# Deepfake Detection Using Multimodal Biometric Fusion

**Neenamol Joseph[1], Marteena Noble[2], Nehala Naseer[3]**

**[1]Assistant Professor, Department of Computer Science, MG University, Santhigiri College of Computer Sciences, Vazhithala, Kerala, India**

**[2,3]Student, Department of Computer Science, MG University, Santhigiri College of Computer Sciences,Vazhithala, Kerala, India**

**DOI: https://doi.org/10.51584/IJRIAS.2026.11100069**

**Received: 27 October 2026; Accepted: 02 November 2026; Published: 07 November 2026**

## ABSTRACT

The rapid advancement in using AI techniques for creating realistic synthetic media, commonly known as deepfakes, possess the risks of misinformation, identity fraud, and social manipulation. This paper systematically reviews the repercussions of deepfakes and addresses these critical issues through an innovative multimodal approach that integrates multiple biometric traits, unlike unimodal detection, which faces bottlenecks due to its vulnerability to spoofing. To achieve a combined analysis, this model integrates visual features, auditory features, and behavioral features. This also compares various datasets used and fusion strategies. Detecting deepfakes using a multimodal approach isn't easy and gives rise to challenges such as complexity in fusion, limited datasets, and high computing power. This study serves as a foundation for future advancements and can be expanded to be part of apps, social media tools, digital forensics and even help in governance, making it more reliable.

**Index Terms:** CyberSecurity, DeepFake, DeepFake Detection, Feature Extraction, Fusion Strategies, Multimodal Biometric Fusion.

## INTRODUCTION

The rise of deepfake technology in recent years has resulted in serious problems for a number of industries, including cybersecurity, politics, and the media. Synthetic media produced with sophisticated artificial intelligence methods, especially deep learning algorithms, are known as "deepfakes" , and they have the capacity to widely manipulate and take advantage of people. Deepfakes can be used to manipulate people and spread false information in a variety of contexts. Digital de-aging, realistic character reconstruction, and improved visual effects are all made possible by deepfake. Several approaches have been proposed by researchers to solve this problem that focus on a unimodal method, meaning they consider only one type of content, either audio or video. However, these unimodal methods face limitations in effectively capturing the complexity of advanced deepfakes. Although they demonstrate strong performance, detectors that depend entirely on video can be deceived by synthetic audio, and those based only on audio can be misled by manipulated video content. Therefore, they are not reliable for effective deepfake detection. To address these issues, we propose a multimodal analysis framework that combines visual, auditory, and textual features to give a complete view of the media content. This method improves detection capabilities and provides a more durable solution against the changing landscape of deepfake technologies. In this paper, we discuss the methodology that involves extracting visual, auditory, and behavioral features from the input data. We apply early, late, or mid fusion strategies and evaluate the proposed deepfake detection model using standard datasets and performance metrics.

### Objectives Of The Study

1. To analyse the design of multimodal biometric fusion.

2. To evaluate the advanced fusion techniques used in the multimodal analysis.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

3. To conduct a comparative analysis of unimodal and multimodal systems.

4. To examine the effectiveness of multimodal system in real-word application.

## METHODOLOGY

Multimodal biometric fusion enhances deepfake detection by simultaneously analyzing various biometric traits, enabling more robust and reliable identification of synthetic media manipulations. The proposed methodology is composed of six stages:




### Input

The input to the proposed system consists of digital video samples containing both visual (face) and auditory (voice) modalities. Since deepfakes typically manipulate either facial regions, speech signals, or both, the system processes the video as a multimodal signal.

### Preprocessing

Preprocessing prepares the raw input so that it becomes suitable for analysis. Examples:Frame extraction,Face detection & cropping,Noise removal fromaudio,Resizingimages,Normalization,Synchronizing audio and video.At this stage, we are not extracting meaningful patterns yet.We are only cleaning and standardizing the data.

### Feature Extraction Methods

### Visual Feature Extraction

CNN-Based Texture Analysis:Convolutional Neural Networks (CNNs) are widely used for extracting visual features in deepfake detection.CNNs capture global facial structure and semantic information.Local texture analysis helps detect subtle forgery artifacts such as: Unnatural texture transitions, Local detail distortions.Combining global CNN features + local texture features improves detection accuracy.This fusion enables both macro-level and microlevel analysis of manipulated regions.

