AI Based DeepFake Image Detection System

Shimpi Sneha Vijay[1], Dungarwal Riya Sunil[2], Sonavane Srushti Rajesh[3],
Pagar Varuna Santosh[4], Zambare Dattu Bhausaheb[5]
[1,2,3,4,5]*Shri H.H.J.B. Polytechnic, Chandwad*

***Abstract*— Deepfake technology has advanced rapidly in recent years, enabling the generation of highly realistic synthetic images that are difficult to distinguish from authentic ones. While such technology has beneficial applications, it also poses serious threats including misinformation, identity theft, and digital forgery. This paper presents a Deepfake Image Detection System based on deep learning techniques to identify manipulated or synthetic facial images. The proposed system leverages convolutional neural networks (CNNs) to automatically learn discriminative features from images, such as texture inconsistencies, facial artifacts, and frequency-domain anomalies. Experimental results demonstrate that the proposed approach achieves high detection accuracy on benchmark deepfake datasets, indicating its effectiveness in combating image-based deepfake threats.**

***Index Terms*— Convolutional Neural Network, Deep Learning, Deepfake Detection, Image Forensics, Synthetic Media.**

I. INTRODUCTION

The rapid development of artificial intelligence and generative models has led to the emergence of deepfakes, which are synthetic media generated using techniques such as Generative Adversarial Networks (GANs) and autoencoders. Deepfake images can convincingly replicate human faces, making it increasingly difficult for humans to differentiate between real and manipulated content. This has raised serious concerns in areas such as social media, digital journalism, legal evidence, and cybersecurity.

Traditional image forensics methods rely on handcrafted features and statistical analysis, which often fail to generalize across different deepfake generation techniques. In contrast, deep learning-based approaches have shown superior performance due to their ability to automatically extract high-level and low-level features from large-scale data. This paper focuses on the design and implementation of a deep learning-based deepfake image detection system that can accurately classify images as real or fake.

II. METHODOLOGY

The The proposed Deepfake Image Detection System consists of four main stages: data collection, preprocessing, feature extraction, and classification.

A. Dataset Collection

The system is trained and evaluated using publicly available deepfake image datasets containing both real and manipulated facial images. These datasets include variations in lighting conditions, facial expressions, poses, and image resolutions, ensuring robustness of the model.

B. Image Preprocessing

Preprocessing steps include face detection, resizing of images to a fixed dimension, normalization of pixel values, and data augmentation techniques such as rotation and flipping. These steps help improve model generalization and reduce overfitting.

C. Feature Extraction Using CNN

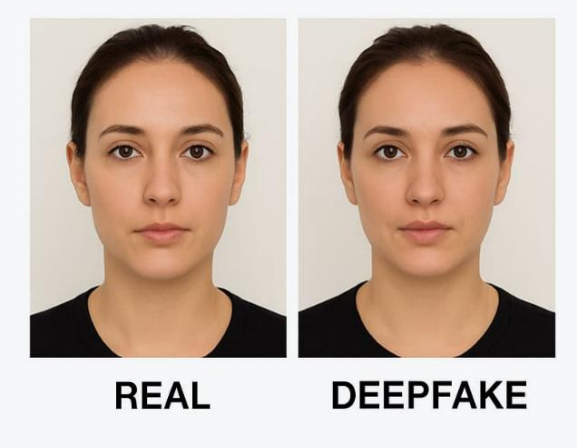
A Convolutional Neural Network (CNN) is employed to extract discriminative features from input images. The CNN automatically learns spatial hierarchies of features, capturing subtle artifacts introduced during deepfake generation, such as unnatural skin textures and blending errors.

D. Classification

The extracted features are passed through fully connected layers followed by a softmax activation function to classify the image as either real or deepfake. Binary cross-entropy loss is used during training, and the model is optimized using the Adam optimizer.

