iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67377

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks

Arifa ok[1], Mr. Pramod K[2]
[1]*MCA Scholar,* [2]*Associate Professor, Department of MC, Nehru College of Engineering and Researce Centre, pampady*

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, Face Forensics++, Computer vision.***

I. INTRODUCTION

Various techniques have been employed to successfully and efficiently identify fake videos. The significant challenges posed by the vast scale and high-dimensionality of deepfake videos are primarily addressed by deep learning methods. Current systems have acknowledged the swift advancement of social media, where users depend on these platforms for the latest news. Consequently, social media sites like WhatsApp, Twitter, Facebook, and YouTube work to filter out fake videos and misleading information from the extensive user-generated content. There exists a potential hazard with the Manufactured Deepfake, which refers to a deep learning-based technique that replaces the primary individual in a video with the facial images of a target person, resulting in a video showcasing the target person supposedly doing or saying things initially expressed by the main individual. The harmful implications of deepfake techniques stem from their capability to produce videos that slander public figures and create confusion and turmoil in financial markets by disseminating false information, thereby misleading the public. The objective of deepfake technology is to generate convincing fraudulent videos that may be challenging to differentiate from authentic ones. Although this technology could serve legitimate purposes, it also brings considerable obstacles in recognizing the spread of misinformation and various malicious content. As the prevalence of deepfakes grows, there is an increasing necessity for robust methods of deepfake detection to safeguard society at large. Deep learning videos are likely to be circulated and shared across social media channels. Operating in these areas presents several difficulties, including (i) identifying the most significant features, (ii) handling videos with greater diversity and dimensionality, and (iii) selecting the appropriate DL model. One commonly used deep learning technique is the convolutional neural network (CNN), favored for its advanced capability to automatically identify both low and high-level features from datasets. The proposed framework consists of three components: preprocessing, detection, and prediction. The preprocessing phase includes the extraction of frames, face detection, alignment of faces, face cropping, as well as cropping the eyes and nose. During the detection phase, a CNN-based architecture is employed for detecting eye and nose features, while a combination of CNN and vision transformer is utilized for comprehensive face detection. In the prediction stage, a majority voting strategy is applied by integrating the outcomes of the three models that utilize three distinct features, resulting in three separate predictions.

II. LITERATURE REVIEW

Andreas et al [1] explore the authenticity of advanced image manipulation techniques and the challenges associated with detecting these alterations, whether through automated methods or human observation. After gathering data, it undergoes manipulation, and then the authenticity of the image is assessed using convolutional neural networks (CNNs).

Yuezun Li et al [2] emphasize the necessity of creating and assessing Deep Fake detection algorithms, which require extensive datasets. However, existing Deep Fake datasets often lack visual quality and do not accurately reflect the Deep Fake videos shared online. The emergence of deep neural networks (DNNs) has simplified and accelerated the creation of convincing fake videos. This study introduces a new, large, and demanding dataset for Deep Fake videos, named Celeb-DF3, aimed at fostering the advancement and assessment of Deep Fake detection algorithms.

Brian et al [3] highlight that the DFDC currently represents the largest publicly accessible dataset of face-swap videos, which contains over 100,000 clips featuring more than 3,426 paid actors. This dataset was generated using a combination of Deep Fakes and both GAN-based and non-learning techniques.

Ricard et al [4] reported that by analyzing a low-resolution video sequence from the FaceForensics++ dataset, their method achieves a 90% accuracy rate in identifying manipulated videos. They address the challenge of detecting artificial images, particularly fake faces, by proposing a new machine learning approach based on classical frequency analysis of images that identifies different behaviors at high frequencies.

Ruben et al [5] provide a detailed review of techniques for detecting and creating altered face images, including those involving Deep Fakes. In particular, they categorize face manipulation into four types: i) full face alterations; ii) identity switching; iii) characteristic modifications; iv) expression changes.

Nicol'o et al [6] tackle the issue of detecting face alterations in video sequences created with modern facial manipulation methods. Utilizing over 10,000 videos, they employ a CNN method to identify false videos.

Wanying Ge et al [7] introduce the application of SHapley Additive exPlanations (SHAP) for uncovering new insights into detection methods. This paper presents a visualization tool called SHAP that aids in interpreting the output of machine learning models by illustrating the impact of each feature on predictions, thereby enhancing understanding. By evaluating the contribution of every feature to the output, it can elucidate the predictions made by any model.