### Auditory Feature Extraction

Spectrogram Analysis & Speaker Embedding:Spectrograms are used to analyze frequency-time patterns in speech signals.Speaker embeddings capture unique vocal identity characteristics.Fusion of physiological and acoustic features helps detect: Speech synthesis attacks, Voice conversion attacks. Limitations: Requires high-quality speech signals, Performance drops in low signal-to-noise environments, High computational cost may affect real-time implementation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

**Behavioral Feature Extraction**

Behavioral feature extraction focuses on detecting unnatural motion patterns such as eye blinking, head movement, and lip synchronization. LSTM (Long Short-Term Memory):These networks play a crucial role by modeling temporal dependencies across video frames, helping detect irregular motion sequences over time. 3D-CNNs:They are important because they capture both spatial and temporal features simultaneously, enabling analysis of motion continuity and behavioral realism.

**Fusion**

**Late Fusion**

Two separate classifiers: CV → Visual classifier , CA → Audio classifier Each classifier produces its own prediction score:

$$\hat{y}V = CV(fV)\ ,\ \hat{y}A = CA(fA)$$

Final score is obtained by averaging both outputs:

$$\hat{y}_{AV} = (\hat{y}_V + \hat{y}_A) / 2$$

Advantage: Simple and flexible, Each modality works independently.

**Mid Fusion**

Two separate classifiers for visual and audio.Instead of combining final scores, their feature embeddings (before final layer) are extracted.These embeddings are concatenated into a multimodal vector. A fully connected layer produces the final prediction. Important:Trained endto-end, Both modalities influence each other during learning.

**Early Fusion**

Only one classifier (CAV). Visual and audio feature vectors are concatenated at the beginning:

$$f_{AV} = [f_V, f_A]$$

The combined feature is directly fed into a single classifier. Advantage: Learns strong audiovisual correlations early, Potentially better integration of modalities.

**Classification**

After feature extraction and fusion (Early, Mid, or Late), the fused feature vector is fed into a classifier to determine whether the input is real or fake. Common Classifiers: Fully Connected (Dense) layers, Softmax or Sigmoid activation, CNN, LSTM (for temporal patterns), or hybrid deep networks. Working: The fused feature $fAV$passes through dense layers, and the final layer outputs a probability score (e.g., 0 = Real, 1 = Fake, or 0.85 = 85% fake).

$$\hat{y}_{AV} = C(f_{AV})$$

Where $f_{AV}$ is the fused feature and $\hat{y}_{AV}$ is the predicted score.

**Output**

The output module converts the classifier score into a final decision.

Output Includes: Binary decision (Real / Fake), Confidence score, Performance metrics during evaluation: Accuracy, Precision , Recall, F1-score, ROC-AUC.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

DataSets and Evaluation Metrics

Key Datasets for Multimodal Deepfake Detection: IAV-DF (2025) – 10k+ real and 247k+ fake videos; focuses on cross-cultural audio-visual biometric verification; evaluated using Accuracy, F1-score, and AUC.

DECRO (2023) – Cross-lingual audio-visual dataset; supports multilingual deepfake detection with strong generalization.

DeepfakeBench – Standardized benchmark covering multiple generation techniques and modalities; improves model comparability.

PUDD (2024) – Prototype-based multimodal framework; robust against unknown generation techniques.

Training Datasets (Monomodal): FaceForensics++ – Visual-only deepfake dataset, ASVspoof 2019 (LA) – Audio-only spoofing dataset.

Evaluation Datasets (Multimodal):FakeAVCeleb, DFDC (multimodal subset), VidTIMIT, DeepfakeTIMIT, and TIMIT-TTS — used to test cross-dataset robustness and generalization.

**Table 1.** Evaluation Metrics

| Accuracy | F1-Score | Area Under Curve | Cross-Dataset Detection Performance | Processing Efficiency |
|---|---|---|---|---|
| Measures the overall detection performance of the model, suitable for balanced datasets. | It comprehensively takes into account precision and recall, and is suitable for imbalanced datasets. | Assesses the model's classification performance across different thresholds, applicable to binary classification tasks. | Measures the model's generalization ability on unseen datasets, a critical metric for deepfake detection. | Evaluates the computational complexity and real-time detection capability of the model, applicable to practical application scenarios. |

## Challenges

Multimodal deepfake detection faces several challenges. The limited availability of large, diverse multimodal datasets restricts proper training and cross-dataset generalization. Additionally, audio–video synchronization issues can negatively affect fusion performance. The use of complex models like 3D CNNs and LSTMs also increases computational cost, making real-time deployment difficult.

## Global Case Studies

1. Pavel Korshunov & Sebastien Marcel (2019): The authors worked on audio-visual inconsistencies in the detection of deepfakes, using facial features and speaker characteristics. This research is considered to be one of the first attempts at multimodal biometric fusion in the detection of deepfakes.

