



(RESEARCH ARTICLE)



Multi Modal DeepFake Detection Using CNN

Saritha Banoth, Bhavana Chandragiri *, Priyamvadha Ramadugala, Harshavardhan Oraganti and Jayanth Konapakula

Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 622-630

Publication history: Received on 25 March 2025; revised on 02 May 2025; accepted on 04 May 2025

Article DOI: <https://doi.org/10.30574/wjaets.2025.15.2.0564>

Abstract

The proliferation of deepfake content has raised serious concerns about the authenticity of digital media, with implications spanning from social misinformation to cybersecurity breaches. In this study, we propose a multi-model deepfake detection framework that integrates multiple convolutional neural networks (CNNs) to improve classification accuracy and robustness. Each model within the ensemble is trained to detect unique facial distortions and inconsistencies introduced by deepfake generation techniques. By leveraging the strengths of diverse architectures, the system effectively identifies manipulated media across varying formats and qualities. Experimental results on publicly available datasets demonstrate that the proposed multi-model approach outperforms single-model baselines in both precision and generalization. This work contributes to the growing field of AI-based media forensics by offering a scalable and effective solution to combat the evolving challenge of deepfakes.

Keywords: Deepfake Detection; Multi Modal Ensemble; Convolutional Neural Networks; Media Forensics; Digital Security; Fake Media Identification; Deep Learning

1. Introduction

Traditional deepfake detection methods often rely on a single model to identify visual or audio inconsistencies. However, as deepfake techniques become more advanced and harder to spot, these single-model systems tend to fall short—especially when dealing with new types of forgeries or variations in lighting, expressions, and video quality. To address this challenge, our research introduces a multi-model approach that combines the strengths of several deep learning models, particularly convolutional neural networks (CNNs). These models are trained to identify different types of tampering, such as distortions, blending errors, or unnatural movements. By working together, these models create a more accurate and dependable system. This approach offers a practical, scalable solution to help detect deepfakes more effectively in real-world applications like media, security, and digital forensics.

In addition to leveraging multiple deep learning models, our approach incorporates diverse input modalities such as frame-level images, motion cues, and audio signals. Each modality contributes unique information that enhances the system's ability to spot deepfakes under varied conditions. For instance, subtle audio-visual mismatches or inconsistencies in facial dynamics across frames might go undetected by a single model, but can be effectively captured through this integrated framework. This fusion not only improves detection accuracy but also increases the system's robustness to manipulation techniques that aim to bypass traditional detectors. Ultimately, this method lays the groundwork for building adaptive and resilient deepfake detection tools that keep pace with rapidly evolving synthetic media technologies.

* Corresponding author: Bhavana Chandragiri

1.1. Problem Statement

The increasing sophistication of deepfake technology has introduced a major challenge in maintaining the integrity of digital media. Powered by deep learning techniques such as GANs (Generative Adversarial Networks), deepfakes can convincingly alter faces, voices, and entire video scenes to the point where it becomes difficult for the human eye—or even traditional software—to distinguish between real and fake content. While the technology itself has creative potential in entertainment and education, its misuse poses serious risks to public trust, personal privacy, and social stability.

One of the core issues with current deepfake detection systems is their limited ability to generalize. Many rely on single-model architectures trained on specific datasets, which often fail to perform well when tested on deepfakes generated by newer, more advanced algorithms. Additionally, variations in lighting, resolution, facial expressions, and compression levels further degrade the accuracy of detection. This creates a pressing need for more adaptable and resilient detection frameworks.

Another challenge is the real-time application of deepfake detection in practical scenarios—such as live video streams, social media monitoring, or legal investigations—where accuracy and speed are both crucial. An effective solution must not only detect forgeries but also adapt to the evolving landscape of generative techniques.

This paper addresses these issues by proposing a multi-model deepfake detection system that leverages multiple convolutional neural networks (CNNs), each trained to focus on different artifacts introduced during deepfake generation. By combining these models, the system aims to improve detection accuracy, reduce false positives, and enhance generalizability across a wider range of inputs. The ultimate goal is to provide a reliable, scalable solution that can support organizations, platforms, and individuals in identifying manipulated content and preserving digital truth.