Chunlei Peng et al [8] suggest that assigning distinct scores to both authentic and fabricated face data can improve the model's ability to recognize complex samples with greater nuance. They propose considering the concept of perceptual forgery fidelity, given the intricate nature of facial quality data distributions in the real world. This study replaces traditional binary classification with forgery fidelity scores, mapping facial data of various attributes to discrete values.

Tianchen et al [9] base their work on the premise that unique source features in images can be retained and recovered even after utilizing advanced Deep Fake generation methods. They assert that various source features can be identified at different locations within the manipulated image. By extracting local source features and assessing their self-consistency, it becomes possible to detect counterfeit images.

III. BACKGROUND

This section introduces the main concepts of the methods used, CNN and vision transformer.

A. Convolutional Neural Network

Convolutional Neural Networks (CNNs) are deep learning algorithms frequently utilized in computer vision applications, like image classification and object detection. They are structured to automatically learn and identify significant features from input data, especially images. The design and operation of CNNs are inspired by the structure of the visual cortex found in animals. A key component of CNNs is the convolutional layer, which conducts convolution operations on the input data with a collection of learnable filters or kernels. This convolutional layer utilizes these filters on the input data to identify patterns and features across different spatial areas. It captures local relationships and spatial hierarchies, enabling the network to learn intricate representations of the input images. CNNs commonly incorporate pooling layers to minimize spatial dimensions and maintain the most pertinent information. Figure 1 illustrates the architecture of a CNN. Additionally, CNNs contain fully connected layers that make predictions based on the features learned. These layers take the convolutional layers' output, flatten it, and send it through one or more fully connected layers, ultimately generating the final classification or regression result.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

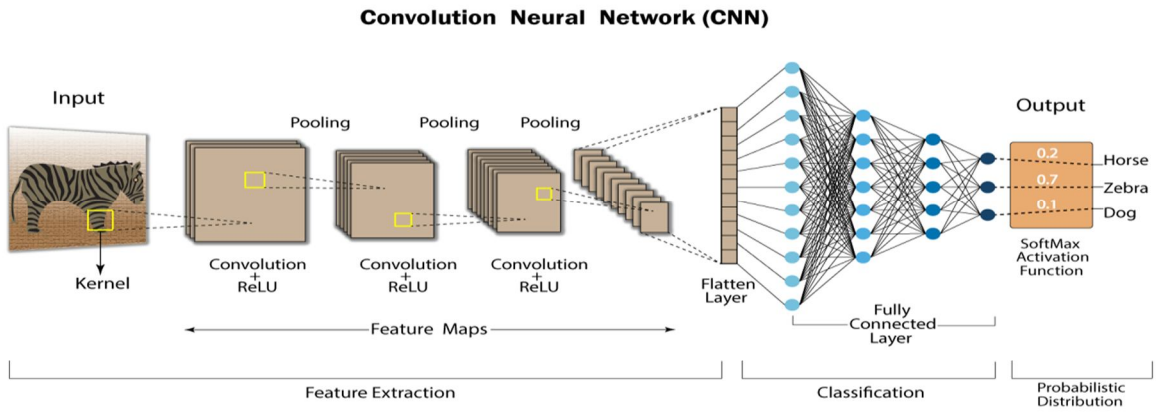


Fig.1 An overview of CNN architecture

B. Vision Transformers

Vision transformers (ViTs) are a specific variety of neural network architecture created for the task of image recognition. In contrast to conventional convolutional neural networks (CNNs), which utilize convolutions for image data processing, ViTs adopt the transformer framework that was initially developed for natural language processing (NLP). In the ViT model, rather than analyzing the entire image simultaneously, ViTs segment the image into smaller patches. Each of these patches is regarded as a token, similar to the approach taken with words in NLP models. Every image patch is flattened into a vector, and a learnable positional embedding is incorporated to preserve spatial information (the location of the patch within the image). This transforms the image into a series of embeddings that the transformer model can then process. These embeddings are transmitted through several transformer layers, which employ mechanisms such as self-attention to analyze the relationships between the patches.

