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 6, Issue 5, October 2026
Impact Factor: 8.2

Deepfake Detection using Convolutional Vision Transformers and Convolutional Neural Networks

Sanskruti Walke[1], Shradha Wadkar[2], Lavanya Yengupatla[3], Prof. P. N. Waghule[4], Prof. P. B. Todmal[5]
Degree Students, Department of Computer Engineering[1,2,3]
Faculty, Department of Computer Engineering[4,5]
Adsul Technical Campus, Chas, Ahmadnagar, India

Abstract: *Deepfake technology has rapidly advanced in recent years, creating highly realistic fake videos that can be difficult to distinguish from real ones. The rise of social media platforms and online forums has exacerbated the challenges of detecting misinformation and malicious content. This study leverages many papers on artificial intelligence techniques to address deepfake detection. This research proposes a deep learning (DL)-based method for detecting deepfakes. The system comprises three components: preprocessing, detection, and prediction. Preprocessing includes frame extraction, face detection, alignment, and feature cropping. Convolutional neural networks (CNNs) are employed in the eye and nose feature detection phase. A CNN combined with a vision transformer is also used for face detection. The prediction component employs a majority voting approach, merging results from the three models applied to different features, leading to three individual predictions. The model is trained on various face images using FaceForensics?? and DFDC datasets. Multiple performance metrics, including accuracy, precision, F1, and recall, are used to assess the proposed model's performance. The experimental results indicate the potential and strengths of the proposed CNN that achieved enhanced performance with an accuracy of 97%, while the CViT-based model achieved 85% using the FaceForences?? dataset and demonstrated significant improvements in deepfake detection compared to recent studies, affirming the potential of the suggested framework for detecting deepfakes on social media. This study contributes to a broader understanding of CNN-based DL methods for deepfake detection.*

Keywords: Convolutional neural network Convolutional vision transformer Deepfake detection Face recognition MTCNN FaceForensics?? Computer vision

I. INTRODUCTION

Deepfake is a deep learning-based approach that can add facial images of a target person instead of the main person in a video to create a video of the target person doing or saying things the main person did or said. Deepfake methods generate harm because they can be used to create videos defaming celebrities, creating confusion and chaos in economic markets by developing fake news and misleading individuals. In 2017, the first deepfake video was released when a Reddit user changed superstars' faces into porn videos. Thus, several methods were introduced to detect deepfake videos. The technologies used to edit pictures, videos, and voice-overs are developing rapidly. Techniques for creating and manipulating digital content and technical knowhow are also easily accessible. It is now possible to easily create hyper-realistic digital pictures using a few tools and straightforward how-to guides that are easily available online [9]. Deepfake technology aims to create compelling fake videos that can be difficult to distinguish from real ones. While this technology has the potential to be used for legitimate purposes, it also presents significant challenges in detecting the spread of misinformation and other forms of malicious content. As the use of deepfakes continues to increase, there is a need for effective methods of deepfake detection to protect society as a whole [27].

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

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 6, Issue 5, October 2026
Impact Factor: 8.2

Since the spread of deepfake, many researchers have researched deepfake algorithms to search for weaknesses in Extended author information available on the last page of the article 123 Neural Computing and Applications (2024) 36:19759–19775 https://doi.org/10.1007/s00521-024-10181-7(0123456789().,-volV)(0123456789(). ,- volV) those algorithms. However, each of the existing solutions failed to work when either the progress of the current algorithms, the emergence of new algorithms, the unavailability of a large amount of data, or the stage of data processing applied [34].

II. BACKGROUND

The vision transformer (ViT) model architecture, introduced in research paper [6] extends the Transformer architecture introduced in paper [28] to effectively handle the image domain. Developed by the Google Research Brain Team, the ViT model modifies the implementation to handle image data. In the ViT model, an input image is divided into visual tokens and patches of the image. These visual tokens are embedded into fixed-dimensional encoded vectors, and the position information of each patch is also embedded and combined with the encoded vectors. The transformer encoder network processes this combined representation, similar to text inputs [31].

