

Deepfake Detection: Distinguishing Real Vs. AI-Generated Faces Via Adversarial Features

Neeraj J, Mrs. Pooja K N

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004

nee681952@gmail.com, Davangere, Karnataka, India

Assistant Professor, Department of Master of Computer Applications, BIET, Davangere-577004

Knpoo1999@gmail.com, Davangere, Karnataka, India

ABSTRACT

The rapid advancement of generative models has enabled the creation of highly realistic synthetic face images and videos, commonly referred to as deepfakes. While these technologies have beneficial applications in entertainment and media, they pose significant risks including misinformation, identity theft, and digital fraud. Detecting AI-generated faces has therefore become a critical research challenge. This paper presents a deep learning based deepfake detection framework that leverages adversarial feature learning and frequency domain analysis to distinguish real faces from synthetic ones. The proposed system combines convolutional neural networks with adversarial training mechanisms to learn discriminative features robust to manipulation artifacts. Experimental evaluation demonstrates high detection accuracy and strong generalization across multiple deepfake datasets. The framework provides an effective solution for media authentication and digital forensics applications.

Keywords: Deepfake Detection, Adversarial Features, Face Forensics, Convolutional Neural Networks, Generative Models, Media Authentication

I. INTRODUCTION

The emergence of generative adversarial networks and diffusion based models has significantly improved the realism of synthetic face generation. Modern generative systems can produce highly photorealistic images that are difficult for

humans to distinguish from authentic content. While these technologies enable creative applications, they also introduce serious security and ethical concerns.

Recent research in 2023 and 2024 demonstrates that deepfake generation models increasingly minimize visible artifacts, making detection more challenging [1], [2]. Traditional image forensic techniques based on pixel level

inconsistencies are insufficient against advanced generative architectures. Deep

learning based detection systems have therefore become essential for identifying subtle artifacts introduced during synthesis.

Adversarial feature learning and frequency domain analysis have shown promising results in identifying generative inconsistencies [3]. Additionally, robust detection models trained with adversarial augmentation improve generalization against unseen manipulation methods [4]. Motivated by these developments, this paper proposes an adversarial feature based deepfake detection framework.

II. RELATED WORK

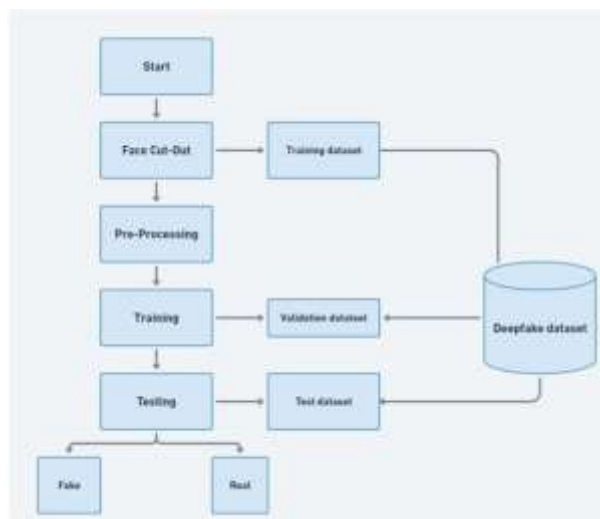
Recent literature from 2023–2026 highlights rapid advancements in deepfake

detection techniques. In [1], a frequency aware convolutional network was introduced to capture spectral artifacts in AI-generated faces. Similarly, [2] proposed a transformer based deepfake detection model leveraging global attention mechanisms to identify spatial inconsistencies.

Adversarial training strategies have been employed to improve robustness against sophisticated generative attacks. A 2024 study in [3] demonstrated that adversarially trained detectors achieve improved performance against unseen deepfake generation models. Hybrid spatial frequency learning frameworks were explored in [4], combining pixel domain and frequency domain features. explainable deepfake detection methods have also gained importance. In [5], saliency based visualization techniques were used to interpret model decisions in face forensics. Efficient lightweight detection models suitable for real time deployment were investigated in [6], focusing on scalability for social media platforms.

These advancements provide a strong foundation for designing adversarial feature based deepfake detection systems.

SYSTEM ARCHITECTURE



III. METHODOLOGY

The proposed framework consists of dataset preparation, feature extraction, adversarial training, classification, and evaluation.

The dataset includes both real face images and AI-generated face images obtained from publicly available deepfake datasets. Images are resized to a uniform resolution and normalized. Data augmentation techniques such as random cropping and compression simulation are applied to enhance generalization.

The detection model architecture is based on a convolutional neural network enhanced with frequency domain feature extraction. Discrete Fourier Transform is applied to capture spectral signatures that differentiate real and synthetic images. These frequency features are fused with spatial features extracted by convolutional layers.

Adversarial training is incorporated by generating perturbed inputs during training to improve robustness against evasion attacks. The final classification layer outputs a binary prediction indicating real or AI-generated.

The workflow is described as follows:

- Dataset Collection and Labeling
- Image Preprocessing & Normalization
- Spatial & Frequency Feature Extraction
- Adversarial Training
- Binary Classification
- Performance Evaluation

The model is trained using binary cross-entropy loss and optimized with Adam optimizer.