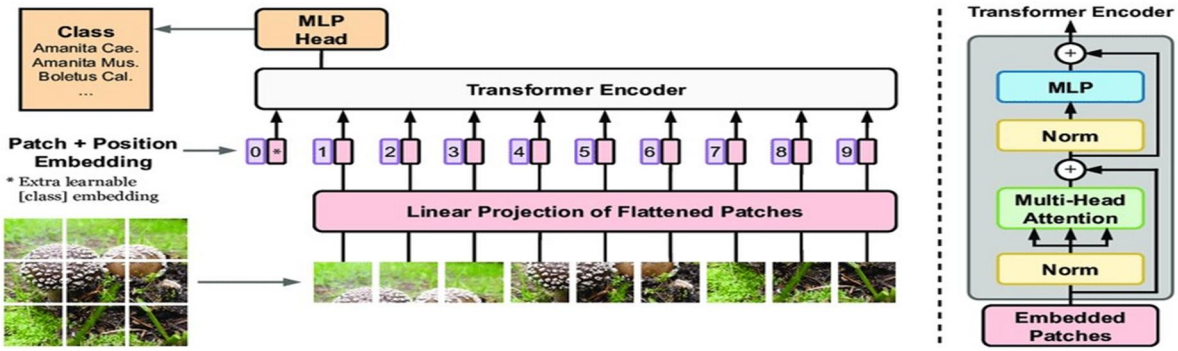


Fig.2 ViT architecture

IV. 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. Consequentl predictions can be made independently for each pathway or using the majority voting approach.

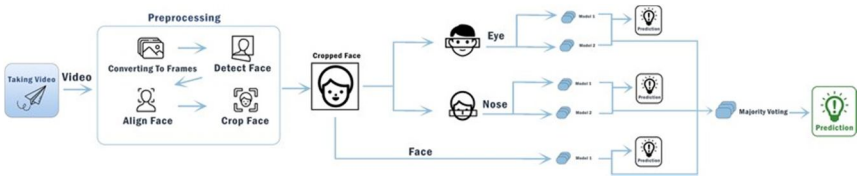


Fig.3 system architecture for preprocessing ,and detection and prediction

A. The Preprocessing Component

3.1 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 fore ground, and then aligning them accordingly.

B. 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 com posed 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 comprehensively. The assessment of machine learning model performance is carried out using K-fold cross validation.

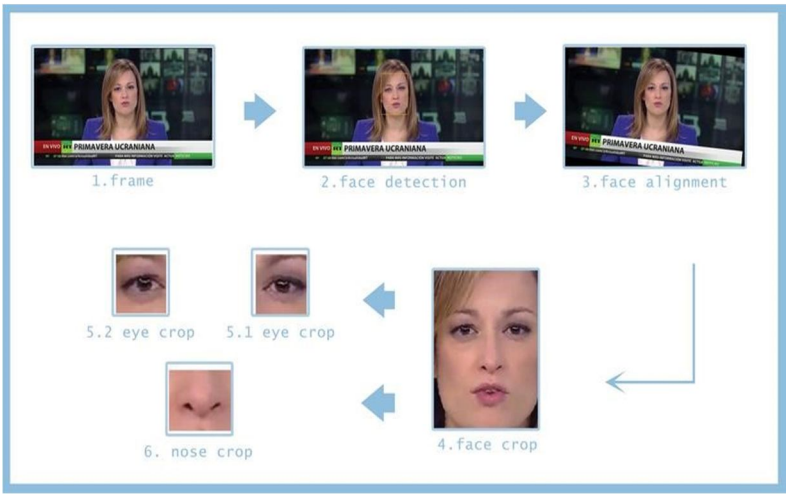


Fig.4 example of how our preprocessing steps work

1) CNN-based architecture for eye and nose regions

(Model A)

Model A is a deep learning architecture comprising 12 layers that adopts a CNN-based methodology. It is structured with three blocks, each containing three Conv2D layers that use ReLU activation to introduce nonlinearity. To improve performance and mitigate overfitting, the model integrates batch normalization, max pooling, and dropout layers. The architecture is designed for input images sized at 50 and is trained using features from the eye and nose. The dataset is divided into 80% for training purposes and 20% for testing. The kernel dimensions are set to (3, 3), the pooling dimensions are (2, 2), and the dropout rate is established at 0.3. This architecture incorporates a fully connected dense layer containing 512 units, followed by a dropout layer and an output layer with two dense units utilizing the softmax activation function. The model employs the Adam optimizer with a learning rate of 0.0001 and undergoes training for a total of 100 epochs. The loss function used is sparse categorical cross-entropy. Our model is illustrated in Fig. 5 and depicted in the accompanying pseudocode.