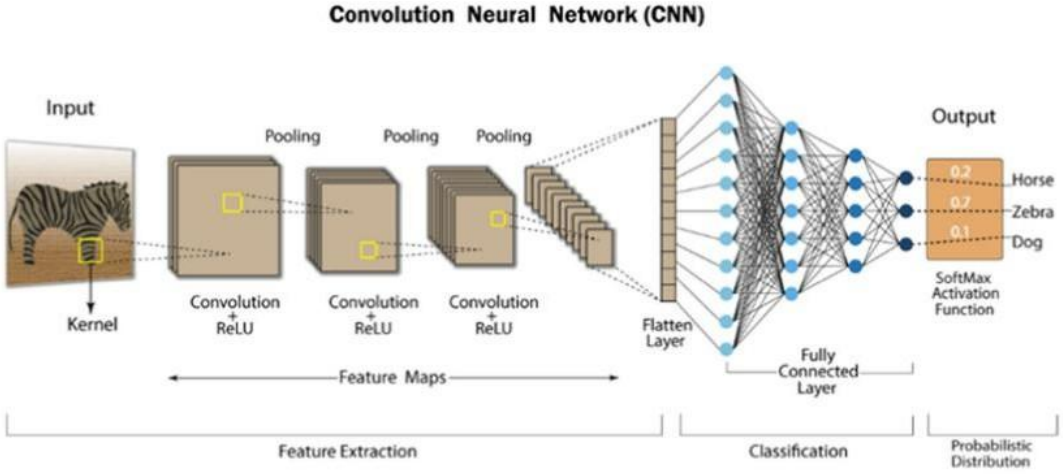
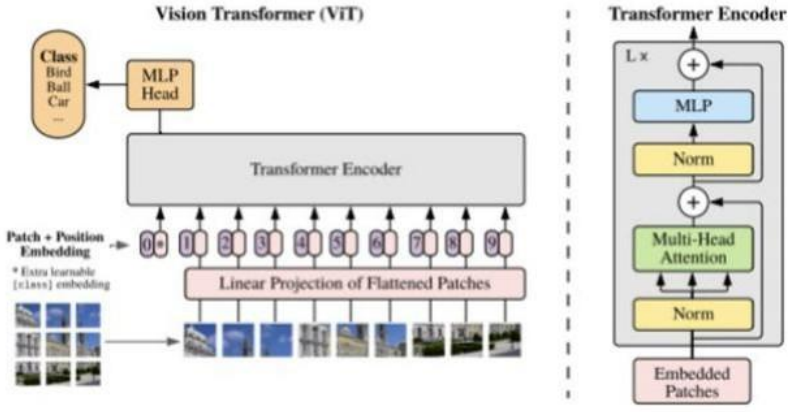


Fig. 1 An overview of CNN architecture

Fig. 2 ViT architecture [19]



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 6, Issue 5, October 2026
Impact Factor: 8.2

III. METHODOLOGY

Our approach consists of three primary phases: preprocessing, detection, and prediction. These phases are illustrated in Fig. 3. Within the preprocessing phase, we extract frames from the video, improve each frame's quality, distinguish the background from the foreground, and then align them accordingly. The subsequent stage is detection, during which the regions encompassing the face, nose, and eyes are identified and cropped from the frame. The cropped face then undergoes detection through three distinct pathways: the first focuses on eye detection, the second on nose detection, and the third on face detection. Within both eye and nose pathways, the eyes and nose are extracted from the face, and after cropping, they are passed to two models, A and B. Each model utilizes a different architecture and possesses a unique layer configuration, which will be expounded upon in the subsequent sections. The outcomes of these models are integrated into the final prediction. The face is directed to model C in the face pathway, which employs an alternative architecture and layers' number. The results of this model contribute to the overall prediction. To ensure reliability, despite the capability of the eye, nose, and face pathways to generate predictions individually, we implement a majority voting approach to consolidate all results into a single outcome. Consequently,

The preprocessing component

The initial data preparation phase involves converting the raw dataset into suitable formats for training, validation, and testing purposes. Our model training and evaluation were conducted using the FaceForensics?? dataset, which comprises authentic and manipulated facial videos. Within the preprocessing stage, four distinct subcomponents are employed: frames extraction, improving each frame's quality, distinguishing the background from the foreground, and then aligning them accordingly. Figure 4 shows an example of how the preprocessing work

