International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

AI Powered Multimodal Deepfake Detection

Anamika Suresh[1], Jayalakshmi M[2], Krishnapriya R[3], Sandra Sabu[4], Dr. Rekha K S[5]
Department of CSE, College of Engineering Kidangoor, Kottayam, Kerala, India

***Abstract*— Deepfakes pose a significant threat to digital trust by generating highly realistic fabricated media that depict false actions or statements. Detecting such manipulated content has become increasingly challenging due to the growing advancement of deepfake generation techniques. This paper proposes a multimodal deepfake detection system that leverages both audio and visual modalities to capture inconsistencies between speech and facial movements. Visual features are extracted using ResNet-50, while audio representations are derived from Mel-spectrograms processed through a Convolutional Neural Network (CNN). The extracted features are fused into a unified representation and subsequently used for classification. The system combines Grad-CAM based visual interpretability with a rule-based module to generate textual explanations from model outputs. This approach enables more reliable deepfake detection by integrating cross-modal analysis with explainable outputs, thereby increasing trust and usability in real-world applications.**

***Keywords*—Deepfake Detection, Multimodal Learning, Audio Visual Analysis, Explainable AI, Grad-CAM.**

I. Introduction

The growing realism of deepfake media has made it increasingly difficult to verify the authenticity of digital content, as synthetic manipulations can closely mimic natural facial expressions and speech. Although such technologies have legitimate applications in entertainment and media production, their misuse for political misinformation, financial fraud, and identity impersonation poses serious threats to digital trust and information integrity. As deepfake generation continues to advance in both realism and accessibility, traditional and single-modality detection approaches are becoming less effective, highlighting the need for more robust detection frameworks.

Existing deepfake detection approaches primarily analyze either visual artifacts or audio inconsistencies in isolation. While these unimodal methods can detect coarse manipulations, they often fail when forgeries are subtle or when relevant cues span both audio and visual modalities. In addition, most deep learning-based detectors operate as black boxes, producing predictions without explaining the reasoning behind them, limiting their reliability in forensic and legal contexts where transparent decision-making is essential.

To address these limitations, this paper presents a multimodal deepfake detection system that jointly analyzes visual and audio information from input videos to identify cross-modal inconsistencies between facial movements and speech. The system processes videos through two parallel branches: one for visual features from frames and one for audio features from frequency representations. The extracted features are fused into a unified representation and passed through a classifier to produce a real or fake prediction.

To ensure transparency, Grad-CAM is applied independently to both branches, generating attention heatmaps that highlight the regions most responsible for the classification. A rule-based module then translates these outputs into human-readable textual explanations. By combining multimodal feature analysis with interpretable outputs, the system provides a reliable and practical solution for real-world deepfake detection.

II. Related Works

Deepfake detection has evolved significantly as the threat of manipulated media has grown. Early approaches primarily focused on analyzing visual artifacts such as facial inconsistencies, unnatural textures, and frame-level distortions using CNN-based classifiers. While effective against obvious manipulations, these visual-only methods proved increasingly inadequate as generative models advanced in realism, highlighting the need for more comprehensive detection frameworks.

Audio-based detection approaches emerged to complement visual methods by identifying anomalies in speech patterns and acoustic features. Mel spectrogram representations processed through CNN networks demonstrated reasonable performance in detecting synthetic speech by capturing spectral and temporal irregularities [3]. However, audio-only systems remain vulnerable when cross-modal inconsistencies between speech and facial movements carry the primary forgery signal, reinforcing the need for jointly analyzing both streams [5].

The shift toward multimodal detection was motivated by the limitations of single-stream approaches. Salvi et al. [5] demonstrated that early fusion of facial texture features with voice based representations consistently outperforms both unimodal and late fusion strategies, establishing feature-level fusion as a strong baseline for multimodal detection.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Building on this foundation, Muppalla et al. [3] aligned spatial-temporal visual features with Mel spectrogram-based audio representations, confirming that cross-modal integration significantly improves robustness against subtle manipulations. Alsaeedi et al. [2] further extended multimodal detection by incorporating emotional congruence between lip movements and voice as an additional discriminative cue, demonstrating that affective inconsistencies can reveal manipulation even when individual modalities appear authentic.

Cai et al. [4] introduced a large-scale benchmark for content driven audio-visual forgery detection and localization, moving evaluation beyond binary classification toward spatial and temporal artifact identification. Their work demonstrated that cross-modal inconsistencies can expose subtle manipulations that single-stream detectors consistently miss, advancing the field toward more granular detection capabilities.

The availability of standardized benchmark datasets has been essential in driving progress in this area. Datasets such as FaceForensics++, Celeb-DF, and the Deepfake Detection Challenge have served as primary benchmarks for visual and multimodal detection methods [5]. FakeAVCeleb has emerged as the standard benchmark for audio-visual evaluation, explicitly covering four manipulation categories that enable comprehensive cross-modal assessment [2, 6]. The proposed system follows this established practice and is evaluated on FakeAVCeleb.

More advanced architectures introduced attention mechanisms to better capture cross-modal relationships. Katamneni and Rattani [6] proposed MIS-AVoiDD, which learns both modality invariant and modality-specific representations to bridge the distributional gap between audio and visual data. Their subsequent work [8] introduced contextual cross-modal attention that dynamically weights each modality frame by frame, allowing the model to focus on whichever stream carries the more reliable forgery signal.

Fine-grained detection approaches targeted localized temporal and spatial inconsistencies rather than global synchronization cues. Astrid et al. [9] introduced a framework capturing local temporal misalignments at the frame and phoneme level, demonstrating superior generalization over global approaches. This insight motivated the use of Grad-CAM in the proposed system to spatially localize manipulation artifacts in both video frames and Mel spectrograms.

Explainability has become a critical requirement, particularly for forensic and legal applications where decisions must be transparent and verifiable.

Abir et al. [7] demonstrated that combining CNN classifiers with Grad-CAM in the image domain produces interpretable outputs without sacrificing accuracy. Mansoor and Iliev [1] extended this to video-based detection, showing that transparent models significantly improve user trust in high stakes applications. Hondru et al. [10] introduced ExDDV, the first benchmark for explainable deepfake detection in video, with annotations linking artifacts to human-readable descriptions.

Despite these advances, most existing systems address detection accuracy and explainability independently. Few systems combine cross-modal analysis with interpretable outputs across both modalities simultaneously. This gap motivates the proposed system, which jointly analyzes visual and audio features and provides Grad-CAM-based attention maps along with rule-based textual explanations, offering both reliable detection and human-understandable reasoning for real-world applications.

III. Methodology

The proposed system follows an end-to-end pipeline that processes raw video input to produce a classification decision along with an explainable output. The system separates the input video into audio and visual streams, extracts features from each stream independently using dedicated deep learning models, and fuses the extracted embeddings into a unified representation for classification. Transparent explanations are then generated through an explainable AI module, enabling human-understandable reasoning for every prediction.

System Architecture

The system has been mainly divided into five modules, as shown in Fig. 1.

1. Input and Data Preprocessing Module: The system accepts a raw input video and separates it into visual frames and an audio track. Both streams are resized, normalized, and standardized to prepare them for feature extraction.

2. Visual Feature Extraction Module: Video frames are processed using a pretrained ResNet-50 model to capture facial manipulation artifacts such as texture inconsistencies and micro-expressions, producing a visual feature embedding vector.

3. Audio Feature Extraction Module: The audio track is converted into a Mel spectrogram and processed through a CNN to extract spectral and temporal features that distinguish real and synthetic speech, producing an audio embedding vector.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

4. Fusion and Classification Module: The visual and audio embeddings are concatenated into a unified multimodal representation and passed through a fully connected classifier to produce a binary real or fake prediction with a confidence score.

5. Explainable AI (XAI) and Output Module: Grad-CAM is applied independently to both branches to generate attention heatmaps, and a rule-based module converts these outputs into human-readable textual explanations alongside the prediction.

IV. Implementation And Results

A. Implementation Setup

The system is developed using Python and Django as the backend framework, with HTML and CSS for the frontend user interface. Deep learning models are built and trained using TensorFlow, while audio processing is handled through Librosa and MoviePy. Grad-CAM heatmaps are generated using OpenCV, and SQLite serves as the database for user authentication and session management. The system runs on a standard machine equipped with a multi-core processor, 8GB RAM, and sufficient storage capacity, ensuring smooth execution of deep learning models and efficient processing of audio-visual data for real-time deepfake detection.

B. Dataset

The system is trained and evaluated on the FakeAVCeleb v1.2 dataset, which is specifically designed for multimodal deepfake detection research. The dataset covers four manipulation categories — RealVideo-RealAudio, FakeVideo-FakeAudio, FakeVideo-RealAudio, and RealVideo-FakeAudio — enabling comprehensive evaluation across all possible combinations of audio-visual forgery. The dataset is divided into 80% for training and 20% for validation using stratified sampling to ensure balanced class distribution.

C. System Walkthrough

The system is accessible through a web-based interface that guides the user through the detection process. The user first registers and logs in through the authentication pages before accessing the Video Analysis page. A video file is uploaded through the drag-and-drop interface, and the system validates the file format before initiating the analysis pipeline. The uploaded video is then processed through the five modules described in the methodology section to generate a Detection Report.

Fig. 2 shows the video upload interface with a loaded video file ready for analysis, displaying the filename, file size, and video resolution alongside a preview of the uploaded content. Once the user initiates the analysis by clicking the Run Engine Analysis button, the system simultaneously extracts video frames and the audio track, processes each through the respective feature extraction branches, and combines the outputs through the fusion classifier. The Detection Report shown in Fig. 3 presents the final verdict, prediction confidence score, risk level, and the model used for classification. A confidence score above 0.96 is classified as fake and assigned a High Risk level displayed in red, while a confidence score below 0.96 is classified as real and assigned a Low Risk level displayed in green, providing users with an immediate visual indication of the severity of the detected manipulation.

The Grad-CAM attention heatmaps for both the video frame and the Mel spectrogram are displayed alongside the verdict as shown in Fig. 4, visually highlighting the regions that most influenced the classification decision. Red regions in the heatmap indicate high activation areas where the model detected the strongest evidence of manipulation, while blue regions indicate low impact areas that contributed minimally to the classification.

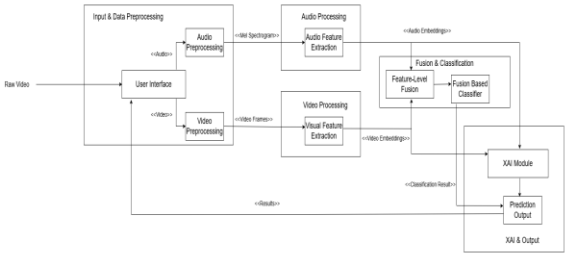

Figure 1: System Architecture

Fig. 5 shows the XAI Explanation page, which presents the rule-based textual observations generated from the model outputs. These observations include indicators such as the percentage of frames exceeding the fake threshold, frame-level confidence variance, detection of strong visual manipulation cues, and identification of cross-modal inconsistencies between video and audio, helping users understand the specific reasons behind the prediction in a clear and accessible manner.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Figure 2: Video Upload Page

Figure 3: Detection Report Page

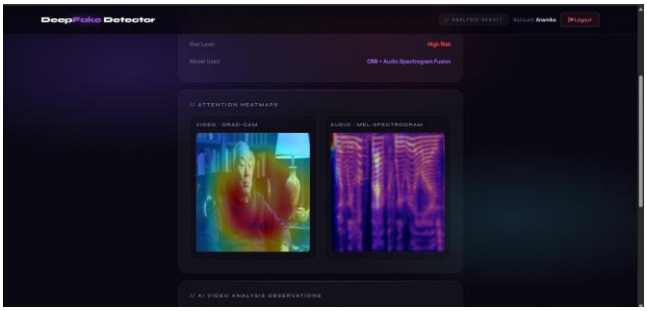

Figure 4: Grad-CAM Attention Heatmaps for Visual and Audio Branches

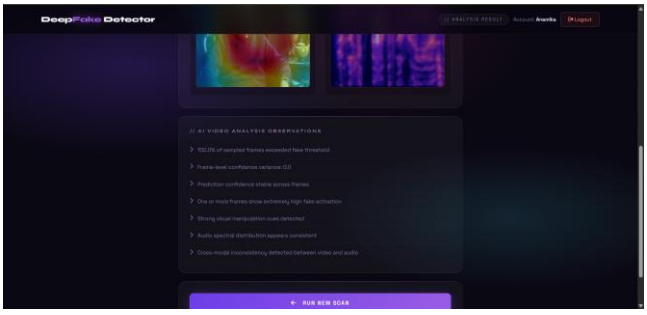

Figure 5: XAI Explanation Page showing Rule-Based Observations

V. CONCLUSION

The Multimodal Deepfake Detection System is an advanced AI powered tool designed to accurately identify manipulated media by analyzing both visual and audio modalities. By integrating ResNet-50 for visual feature extraction and a CNN-based Mel spectrogram model for audio analysis, the system captures cross-modal inconsistencies that unimodal approaches would miss.

The incorporation of Grad-CAM based Explainable AI bridges the critical gap between model accuracy and forensic transparency, allowing users to understand why a video is classified as fake. The web-based Django interface ensures accessibility and real-world usability of the system.

In the future, the proposed system can be further enhanced by incorporating transformer-based architectures for improved cross-modal attention, training on larger and more diverse datasets such as the full DFDC corpus, and optimizing the system for real-time detection in live video streams and social media monitoring platforms.

REFERENCES

[1] N. Mansoor and A. I. Iliev, “Explainable AI for DeepFake Detection,” Applied Sciences, vol. 15, article 725, 2025.

[2] A. Alsaeedi, A. AlMansour, and A. Jamal, “Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 6, 2025.

[3] S. Muppalla, S. Jia, and S. Lyu, “Integrating Audio-Visual Features for Multimodal Deepfake Detection,” arXiv preprint arXiv:2310.03827, Oct. 2023.

[4] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, “Glitch in the Matrix: A Large Scale Benchmark for Content Driven Audio-Visual Forgery Detection and Localization,” Computer Vision and Image Understanding, vol. 236, article 103818, 2023.

[5] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, “A Robust Approach to Multimodal Deepfake Detection,” Journal of Imaging, vol. 9, no. 6, article 122, 2023.

[6] V. S. Katamneni and A. Rattani, “MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection,” arXiv preprint arXiv:2310.02234, Oct. 2023.

[7] W. H. Abir et al., “Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods,” Intelligent Automation & Soft Computing, vol. 35, no. 2, pp. 2151–2169, 2023.

[8] V. S. Katamneni and A. Rattani, “Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization,” arXiv preprint arXiv:2408.01532, Aug. 2024.

[9] M. Astrid, E. Ghorbel, and D. Aouada, “Detecting Audio-Visual Deepfakes with Fine-Grained Inconsistencies,” arXiv preprint arXiv:2408.06753, Oct. 2024.

[10] V. Hondru, E. Hogea, D. Onchis, and R. T. Ionescu, “ExDDV: A New Dataset for Explainable Deepfake Detection in Video,” University of Bucharest and West University of Timisoara, Romania, 2024.