Algorithm 1 Model A (convolutional neural network with batch normalization and dropout)
1: Input: Image of size 50x50x3
2: Convolutional Block 1:
3: Convolutional layer (3 filters, 3x3 kernel, ReLU activation, padding) 4: Batch normalization
5: Convolutional layer (32 filters, 3x3 kernel, ReLU activation, padding
6: Batch normalization
7: Convolutional layer (32 filters, 3x3 kernel, ReLU activa- tion, padding)
8: Batch normalization
9: Max pooling (2x2 pool size, 2x2 strides)
10: Dropout (rate 0.3)
11: Convolutional Block 2:
12:... (Replicate structure for Convolutional Block 2)
13: Convolutional Block 3:
14:... (Replicate structure for Convolutional Block 3)
15: Flatten the output of the convolutional layers
16: Dense layer (512 units, ReLU activation)
17: Dropout (rate 0.3)
18: Dense layer (2 units, softmax activation)

2) CNN-based architecture for eye and nose regions (Model B)

Model B exhibits a more streamlined architecture than Model A, comprising six layers encompassing three blocks of Conv2D layers. These layers employ the ReLU activation function and incorporate Max Pooling and Dropout layers. The training of Model B is conducted on eye and nose features, utilizing a 50-pixel image size, and follows the same dataset partition as Model A. Similarities persist in terms of kernel size, pool size, activation function, dropout rate, and optimizer shared between Model A and Model B. Specifically; Model B undergoes 150 epochs of training for the eye region and 200 epochs for the nose region.

While both models adhere to a similar framework, Model A boasts additional layers and integrates batch normalization layers. The training process remains consistent across both models, with minor epoch adjustments applied to specific regions of interest (eye and nose).

Algorithm 2 Model B (simple convolutional neural network)
1: Input: Image of size 50x50x3
2: Convolutional layer (32 filters, 3x3 kernel, ReLU activa- tion)
3: Max pooling (2x2 pool size)
4: Dropout (rate 0.3)
5: Convolutional Layer 2:
6: (Replicate structure for Convolutional Layer 2)
7: Convolutional Layer 3:
8:... (Replicate structure for Convolutional Layer 3)
9: Flatten the output of the convolutional layers
10: Dense layer (512 units, ReLU activation)
11: Dropout (rate = 0.3)
12: Dense layer (2 units, softmax activation)

C. The Predicting Component

To determine the authenticity of a video, we employed a majority voting approach by merging the results obtained from three models applied to three different features, which resulted in a total of three individual predictions.

By con side ring the collective opinion of multiple models, our approach aims to enhance the accuracy and robustness of deepfake detection. This comprehensive method considers various aspects and characteristics of the video, increasing Neural Computing and Applications.

V. DISCUSSION

44.1 Overview of the existing deepfake detection techniques Deepfake detection has been a hot topic in the research community, and several techniques have been proposed. Some existing deepfake detection techniques are based on ML algorithms, such as CNNs, RNNs, and autoencoders. These techniques work by extracting features from the deepfake images or videos and comparing them with the features of the original images or videos. However, these techniques have several limitations, such as the need for large training data , the susceptibility to adversarial attacks , and the inability to detect unseen deepfakes.

A. Advantages of the Proposed Methodology

Our paper proposes a novel technique for deepfake detection that combines three models based on different features, including the entire face, eyes, and nose. While this com bination of multiple models only slightly affected overall accuracy, it improves the accuracy of deepfake detection, reducing the impact of weaknesses in a single algorithm. Additionally, we develop a customized data processing stage for each model to detect deepfakes with high reliability. Our proposed technique also benefits from the large amount of data used for training, including datasets like FaceForensics++.

B. Limitations and Future Work

Our proposed technique has certain limitations, such as the need for high-computational resources for training and inference. Additionally, the technique may not be effective in detecting deepfakes that involve changes in parts of the face other than the eyes, nose, and entire face. Future research could focus on developing methods that require less data while maintaining high accuracy rates. We also plan to investigate the use of other features for deepfake detection.