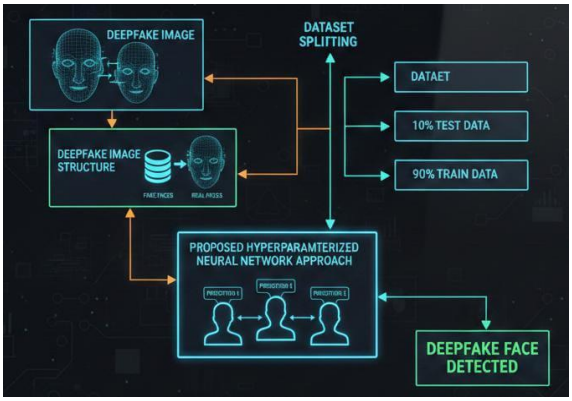
III. OBJECTIVES

The objective of this project is to develop a dependable and efficient deepfake image detection system using deep learning approaches. The work aims to understand the nature of deepfake image generation and the challenges associated with identifying manipulated facial content. By designing and implementing a Convolutional Neural Network (CNN), the system seeks to automatically learn meaningful facial features and artifacts that arise due to image manipulation. Another key objective is to apply suitable image preprocessing and enhancement techniques so that the input data is consistent and supports effective model training. The proposed system focuses on accurately classifying facial images as either genuine or fake while minimizing incorrect predictions. In addition, the project aims to evaluate the performance of the detection model using standard metrics such as accuracy, precision, recall, and F1-score. Overall, the objective is to build a scalable and adaptable framework that can support real-world applications such as digital media verification, content authentication, and misinformation control.



IV. SYSTEM ARCHITECTURE

The system architecture comprises an input image layer, multiple convolution and pooling layers for feature learning, dense layers for classification, and an output layer indicating the probability of an image being fake or authentic. The architecture is designed to balance detection accuracy and computational efficiency, making it suitable for real-time applications.



V. PERFORMANCE EVALUATION

The performance of the proposed system is evaluated using metrics such as accuracy, precision, recall, and F1-score. Experimental results show that the CNN-based model achieves high accuracy in detecting deepfake images and outperforms traditional machine learning approaches. The system also demonstrates robustness against different deepfake generation techniques.

VI. CONCLUSION

This paper presented a deep learning-based Deepfake Image Detection System capable of accurately identifying manipulated facial images. By leveraging convolutional neural networks, the proposed approach effectively captures subtle artifacts introduced during image manipulation. Future work may include extending the system to video-based deepfake detection and improving robustness against emerging generative models.

VII. APPLICATION

The Deepfake Image Detection System has wide-ranging applications across multiple domains where the authenticity of visual content is critical. In social media platforms, the system can be used to automatically identify and flag manipulated images before they are widely shared, thereby reducing the spread of misinformation and fake news. In the field of digital forensics, it can assist law enforcement agencies and investigators in verifying the originality of image-based evidence used in legal and criminal investigations. The system is also applicable in identity verification and biometric security systems,

where detecting forged facial images is essential to prevent impersonation and fraud. In journalism and media organizations, the proposed approach can support content verification workflows by ensuring that published images are genuine and trustworthy. Additionally, the system can be integrated into cybersecurity frameworks to protect individuals and organizations from image-based social engineering attacks. Overall, the proposed deepfake detection system contributes to enhancing digital trust, media authenticity, and information security in real-world environments.

REFERENCE

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images" 2019 IEEE/CVF International Conference on Computer Vision (ICCV).

[2] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, Baining Guo, "Face X-Ray for More General Face Forgery Detection"2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, Isao Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network" 2018 IEEE International Workshop on Information Forensics and Security (WIFS).

[4] Ricard Durall,Margret Keuper, Franz-Josef Pfreundt, Janis Keuper, "Unmasking DeepFakes with simple Features",Fraunhofer ITWM, Germany,IWR, University of Heidelberg, Germany,Fraunhofer Center Machine Learning, Germany,Data and Web Science Group, University Mannheim, Germany.

[5] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, Peyman Najafirad, Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models.

[6] Md Shohel Rana, Mohammad Nur Nobi,Beddhu Murali,Andrew H. Sung, Deepfake Detection: A Systematic Literature Review IEEE Access 10:1-1.

[7] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In IEEE Workshop on Information Forensics and Security, WIFS 2017, Rennes, France, December 2017.

[8] Y. Rao and J. Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. In Information Forensics and Security (WIFS), 2016 IEEE International Workshop on, pages 1–6. IEEE, 2016. [12] J. A. Redi, W. Taktak, and J.-L. Dugelay. Digital image forensics: a booklet for beginners. Multimedia Tools and Applications, 51(1):133–162, 2011.

[9] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, ¨ and M. Nießner. Faceforensics: A large-scale video dataset for

[10] forgery detection in human faces. arXiv preprint arXiv:1803.09179, 2018.