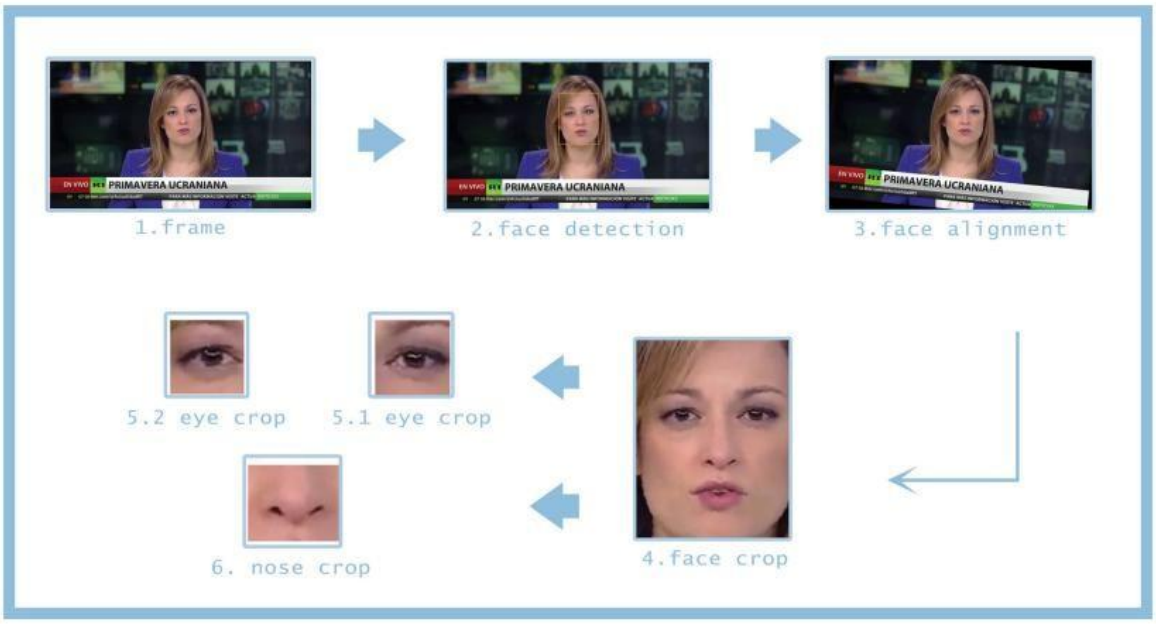


Fig. 4 Example of how our preprocessing steps work

The detection component

Frames extraction entails isolating individual frames from video files, while face detection leverages multitask cascaded convolutional networks (MTCNNs) to pinpoint faces within each frame. Face alignment corrects variations in head pose and facial expression by standardizing the alignment of each face. Subsequently, face cropping trims the aligned face images to a consistent size. Meanwhile, the extraction and cropping of eyes and nose involve identifying and isolating corresponding regions from the aligned face images. The proposed model for identifying deepfakes is

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 6, Issue 5, October 2026
Impact Factor: 8.2

composed of three primary models. These models encompass a CNN-based design tailored for extracting features related to the eyes and nose, an additional CNN-based structure serving the same purpose, and a fusion of a CNN module with a ViT module to analyze the entire face

IV. CONCLUSION

In this study, we introduced a groundbreaking method for deepfake detection, leveraging a fusion of distinct facial features and a comprehensive dataset enhanced by meticulous preprocessing. Our strategy entailed the development of a composite model, integrating three sub-models, each specializing in the recognition of deepfakes by analyzing specific facial elements: the entire face, the eyes, and the nose. Our tailored data processing techniques for each submodel further strengthen this multifaceted approach, circumventing the constraints typically encountered in singlealgorithm detection methods. Our training regimen utilized an expansive array of facial images from the most extensive dataset, such as FaceForensics??. This extensive dataset was pivotal in refining our model/s ability to discern physical anomalies indicative of deepfakes. The empirical evidence from our tests revealed a significant enhancement in accuracy and efficiency over existing Table 6 Comparison between recent deepfake detection techniques References Algorithms Features used Accuracy Dataset Limitations Our Proposed CNNs and CViT Entire face, Eyes, Nose 97% and 85% FaceForences?? High-computational resources CViT-based Technique [30] CViT Entire face 69% FaceForences?? Limited to entire face region CNN-based Technique [32] CNN Eye Region 90% FaceForences?? Limited to eye region Neural Computing and Applications (2024) 36:19759–19775 19773 123 deepfake detection methods, thereby establishing the superiority of our approach. A standout feature of our method is its robust performance across diverse scenarios, encompassing various environmental conditions and facial orientations, illustrating its practical applicability in realworld settings. This adaptability underscores our model's ability to identify deepfakes with high physical fidelity, an essential attribute in the current digital era. The implications of our work are far-reaching, addressing the pressing demand for reliable deepfake detection to thwart the proliferation of misinformation and other harmful digital content.

V. ACKNOWLEDGMENT

Automated quality inspection has historically relied on Conventional Convolutional Neural Networks (CNNs) (such as ResNet or YOLO architectures) due to their robust local feature extraction capabilities. However, as highlighted in comprehensive industrial surveys like the IEEE Review of CNNs for Defect Detection, traditional CNNs face structural limitations in modeling long-range spatial dependencies due to their restricted receptive fields.

To address this constraint, researchers have increasingly deployed Vision Transformers (ViTs), which treat image patches as sequences to capture global self-attention across entire surfaces.

Comparative studies, such as the evaluation of surface defects published in the Journal of Energy Storage, demonstrate that while ViT models frequently surpass CNNs in complex defect classification precision, they present trade-offs regarding computational latency and the volume of training data required."

REFERENCES

[1]. Arif MS, Mukheimer A, Asif D (2023) Enhancing the early detection of chronic kidney disease: a robust machine learning model. Big Data Cogn Comput 7(3):144J.

[2]. Al-Hussein M, Venkataraman S, Jawahar C (2020) Deepfake detection for video: an open source challenge. arXiv preprint arXiv:2006.06058.

[3]. . Altwaijry N, Al-Turaiki I (2021) Arabic handwriting recognition system using convolutional neural network. Neural Comput Appl 33(7):2249–2261.

[4]. Karandikar A, Deshpande V, Singh S, Nagbhidkar S, Agrawal S (2020) Deepfake video detection using convolutional neural network. Int J Adv Trends Comput Sci Eng 9(2):1311–1315.