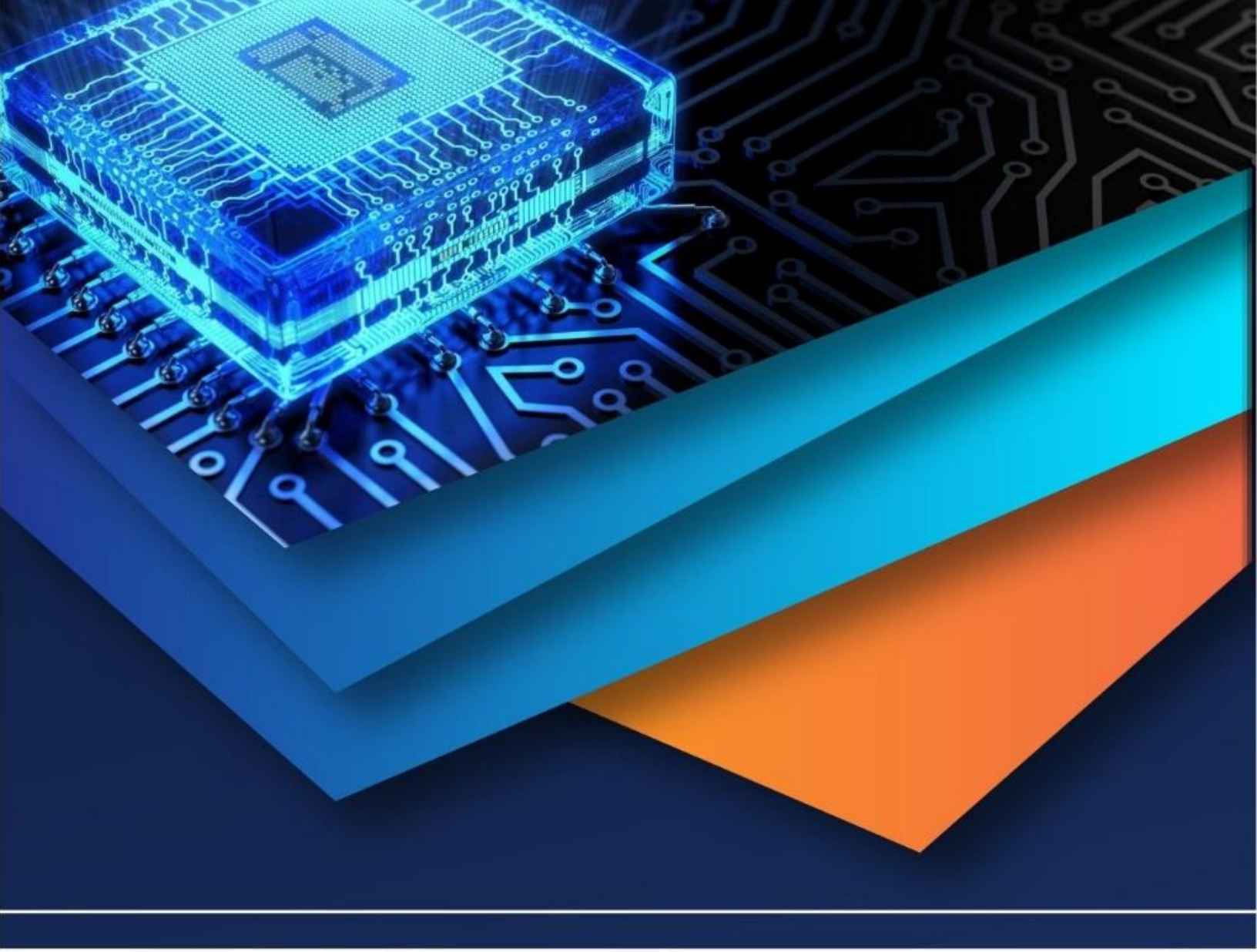
VI. CONCLUSION

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 sub model further strengthen this multifaceted approach, circumventing the constraints typically encountered in single algorithm 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 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 real world 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. The application of our approach has the potential to safeguard individuals, organizations, and society at large from the adverse impacts of deepfakes, thereby contributing significantly to digital security and integrity. Although our results are promising, we recognize the scope for fur there enhancement. Future research could delve into integrating additional facial features or employing alternative datasets, aiming to augment the physical accuracy and operational efficiency of deepfake detection. Such advancements will fortify our method's effectiveness and contribute to the broader field of digital media authenticity.

