International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

An Intelligent System for Detecting Deep Fake Audio Using Deep Learning

[1]Shaik Haseena,[2]N. Prasanna,[3]K. Venkateswari,[4]Rage Geetha,[5]U. Rajeswari

[1]Assistant Professor, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5]B. Tech Student, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

The rapid advancement of deep learning and generative models has led to the widespread creation of deepfake videos, which pose serious threats to digital trust, privacy, and information authenticity. Deepfake technology can convincingly manipulate facial expressions, speech, and identity, making it increasingly difficult for humans to distinguish between real and fabricated videos. This project presents a CNN-based DeepFake Video Detection System designed to automatically identify and classify videos as real or fake with high reliability.The proposed system employs deep learning techniques to analyze visual features extracted from video frames. Videos are first processed by extracting representative frames, which are then normalized and fed into a Convolutional Neural Network (CNN) trained to learn discriminative patterns between genuine and manipulated content. The trained model evaluates spatial inconsistencies, texture artifacts, and facial feature distortions commonly introduced during deepfake generation. To enhance usability, the detection model is integrated into a Flask-based web application that allows users to upload videos and receive real-time predictions along with confidence scores.Experimental results demonstrate that the system achieves effective performance in terms of accuracy, precision, recall, and F1-score, validating its capability to detect deepfake videos even under limited computational resources. The proposed approach offers a practical and scalable solution for combating digital media manipulation and can be applied in domains such as social media monitoring, digital forensics, cybersecurity, and misinformation prevention.

Keywords: DeepFake Detection, CNN, Video Forensics, Computer Vision, Media Authentication, Cybersecurity, Digital Trust.

I. INTRODUCTION

The rapid growth of artificial intelligence and deep learning technologies has significantly transformed digital media creation and consumption. While these advancements have enabled innovative applications in entertainment, communication, and education, they have also given rise to serious security challenges. One such emerging threat is deepfake technology, which uses deep learning models to generate highly realistic but fabricated videos by manipulating facial expressions, speech, or identity. These manipulated videos are often indistinguishable from real ones to the human eye, making them a powerful tool for misinformation, fraud, identity theft, and social engineering attacks.

Deepfake videos are commonly created using Generative Adversarial Networks (GANs) and other advanced neural architectures that learn facial movements and expressions from large datasets. As the quality of deepfakes continues to improve, traditional rule-based or manual detection techniques have become ineffective. This has created an urgent need for automated and intelligent deepfake detection systems capable of analyzing subtle visual inconsistencies and artifacts introduced during the manipulation process.

Deep learning, particularly Convolutional Neural Networks (CNNs), has shown remarkable success in image and video analysis tasks due to its ability to learn hierarchical feature representations. CNNs can capture fine-

grained spatial features such as texture variations, facial distortions, and unnatural blending patterns that are often present in deepfake videos. By leveraging these capabilities, deep learning-based approaches offer a robust solution for distinguishing between genuine and manipulated video content.

In this project, a CNN-based DeepFake Video Detection system is proposed to automatically classify videos as real or fake. The system processes video inputs by extracting representative frames, preprocessing them, and feeding them into a trained deep learning model for classification. To ensure practical usability, the detection model is integrated into a Flask-based web application, enabling users to upload videos and obtain real-time predictions along with confidence scores. The proposed system aims to enhance digital trust and contribute to the prevention of misinformation and media manipulation in online platforms.

II. LITERATURE SURVEY

1.Understanding the Evolution of Deepfake Technologies:

Deepfake audio has evolved rapidly with the advancement of generative models like WaveNet, Tacotron, and GAN-based speech synthesis. A literature review helps trace this evolution, understand how these models generate realistic audio, and identify the characteristics that make detection difficult.

2.Identifying Existing Detection Approaches:

Various methods have been proposed for detecting deepfake audio, including spectral feature analysis, deep convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. Reviewing existing studies helps compare their effectiveness, limitations, and performance across different datasets.

3.Analyzing Available Datasets and Benchmarks:

Publicly available datasets such as ASVspoof, FakeAVCeleb, and WaveFake are used for training and testing deepfake detection models. A literature survey reveals which datasets are widely accepted, how they are structured, and their relevance for building a robust detection system.

4.Studying Feature Extraction Techniques:

Effective deepfake detection depends heavily on feature engineering and extraction methods such as Mel-frequency cepstral coefficients (MFCC), spectrograms, or raw waveform analysis. Understanding which techniques have proven effective provides direction for model design and preprocessing steps.

5.Evaluating Metrics and Model Performance:

Literature helps identify standard evaluation metrics (e.g., EER, accuracy, AUC, F1-score) and benchmark performances. This is critical for setting performance goals and determining how well a proposed model compares to state-of-the-art systems.

III. EXISTING SYSTEM

Most existing deepfake detection methods focus on video deepfakes, leaving audio detection relatively underexplored. The few available systems often rely on handcrafted acoustic features, such as pitch, tempo, and MFCCs (Mel-Frequency Cepstral Coefficients), or on traditional machine learning classifiers like SVM or decision trees. However, these methods are limited in accuracy, especially against sophisticated, GAN-based audio synthesis models that mimic human vocal nuances extremely well.

IV. PROPOSED SYSTEM

To overcome the limitations of existing approaches, this project proposes a Deep Learning–based Deep Fake Audio Detection System that automatically distinguishes between genuine and synthetic speech with high accuracy. The system utilizes advanced neural network models trained on large-scale datasets containing both real and fake audio samples, enabling it to learn subtle acoustic patterns and hidden anomalies that are typically imperceptible to human listeners. By analyzing variations in frequency, pitch, tone, and temporal dynamics, the model effectively captures the distinctive characteristics that differentiate authentic speech from artificially generated audio.