2. Literature review

Early approaches to deepfake detection primarily focused on visual artifacts and inconsistencies in facial features using single-stream convolutional neural networks (CNNs). Methods such as MesoNet and XceptionNet achieved notable success by detecting anomalies in compression artifacts or facial textures. However, these models often struggle when deepfakes are generated using advanced techniques that minimize visible distortions. Other approaches have explored frequency-based analysis or used temporal inconsistencies across video frames to improve detection accuracy. Despite these innovations, models that rely on a single modality—such as image frames alone—often lack the robustness required for generalization across diverse datasets and real-world scenarios.

To address these limitations, recent studies have investigated multi-modal and ensemble-based deepfake detection systems. Some researchers have proposed combining spatial, temporal, and audio features to identify cross-modal inconsistencies that deepfakes fail to reproduce accurately. For instance, models integrating lip-sync analysis with audio spectrogram classification have shown improved performance against audio-visual forgeries. Additionally, ensemble learning techniques that aggregate predictions from multiple models have proven effective in enhancing generalization and reducing false positives. These developments highlight a growing trend toward hybrid architectures that leverage complementary strengths of various detection strategies, setting the stage for more resilient and scalable solutions to combat synthetic media.

3. Existing System

Initial research on deepfake detection focused on identifying superficial artifacts introduced by early-generation models. These included irregularities in eye blinking, inconsistent facial expressions, or unnatural head movements. Li et al. (2018) demonstrated that early deepfakes often failed to mimic natural eye-blinking patterns, making it a key cue for detection. These rule-based or handcrafted approaches were easy to implement but lacked adaptability, quickly becoming obsolete as generative techniques matured and overcame such flaws.

With the evolution of deepfake techniques, researchers began adopting deep learning methods—particularly Convolutional Neural Networks (CNNs)—for more sophisticated detection. Architectures such as XceptionNet and

EfficientNet showed high accuracy on benchmark datasets like FaceForensics++, Celeb-DF, and DFDC. These models learned to detect visual artifacts such as texture mismatches, boundary inconsistencies, and unnatural facial blending. However, they often suffered from poor generalization, performing well on the datasets they were trained on but struggling with unseen formats, compression levels, or real-world data.

To address the limitations of single-model systems, recent work has focused on ensemble and multi-model approaches. These systems combine the strengths of different models, each trained to detect distinct forgery cues—such as spatial artifacts, temporal anomalies, or frequency-domain inconsistencies. Research shows that ensemble methods improve detection robustness and adaptability across datasets. However, standardized integration methods and real-time deployment capabilities are still developing. Our proposed work builds upon this concept, integrating multiple CNNs into a unified framework to enhance accuracy and generalizability in detecting deepfake content.

4. Proposed System

This project introduces a comprehensive multi-modal deepfake detection system capable of identifying manipulated images, videos, and audio using an ensemble of deep learning models. To ensure a scalable and reliable framework, the system is organized into key functional components.

4.1. Core Components of the System

The proposed framework is composed of three primary modules: the Image Analysis Module, Video Forensics Module, and Audio Authentication Module. Each module is tailored to the specific modality and employs specialized deep learning architectures to identify forgery cues.

- The Image Analysis Module uses convolutional neural networks (CNNs) to detect inconsistencies in facial regions, skin textures, lighting artifacts, and unnatural edges that often occur in fake images.
- The Video Forensics Module combines CNNs with temporal models such as LSTMs or 3D CNNs to analyze frame sequences. This allows the detection of frame-level anomalies, unnatural movements, or temporal lags that indicate manipulation.
- The Audio Authentication Module focuses on detecting tampered or AI-generated speech using spectrogram analysis and recurrent neural networks (RNNs). It identifies patterns like pitch irregularities or phase distortions that distinguish real from synthetic audio.

4.1.1. Integration and Decision Layer

Each module outputs its individual prediction, which is then fed into an ensemble-based decision layer. This layer uses voting or weighted averaging to produce a final verdict on whether the input content is authentic or manipulated. This multi-model, multi-modal approach enhances overall accuracy and adaptability to evolving deepfake generation techniques.

5. Methodology

The proposed deepfake detection system follows a structured multi-stage pipeline. It begins with data acquisition, where datasets containing both authentic and manipulated media across various modalities (image, video, and audio) are collected. The next phase is preprocessing, involving operations such as frame extraction from videos, resizing, noise reduction, normalization, and audio spectrogram generation to ensure uniform input formats across all models. Following this, the model architecture is defined, tailored to process different types of inputs through deep learning networks.

At the core of the system are modality-specific CNN models, each trained to detect inconsistencies within a particular type of data—visual frames, motion patterns, or audio signals. Depending on the configuration, a feature fusion stage may be applied to integrate the outputs of these individual models, combining their extracted features into a unified representation. The fused or individual outputs are then passed to a classification and decision layer, which determines the likelihood of an input being real or fake. Finally, evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix are used to assess the performance and robustness of the detection system across different test scenarios.

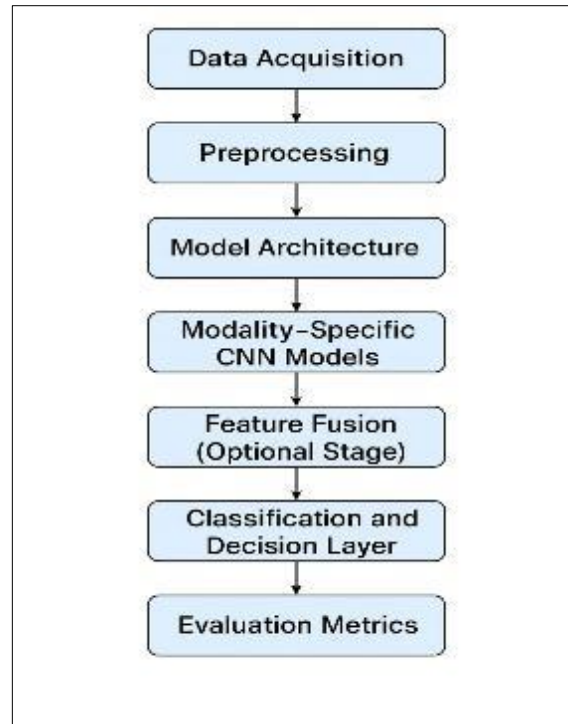


Figure 1 Methodology

5.1. System Architecture

The system begins with an Input Data Stream, which may include audio, video, or standalone image samples. Depending on the input type, it is directed to one of three specialized processing modules: the Image Module, Audio Module, or Video Module.

- The Image Module utilizes convolutional neural networks like InceptionV3 and Xception, which are designed for image classification tasks. These networks extract spatial features from individual frames or image inputs to identify manipulations or inconsistencies.
- The Audio Module processes sound inputs using Mel-Frequency Cepstral Coefficients (MFCCs). These features represent the short-term power spectrum of audio signals and are effective for detecting subtle anomalies in speech patterns or background sounds.
- The Video Module employs a ResNet-50 3D CNN, which captures both spatial and temporal features from video data. This allows the model to detect unnatural facial movements or inconsistencies across consecutive frames.

Once each module extracts relevant features, they are passed to a Feature Fusion Layer, which combines information from all modalities into a unified representation. This fusion allows the model to make more informed and accurate predictions by leveraging complementary strengths of audio, image, and video analyses.

To enhance interpretability, the system integrates an Explainable AI (XAI) component after fusion. This stage provides transparency by explaining the reasoning behind the model's predictions, which is essential for trust in high-stakes applications like security or digital forensics.

Finally, the system delivers a Deepfake Prediction, classifying the input as either real or fake based on the fused and explained model output.