REFERENCES

[1] FaceApp. Accessed: Jan. 4, 2021. [Online]. Available: https://www.faceapp.com/
[2] FakeApp. Accessed: Jan. 4, 2021. [Online]. Available: https://www.fakeapp.org/
[3] G. Oberoi. Exploring DeepFakes. Accessed: Jan. 4, 2021. [Online].Available: https://goberoi.com/exploring-deepfakes-20c9947c22d9
[4] J. Hui. How Deep Learning Fakes Videos (Deepfake) and How toDetect it. Accessed: Jan. 4, 2021. [Online]. Available: https://medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9
[5] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair,A. Courville, and Y. Bengio, ''Generative adversarial nets,'' in Proc. 27thInt. Conf. Neural Inf. Process. Syst. (NIPS), vol. 2.Cambridge, MA, USA:MIT Press, 2014, pp. 2672–2680.
[6] G. Patrini, F. Cavalli, and H. Ajder, ''The state of deepfakes:Reality under attack,'' Deeptrace B.V., Amsterdam, The Nether-lands, Annu. Rep. v.2.3., 2018. [Online]. Available: https://s3.eu-west-2.amazonaws.com/rep2018/2018-the-state-of-deepfakes.pdfVOLUME 10, 2022
[7] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner,''Face2Face: Real-time face capture and reenactment of RGB videos,''in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas,NV, USA, Jun. 2016, pp. 2387–2395, doi: 10.1109/CVPR.2016.262.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, ‘‘Unpaired image-to-imagetranslation using cycle-consistent adversarial networks,’’ in Proc. IEEEInt. Conf. Comput. Vis. (ICCV), Venice, Oct. 2017, pp.2242–2251, doi:10.1109/ICCV.2017.244.
[9] S. Suwajanakorn, S. M. Seitz, and I. K. Shlizerman, ‘‘SynthesizingObama: Learning lip sync from audio,’’ ACM Trans. Graph., vol. 36,no. 4, p. 95, 2017.
[10] L. Matsakis. Artificial Intelligence is Now Fighting Fake Porn. Accessed:Jan. 4, 2021. [Online]. Available: https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/
[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner,‘‘FaceForensics: A large-scale video dataset for forgery detection inhuman faces,’’ 2018, arXiv:1803.09179.
[12] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner,P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, ‘‘Deep video por-traits,’’ ACM Trans. Graph., vol. 37, no. 4, pp. 1–14, Aug. 2018, doi:10.1145/3197517.3201283.
[13] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, ‘‘Everybody dance now,’’2018, arXiv:1808.07371.
[14] T. Karras, S. Laine, and T. Aila, ‘‘A style-based generator architecturefor generative adversarial networks,’’ in Proc. IEEE/CVF Conf. Com-put. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019,pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.
[15] D. Budgen and P. Brereton, ‘‘Performing systematic literature reviews insoftware engineering,’’ in Proc. 28th Int. Conf. Softw. Eng., New York,NY, USA, May 2006, pp. 1051–1052, doi:10.1145/1134285.1134500.
[16] Z. Stapic, E. G. Lopez, A. G. Cabot, L. M. Ortega, and V. Strahonja, ‘‘Performing systematic literature review in software engineering,’’ in Proc.23rd Central Eur. Conf. Inf. Intell. Syst. (CECIIS), Varazdin, Croatia,Sep. 2012, pp. 441–447.
[17] B. Kitchenham, ‘‘Procedures for performing systematic reviews,’’Softw. Eng. Group; Nat. ICT Aust., Keele; Eversleigh, Keele Univ.,Keele, U.K., Tech. Rep. TR/SE-0401; NICTA Tech. Rep.0400011T.1,2004.
[18] B. Kitchenham and S. Charters, ‘‘Guidelines for performing systematicliterature reviews in software engineering,’’ Softw. Eng. Group; KeeleUniv., Durham University Joint, Durham, U.K., Tech. Rep. EBSE-2007-01, 2007.
[19] M. A. Babar and H. Zhang, ‘‘Systematic literature reviews in softwareengineering: Preliminary results from interviews with researchers,’’ inProc. 3rd Int. Symp. Empirical Softw. Eng. Meas., Lake Buena Vista, FL,USA, Oct. 2009, pp. 346–355, doi: 10.1109/ESEM.2009.5314235.
[20] H. Do, S. Elbaum, and G. Rothermel, ‘‘Supporting controlled experimen-tation with testing techniques: An infrastructure and its potential impact



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY