectively captures global contextual information. The study discusses challenges associated with evolving manipulation methods. Findings support the adoption of transformers in digital forensics.

Article [8] "Deepfake Detection Using Convolutional Vision Transformers" by Ahmed Hassan Soudy, Mohamed Abdelmaksoud, and Mohamed Elmogy in 2024: This research presents a convolutional vision transformer framework for deepfake detection. The system integrates preprocessing, feature extraction, and prediction modules. CNN layers identify local facial artifacts, while transformers capture global dependencies. Face alignment and cropping improve input quality. Majority voting is applied for final prediction. Experimental evaluation demonstrates high classification accuracy. The framework performs effectively across diverse datasets. The study highlights the benefits of hybrid architectures.

Article [9] "MEViT: Generalization of Deepfake Detection With Meta Learning and Efficient Vision Transformer" by Van Nam Tran, Duc Anh Nguyen, and Minh Nguyen in 2025: This paper introduces MEViT, a deepfake detection framework combining Efficient Vision Transformers with meta-learning. The model aims to improve cross-dataset generalization. Meta-learning enables rapid adaptation to unseen manipulation techniques. Extensive experiments are conducted on multiple datasets. Results demonstrate improved robustness and detection accuracy. The framework effectively reduces overfitting issues. The study addresses one of the major challenges in deepfake forensics. The proposed architecture enhances practical deployment capability.

Article [10] "The Deepfake Dilemma: Enhancing Deepfake Detection with Vision Transformers" by D. Sumathi, A. Singh, A. Sinha, D. Aditya, and M. R. K. F. in 2025: This study explores the use of Vision Transformers for reliable deepfake detection. The architecture employs attention mechanisms to learn global visual dependencies. Various transformer configurations are evaluated. The framework is tested using benchmark datasets containing manipulated images and videos. Experimental results demonstrate strong classification performance. The model captures complex forgery patterns effectively. The study highlights the advantages of transformer-based approaches. The work contributes to ongoing research in media authentication.

Article [11] "Deepfake Face Recognition using Pretrained Vision Transformers" by Muhammad Fahad, Ahmed Raza, and Muhammad Usman in 2025: This paper investigates pretrained Vision Transformer models for deepfake face analysis. Transfer learning techniques are employed to improve classification accuracy. The framework benefits from pretrained visual representations. Benchmark datasets are used for performance evaluation. Results indicate improved detection capability compared with baseline models. The study demonstrates the effectiveness of transformer-based transfer learning. The approach reduces training complexity and computational cost. The work supports the use of pretrained architectures in forensic applications.

Article [12] "DeepFake Video Detection: Insights into Model Generalisation" by R. Ramanaharan, K. Vidanage, and P. Smith in 2025: This systematic review analyzes modern deepfake detection frameworks and their generalization capabilities. The study evaluates existing CNN, transformer, and hybrid architectures. Cross-dataset performance and robustness are examined extensively. Findings indicate that many models suffer from overfitting and poor generalization. The review identifies major challenges in detecting unseen manipulations. Recommendations are provided for future research directions. The paper highlights the importance of diverse datasets and robust evaluation methods. It serves as a valuable reference for developing generalized deepfake detection systems.

III. Problem Stetement

The rapid advancement of deepfake generation techniques has made manipulated videos increasingly realistic and difficult to distinguish from authentic content. Existing detection systems often struggle to identify subtle forgery artifacts and fail to generalize across different deepfake creation methods. Variations in lighting conditions, facial poses, video compression, image quality, and background environments further reduce detection accuracy. Many traditional approaches rely on handcrafted features or limited visual cues that become ineffective as manipulation technologies continue to evolve. These challenges create a significant need for a robust and reliable deepfake detection system capable of accurately identifying manipulated videos while maintaining strong performance across diverse datasets and real-world scenarios.

IV. Objectives

The primary objective of this study is to develop an accurate and reliable deepfake video detection system using the EfficientNet-B0 and Vision Transformer architectures for identifying manipulated facial content. Another objective is to implement a structured preprocessing pipeline that performs frame extraction, face detection, normalization, and augmentation to improve model performance. The study also aims to train and evaluate the framework using the Fake Videos Dataset obtained from Kaggle to ensure robustness across diverse manipulation techniques. Additionally, the project seeks to integrate the trained model with a Flask-based web application that enables video upload, automated analysis, and user-friendly visualization of detection results.

V. Methodology

1)Data Collection: The dataset used for this project is the Fake Videos Dataset obtained from Kaggle, which contains both real and manipulated videos created using various deepfake generation techniques. The dataset includes videos with different facial expressions, lighting conditions, camera angles, and background environments. Real and fake samples are collected to ensure balanced class distribution during training. The diversity of the dataset helps the model learn a wide range of forgery patterns and improves its ability to generalize to unseen deepfake videos.

2) Data Preprocessing: The collected videos are converted into individual frames at fixed intervals to reduce redundancy while preserving important visual information. Face detection algorithms are applied to identify and crop facial regions from each frame. The extracted faces are resized to a uniform resolution of 224 × 224 pixels and normalized to maintain consistent pixel values. Data augmentation techniques such as rotation, flipping, zooming, and brightness adjustment are applied to increase dataset diversity and reduce overfitting during training.

3) Feature Extraction: Feature extraction is performed using EfficientNet-B0, which serves as the primary convolutional neural network in the proposed framework. The model automatically learns discriminative facial representations from preprocessed images without requiring handcrafted features. It

captures subtle manipulation artifacts such as texture inconsistencies, blending errors, and unnatural facial boundaries. The extracted feature maps provide rich spatial information that is essential for accurate deepfake classification.

4) Model Selection:A hybrid EfficientNet-B0 and Vision Transformer architecture is selected to improve deepfake detection performance. EfficientNet-B0 is responsible for extracting local facial features, while the Vision Transformer captures global contextual relationships through self-attention mechanisms. This combination enables the framework to analyze both fine-grained artifacts and overall facial structures. The selected architecture provides an effective balance between accuracy, robustness, and computational efficiency.

5) Model Training: The selected model is trained using labeled real and fake facial images extracted from the dataset. Binary Cross-Entropy loss is used as the optimization objective for binary classification. The Adam optimizer is employed to update model parameters and improve convergence during training. Multiple training epochs are performed with mini-batch learning, while validation data is used to monitor performance and prevent overfitting.

6) Model Evaluation: The trained model is evaluated using separate validation and testing datasets that were not used during training. Performance is measured using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. These metrics provide a comprehensive assessment of the model's ability to distinguish real videos from manipulated ones. Additional analysis is conducted on misclassified samples to identify limitations and improve future model performance.

7) Integration with Flask:A Flask-based web application is developed to provide an interactive interface for deepfake detection. Users can upload video files through the web interface, which are then processed by the trained model in the backend. The application automatically performs frame extraction, face preprocessing, prediction, and result generation. The final classification result is displayed as either Real or Fake, making the system practical and accessible for real-world use.

VI. SYSTEM ARCHITECTURE

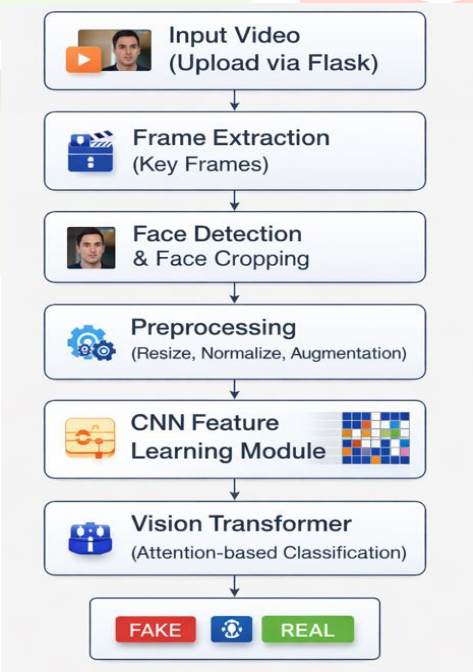

Fig 1: System Architecture of the Proposed Deepfake Video Detection Framework

The system architecture illustrates the complete workflow of the proposed deepfake video detection framework. The process begins when a user uploads a video through the Flask-based web application. The uploaded video is then processed by the frame extraction module, which selects key frames containing important visual information while reducing redundant data. These frames are passed to the face detection and face cropping stage, where facial regions are identified and extracted for further analysis. The cropped face images undergo preprocessing operations including resizing, normalization, and data augmentation

to improve data quality and model robustness. After preprocessing, EfficientNet-B0, functioning as the CNN feature learning module, extracts discriminative facial features and manipulation artifacts such as texture inconsistencies and blending errors. The extracted features are subsequently forwarded to the Vision Transformer, which utilizes attention mechanisms to capture global contextual relationships and enhance classification performance. Finally, the system classifies the input video as either Fake or Real and displays the prediction result through the web interface.

VII. EXPERIMENTAL RESULTS

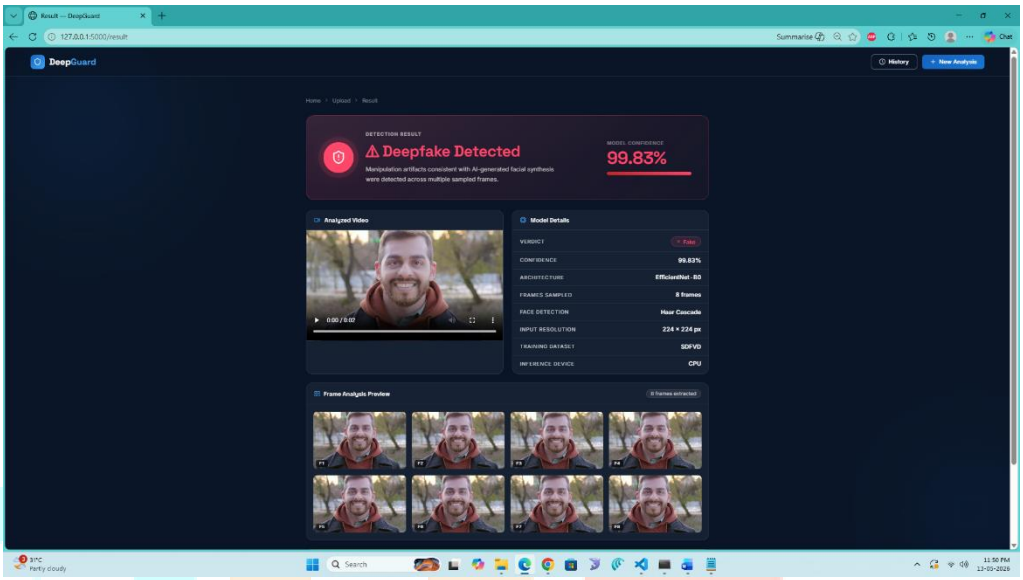

Fig. 2: Deepfake Detection Result Interface Showing Fake Video Classification

The result interface presents the outcome of deepfake video analysis performed by the proposed framework. It provides the final classification label, prediction confidence score, model information, analyzed video preview, and extracted frame samples, allowing users to verify manipulated content through a clear and informative visualization.

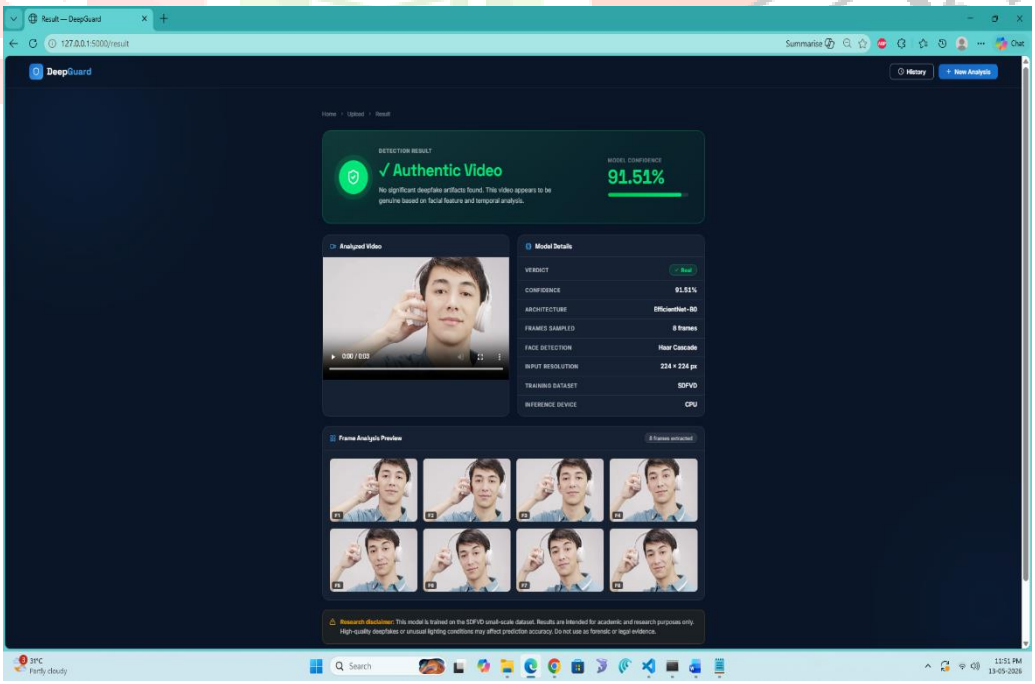

Fig. 2: Deepfake Detection Result Interface Showing Real Video Classification

The result interface presents the outcome of video authenticity analysis performed by the proposed framework. It displays the final classification as a real video along with the confidence score, model

specifications, analyzed video preview, and extracted frame samples, providing users with a comprehensive view of the detection process and verification results.

VIII. CONCLUSION AND FUTURE WORKS

In this research, an effective deepfake video detection framework was developed using a hybrid combination of EfficientNet-B0 and Vision Transformer architectures to identify manipulated facial content with high accuracy. The proposed system incorporated a structured pipeline consisting of video frame extraction, face detection, preprocessing, feature learning, and attention-based classification. EfficientNet-B0 successfully captured subtle manipulation artifacts, while the Vision Transformer enhanced detection by modeling global contextual relationships. Evaluation on the Kaggle Fake Videos Dataset demonstrated reliable performance in distinguishing real and fake videos under varying visual conditions. The integration of a Flask-based web application further improved the practicality of the framework by providing an interactive interface for video upload and automated prediction. The obtained results confirm the effectiveness of combining convolutional and transformer-based learning for multimedia forensics applications. Future work can focus on improving cross-dataset generalization and robustness against newly emerging deepfake generation techniques. The framework may also be extended to support real-time video analysis, audio-visual deepfake detection, cloud deployment, and lightweight mobile implementations. Additional research can explore advanced transformer architectures and larger datasets to further enhance detection accuracy and scalability.

REFERENCES

[1] D. M. Montserrat, H. H. Nguyen, J. Yamagishi, and I. Echizen, “Deepfake Detection with Automatic Face Weighting,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 9351-9355.

[2] M. S. Rana and A. H. Sung, “DeepfakeStack: A Deep Ensemble-Based Learning Technique for Deepfake Detection,” in *Proc. IEEE Int. Conf. Artificial Intelligence and Knowledge Engineering (AIKE)*, 2020, pp. 70-77.

[3] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, S. N. Lim, and Y. G. Jiang, “M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection,” *arXiv preprint arXiv:2104.09770*, 2021.

[4] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, “Combining EfficientNet and Vision Transformers for Video Deepfake Detection,” *arXiv preprint arXiv:2107.02612*, 2021.

[5] J. Li, X. Li, and Q. Ji, “Vision Transformer for Deepfake Detection,” in *Proc. Int. Conf. Pattern Recognition and Artificial Intelligence*, 2022.

[6] Y. J. Heo, Y. H. Kim, and K. R. Park, “Deepfake Detection Algorithm Based on Improved Vision Transformer,” *Applied Intelligence*, vol. 53, no. 4, pp. 4567-4583, 2023.

[7] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, “Deepfake Image Detection Using Vision Transformer Models,” in *Proc. IEEE Int. Conf. Cyber Security and Resilience*, 2024.

[8] A. H. Soudy, M. Abdelmaksoud, and M. Elmogy, “Deepfake Detection Using Convolutional Vision Transformers,” *Neural Computing and Applications*, vol. 36, no. 18, pp. 11245-11261, 2024.

[9] V. N. Tran, D. A. Nguyen, and M. Nguyen, “MEViT: Generalization of Deepfake Detection With Meta Learning and Efficient Vision Transformer,” *IEEE Open Journal of the Computer Society*, vol. 6, pp. 1-12, 2025.

[10] D. Sumathi, A. Singh, A. Sinha, D. Aditya, and M. R. K. F., “The Deepfake Dilemma: Enhancing Deepfake Detection with Vision Transformers,” *Journal of Image Processing and Pattern Recognition Progress*, vol. 12, no. 1, pp. 15-26, 2025.

[11] M. Fahad, A. Raza, and M. Usman, "Deepfake Face Recognition using Pretrained Vision Transformers," in *Proc. Int. Conf. Intelligent Computing and Communication Technologies*, 2025, pp. 221-228.

[12] R. Ramanaharan, K. Vidanage, and P. Smith, "DeepFake Video Detection: Insights into Model Generalisation," *Forensic Science International: Digital Investigation*, vol. 52, pp. 301-315, 2025.