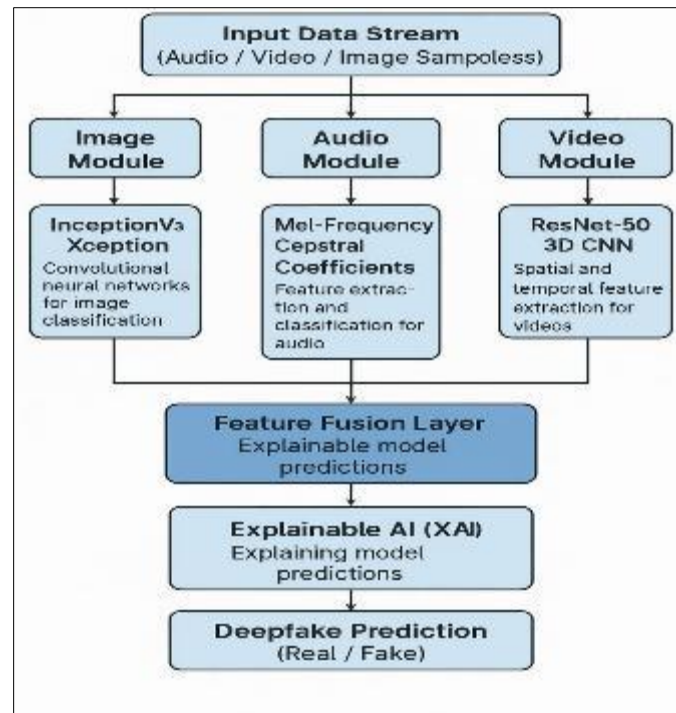


Figure 2 System Architecture

5.1.1. Input Acquisition Layer

This is the first stage of the pipeline, where the system receives various data streams. It supports:

- Images (JPG, PNG, etc.)
- Audio clips (WAV, MP3)
- Videos (MP4 or frame-extracted video input)

Based on the modality, each input is routed to its corresponding processing module (image, audio, or video).

5.1.2. Preprocessing Layer

Each input type undergoes dedicated preprocessing to ensure quality and consistency:

- Image Preprocessing: Frames are resized, normalized, and optionally augmented for noise or compression.
- Audio Preprocessing: Audio is converted into spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs).
- Video Preprocessing: Frames are extracted from video at a fixed frame rate, then resized and normalized for analysis.

5.1.3. Modality-Specific Feature Extraction

Deep learning models specific to each modality are used to extract key features:

- Image Module: Uses CNNs like InceptionV3 or Xception to detect manipulation in facial features or textures.
- Audio Module: Extracts MFCCs and classifies using audio CNNs or RNN-based layers.
- Video Module: Uses 3D CNNs such as ResNet-50 3D to capture temporal inconsistencies and facial distortions across frames.

5.1.4. Feature Fusion Layer

This optional but powerful stage combines features from multiple modalities:

- Fusion Strategy: Concatenation or attention-based fusion is applied to generate a joint representation.
- Purpose: To capture cross-modal relationships and improve the robustness of the final prediction.

5.1.5. Classification and Decision Layer

The fused or individual features are passed through:

- Fully Connected Layers: Followed by softmax or sigmoid activation.
- Output: Binary classification (Real or Fake) with an associated confidence score.

5.1.6. Explainability Layer (XAI)

This layer improves model transparency:

- Grad-CAM for visual explanations on image/video frames.
- Saliency Maps or Feature Attribution for audio data.
- Helps users understand the reasoning behind predictions, increasing trust in the system.

5.1.7. User Interface Layer

Developed using Streamlit for real-time interaction:

- Input Tabs: Users can upload image, video, or audio files.
- Live Results: Predictions and explanations are displayed with visualization.
- Session Support: Users can explore multiple files in a single session.

5.1.8. Evaluation and Testing

Extensive testing was conducted to ensure reliability:

- Metrics Used: Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
- Real-World Testing: Tested with deepfakes of varied lighting, background, and resolutions.
- Cross-Modality Evaluation: Each modality tested independently and in combination.