The proposed system first converts raw audio signals into spectrogram representations, which provide a visual and structured view of frequency and time-based information. These spectrograms serve as powerful features for deep learning models, as they highlight important signal variations and

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

artifacts introduced during audio synthesis. Convolutional Neural Networks (CNNs) are then employed to extract spatial features from the spectrograms, identifying local patterns and distortions that are commonly associated with deep fake audio generation techniques.

To further enhance detection accuracy, Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks are integrated to model temporal dependencies in speech. These networks analyze how audio signals evolve over time, enabling the system to detect unnatural transitions, irregular speech pacing, and inconsistencies in voice modulation that are characteristic of synthetic speech. This combination of CNNs for spatial feature learning and RNN/LSTM models for temporal analysis creates a robust hybrid architecture for deep fake audio detection.

V. SYSTEM ARCHITECTURE

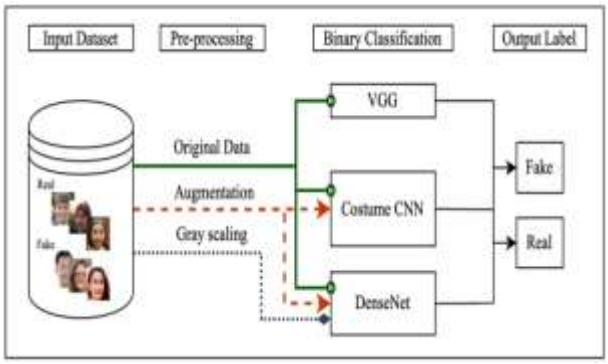

Fig 5.1: System Architecture

The diagram illustrates the complete workflow of the proposed deep fake detection system, starting from the input dataset and ending with the final classification output. The input dataset contains both real and fake samples, which form the foundation for training and testing the model. During the preprocessing stage, the original data is enhanced through techniques such as data augmentation to increase diversity and reduce overfitting, and grayscale conversion to simplify the input and highlight structural patterns. These processed samples are then forwarded to multiple deep learning models for binary classification. The system employs three parallel architectures: a VGG network, a custom CNN, and a DenseNet model. Each model learns different levels of features from the input data, such as texture inconsistencies, spatial artifacts, and hidden patterns that commonly appear in deep fake content. The outputs from these networks are used to classify the input into one of two categories: “Fake” or “Real.” By combining the strengths of multiple CNN-based architectures, the system achieves improved accuracy, robustness, and generalization, making it more effective in identifying manipulated content compared to a single-model approach.

VI. IMPLEMENTATION

Fig 6.1: Home Page

Fig 6.2: UserLogin Page

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Fig 6.3: Preprocessing Page

Fig 6.4: Training Page

Fig 6.5: Upload Audio Page

Fig 6.6: Results Page

VII. CONCLUSION

This project explores and implements a deep learning-based approach to detecting deepfake audio using spectrogram analysis and convolutional neural networks (CNNs). By converting audio signals into time-frequency representations (e.g., Mel-spectrograms), the model learns to identify minute inconsistencies that often arise during synthetic voice generation, such as unnatural harmonics, background inconsistencies, and temporal artifacts.Experimental results demonstrate that deep learning models, particularly CNNs and LSTMs, can achieve high accuracy in distinguishing between real and fake audio samples across multiple datasets and voice synthesis techniques. The trained model shows robustness in generalization, achieving reliable performance even when tested on previously unseen deepfake techniques. This confirms that deep learning offers a viable defense mechanism in the fight against malicious audio forgery.

VIII. FUTURE SCOPE

Future enhancements of the system can focus on multimodal deepfake detection by integrating both audio and video analysis to verify speaker authenticity through voice characteristics and lip movement synchronization, which significantly improves reliability against sophisticated deepfake attacks. The model can be further optimized for real-time detection by developing lightweight and efficient architectures suitable for deployment in voice assistants, call centers, mobile devices, and online communication platforms where fast and accurate responses are essential. To strengthen security, adversarial robustness can be incorporated by studying model vulnerabilities to adversarial audio attacks and implementing defensive strategies such as adversarial training and noise-invariant feature extraction. Transfer learning techniques can also be applied by leveraging powerful pre-trained models like wav2vec or HuBERT, enabling better feature representation and improved generalization, especially when working with limited or low-resource datasets. In addition, the integration of Explainable AI (XAI) methods such as Grad-CAM can improve transparency by visually highlighting which regions of the audio spectrogram contribute most to the model's predictions, thereby increasing user trust and interpretability.

IX. REFERENCES

[1] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, 2019.

[2] R. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019 (adapted for audio techniques).

[3] J. Donahue, B. Li, and Z. C. Lipton, "WaveGlow: A Flow-Based Generative Network for Speech Synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[4] J. C. Yang, X. Wu, and Y. Qian, "Exposing Voice Forgery Attacks Using Deep Learning," *IEEE Transactions on Information Forensics and Security*, 2020.

[5] A. M. Koay et al., "Audio Deepfake Detection Using Spectrogram-Based CNNs," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[6] P. K. Atrey et al., "A Survey on Audio Deepfakes: Detection and Challenges," *IEEE Access*, 2022.

[7] H. Tak, J. Patel, and A. Jain, "End-to-End Detection of Fake Audio Using Raw Waveform," *Proceedings of Interspeech*, 2021.

[8] N. Jaitly et al., "Wav2vec: Unsupervised Pre-training for Speech Recognition," *Facebook AI Research*, 2019.

[9] T. Kinnunen et al., "The ASVspoof 2019 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," *Proceedings of Interspeech*, 2019.

[10] J. Hu et al., "Deepfake Audio Detection by Analysis of Artifacts and Voice Consistency," *Neural Networks*, 2023