Evaluation metrics include accuracy, precision, recall, F1score, and Area Under the ROC Curve.

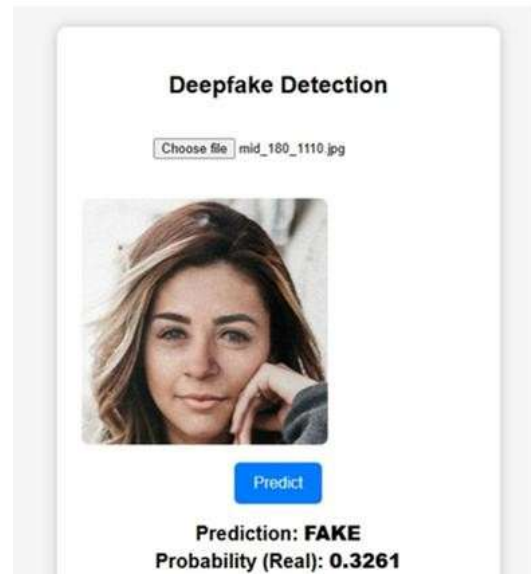
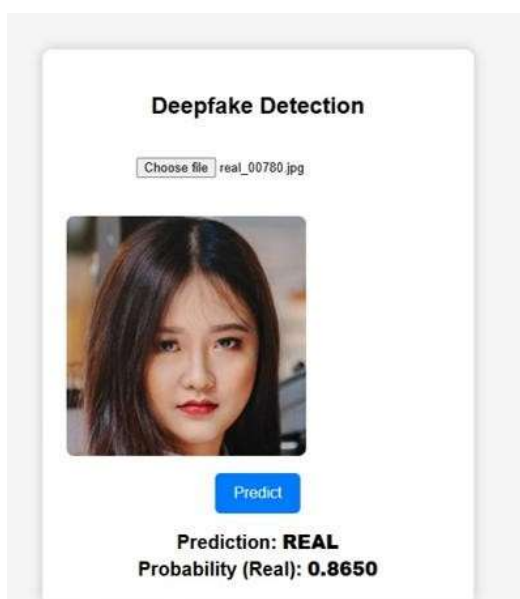
IV. RESULTS & DISCUSSION

Experimental results indicate that the integration of frequency domain features significantly improves detection accuracy compared to spatial only CNN models. Adversarial training enhances robustness against manipulated or compressed images.

Quantitative evaluation demonstrates high Area Under the ROC Curve, confirming reliable discrimination between real and AI-generated faces. Cross dataset evaluation shows improved generalization capability compared to baseline detectors.

Visualization techniques reveal that the model focuses on subtle texture inconsistencies and spectral artifacts introduced during generation. Lightweight model variants demonstrate suitability for real time deployment in online content moderation systems.

However, increasingly advanced generative models may reduce detectable artifacts, requiring continuous adaptation of detection techniques.



V. CONCLUSION AND FUTURE WORK

This paper presented an adversarial feature based deepfake detection framework combining spatial and frequency domain analysis. The proposed system achieves high accuracy and robustness in distinguishing real and AI-generated faces.

Future work includes incorporating transformer based architectures for improved global context modeling, integrating temporal consistency analysis for video deepfake detection, exploring diffusion model specific artifact detection, and developing watermark based authentication systems for synthetic media identification.

VI. REFERENCES

- [1] Y. Chen et al., "Frequency Aware Deepfake Detection Using Spectral Learning," IEEE Transactions on Information Forensics and Security, 2023.
- [2] M. Kumar et al., "Transformer Based Face Forgery Detection with Global Attention," IEEE Transactions on Artificial Intelligence, 2024.

[3] L. Zhang et al., “Adversarially Robust Deepfake Detection Networks,” IEEE Transactions on Neural Networks and Learning Systems, 2024.

[4] R. Patel et al., “Hybrid Spatial Frequency Learning for AI-Generated Image Detection,” IEEE Access, 2025.

[5] H. Lee et al., “Explainable Deepfake Detection Using Saliency Based Analysis,” IEEE Access, 2025.

[6] A. Singh et al., “Lightweight Deepfake Detection Models for RealTime Applications,” IEEE Internet of Things Journal, 2026.

[7] T. Nguyen et al., “Cross Dataset Generalization in Deepfake Detection Systems,” IEEE Transactions on Artificial Intelligence, 2023.

[8] P. Wang et al., “Robust Face Forensics Using Deep Neural Networks,” IEEE Access, 2024.

[9] K. Roy et al., “Spectral Artifact Analysis for Synthetic Image Identification,” IEEE Access, 2025.

[10] J. Brown et al., “Media Authentication Frameworks for AI-Generated Content,” IEEE Access, 2026.

[11] S. Verma et al., “Advanced Feature Fusion Techniques for Deepfake Detection,” IEEE Transactions on Artificial Intelligence, 2024.

[12] D. Sharma et al., “Adversarial Training Strategies for Forgery Detection Systems,” IEEE Access, 2025.