6. Results and Discussion

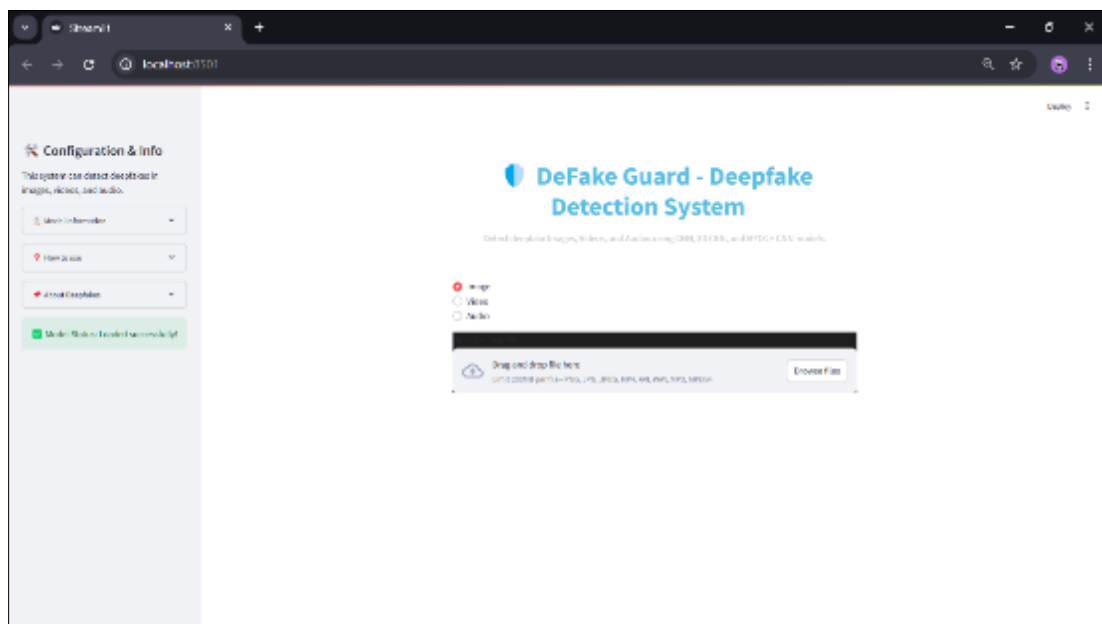


Figure 3 User Interface

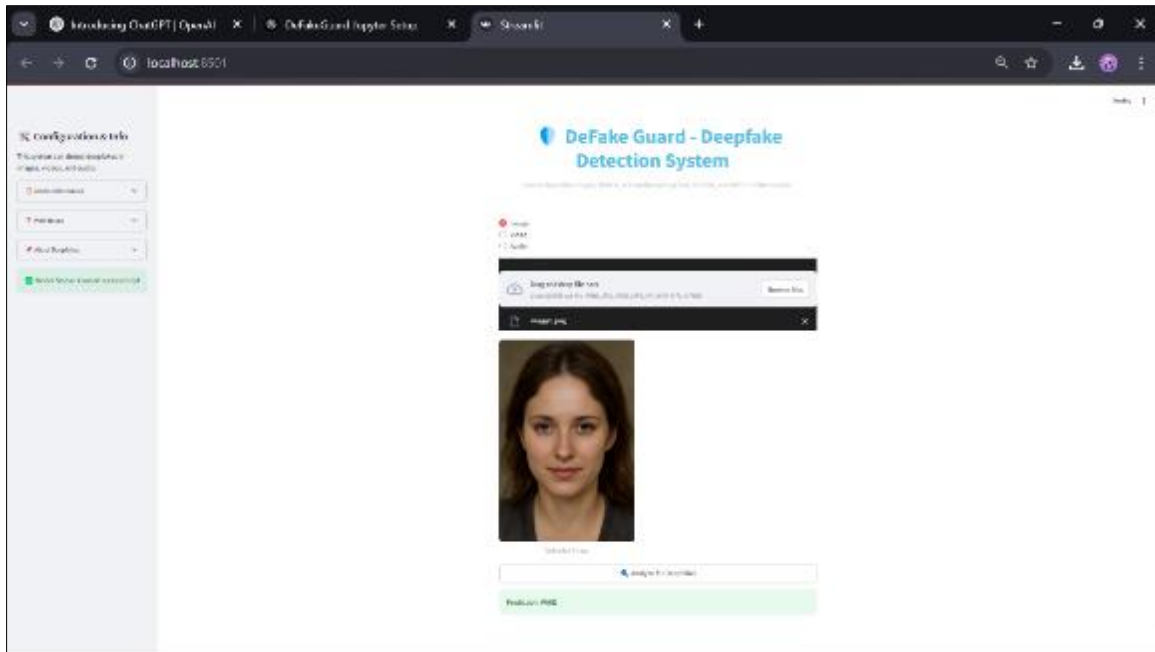


Figure 4 Image Model Prediction

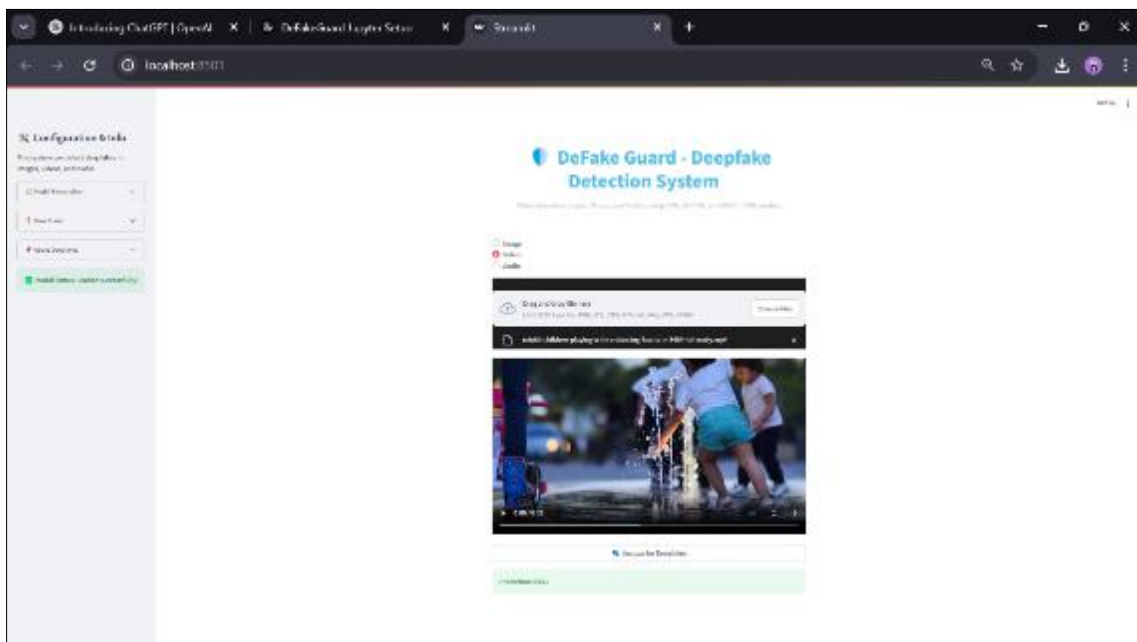


Figure 5 Video Model Prediction

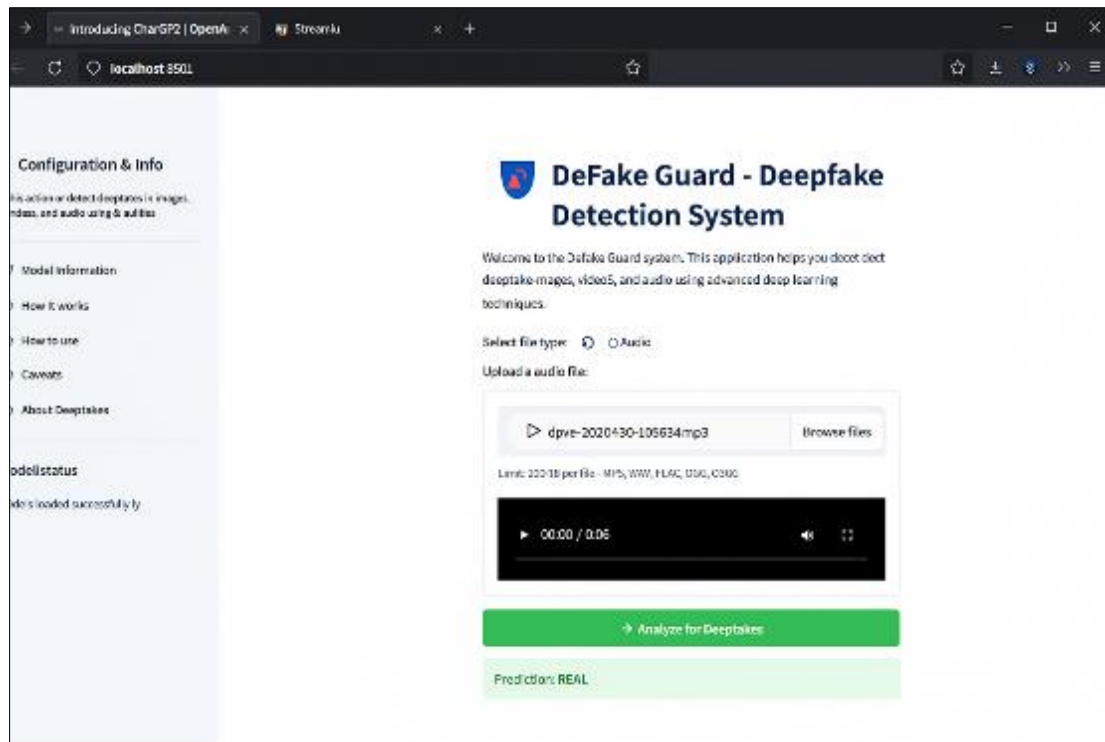


Figure 6 Audio Model Prediction

7. Conclusion

The development of the multi-modal deepfake detection system represents a critical step forward in combating the growing sophistication of synthetic media. By leveraging separate yet collaborative CNN-based pipelines for image, video, and audio analysis, the system achieves high accuracy in detecting manipulations that might escape single-modality approaches. The incorporation of a feature fusion layer enables a comprehensive understanding of cross-modal inconsistencies, while the use of Explainable AI (XAI) techniques enhances model transparency, making predictions more interpretable and trustworthy.

In conclusion, the proposed architecture offers a scalable, reliable, and explainable solution for real-world deepfake detection. Its modular and flexible design allows seamless integration into media forensics, cybersecurity systems, and digital verification platforms. Beyond functioning as a robust detection tool, this project lays the groundwork for future research into more generalized multimodal forgery detection, helping preserve the integrity of digital content across domains.

Compliance with ethical standards

Disclosure of conflict of interest






There is no conflict of interest.

References

- [1] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I., "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE Workshop on Information Forensics and Security (WIFS), pp. 1–7, 2018.
- [2] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C., "The Deepfake Detection Challenge Dataset," arXiv preprint arXiv:2006.07397, 2020.
- [3] Zhang, Y., Zhang, J., Li, X., & Deng, Y., "Multi-modal Deepfake Detection via Clue Fusion," Proceedings of the 29th ACM International Conference on Multimedia, pp. 1031–1039, 2021.

- [4] Jiang, H., Guo, Y., & Sahidullah, M., "Audio Deepfake Detection Using Temporal Convolutional Networks," Interspeech Conference Proceedings, pp. 3261–3265, 2020.
- [5] Nguyen, H. H., Yamagishi, J., & Echizen, I., "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2307–2311, 2019.
- [6] Korshunov, P., & Marcel, S., "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018.
- [7] Mittal, T., Oh, S. J., & Lee, H., "EmoNet: Understanding Emotions from Face Images Using Deep Learning," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2585–2593, 2020.
- [8] Yang, X., Li, Y., & Lyu, S., "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8261–8265, 2019.
- [9] Tariq, S., Lee, H. J., Woo, S., & Choi, S., "Detecting Both Machine and Human Created Fake Face Images in the Wild," Proceedings of the 2nd Deepfake Detection Challenge Workshop, pp. 1–5, 2019.

Author's short biography

| | |
|---|---|
| <p>Mrs. Saritha Banoth Mrs. B. Saritha is currently working as an Assistant Professor in the Department of Computer Science and Engineering. She holds a B.Tech and M.Tech in CSE and is pursuing her Ph.D. Her areas of interest include Machine Learning and Data Science, where she has guided several student projects. With 8 years of academic experience, she is passionate about research and innovation in intelligent systems.</p> |  |
| <p>Bhavana Chandragiri I am Bhavana Chandragiri, A final-year B.Tech student at ACE Engineering College, specializing in CSE (Data Science). With a strong interest in data science and machine learning, constantly exploring new technologies to enhance knowledge and skills. Focused on applying expertise effectively in real-world scenario.</p> |  |
| <p>Priyamvadhya Ramadugala R.Priyamvadhya is a Final-year B.Tech student specializing in CSE (Data Science) at ACE Engineering College, driven by a passion for programming and data-driven solutions. Enthusiastic about learning innovative technologies and continuously developing skills to make a meaningful impact in the field.</p> |  |
| <p>Harshavardhan Oraganti O.Harshavardhan is a final-year Computer Science (Data Science) student at ACE Engineering College, passionate about data science, programming, and emerging technologies. Exploring new concepts and refining skills is exciting, with a strong eagerness to apply knowledge to solve meaningful challenges.</p> |  |
| <p>Jayanth Konapakula K.Jayanth is a final-year B.Tech student specializing in CSE (Data Science) at ACE Engineering College, passionate about programming and Machine Learning. Enjoys learning about innovative technologies and continuously developing my skills to make a meaningful impact in the field.</p> |  |