2. Yuezun Li et al. (2018-2020): The authors created models using the fusion of visual features and noise artifacts, enhancing the detection performance. Their research proved that the fusion of different features can make the detection process more robust against manipulated media.

3. Umur Aybars Ciftci et al. (2020): The authors presented a novel idea for the detection of deepfakes using biological signals such as remote photoplethysmography (rPPG), which can be extracted from facial video signals.

4. Siwei Lyu and others: Their research involved the detection of inconsistencies in eye-blinking, head pose, and physiological signals. These features can make the detection process more efficient using fusionbased detection.

5. Multimodal Transformer and Attention-Based Fusion (2021-2023): Recently, some researchers have presented transformer-based models for the detection of deepfakes using facial landmarks, lip-syncing, voice, and temporal dynamics. The models have shown high detection accuracy against advanced deepfakes using attention and graph neural networks.

**Implementation And Sustainability**

The system processes video input by separating audio and visual data, extracting features using CNNs, 3D-CNNs, and LSTMs. Fusion strategies (early, mid, or late) combine these featuresbefore classification into real or fake. It can be deployed in social media platforms,forensic tools, and verification systems, with optimization techniques enabling real-time performance.

Multimodal fusion improves stability by reducing reliance on a single modality. Even if one signal is noisy or manipulated, others support accurate detection. Cross-dataset evaluation further ensures robustness and generalization across different environments.

**Future Directions**

1.Fine-Grained Biometric Feature Mining: We look at details like micro-expressions, eye movements and heart rates to spot fake videos. These tiny signs are hard to fake so checking them can make our deepfake detection better.

2.Lightweight Detection Models: Models like MobileNet are important for making deepfake detection systems that work in time. To make them lightweight we use techniques like pruning, quantization and knowledge distillation.

3. Interpretability Enhancement: Making our models more understandable can help explain why a video is detected as fake. We can use heatmap regression and 3D key-point detection to achieve to achieve this.

4. Incorporation of RNNs and LSTMs: Adding RNNs and LSTMs to CNNs helps us analyze how the video changes over time. This is useful, for spotting movements of people in the video.

5.Integration of Advanced Preprocessing Techniques: Using techniques like noise reduction and video enhancement can improve our models performance. By reducing noise we can improve the audio and video quality, which in turn makes our model work better.

## CONCLUSION

Deepfake technology is evolving rapidly, creating serious risks in areas such as cybersecurity, media authenticity, and digital governance. As manipulation techniques become more advanced, traditional unimodal detection methods are no longer sufficient. This study highlights how multimodal biometric fusion offers a stronger and more reliable solution by combining visual, auditory, and behavioral features. By analyzing facial textures, speech patterns, and motion dynamics together, the system can detect subtle inconsistencies that single-modality approaches often miss.The integration of advanced models such as CNNs, 3D-CNNs, and LSTMs further strengthens the detection framework by capturing both spatial details and temporal patterns. Although challenges such as limited multimodal datasets, synchronization issues, and computational complexity remain, the multimodal approach demonstrates improved robustness and generalization across diverse datasets.Overall, this work establishes a solid foundation for developing more secure and scalable deepfake detection systems. With further research focused on lightweight architectures, better datasets, and improved interpretability, multimodal biometric fusion can play a vital role in safeguarding digital content and building trust in online media platforms.

## REFERENCES

1. Multimodal Deepfake Detection,Authors: Prof. Sneha G, Prof. Divya S, Lavanya R, Leesha V Kumar, Navyatha A, Nishu Agarwal. Available:https://doi.org/10.22214/ijraset.2024.62936
2. A Research on Deepfake Face Detection Techniques Based on Multimodal Biometric Cross – Verification, Haozhe Wu, School of Computer and Software, Hohai University, Nanjing, Jiangsu, China, ITM Web of Conferences 78, 02017 (2025),CSEIT 2025 Available:https://doi.org/10.1051/itmconf/20257802017
3. Robust Approach to Multimodal Deepfake Detection, Salvi, D.; Liu, H.; Mandelli,S.; Bestagini, P.; Zhou, W.; Zhang, W.;Tubaro, S. A J.Imaging 2023. Available : https://doi.org/10.3390/jimaging9060122
4. From Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges, Authors: Ping Liu, Senior Member, IEEE, Qiqi Tao, and Joey Tianyi Zhou, Senior Member, IEEE. Available: https://arxiv.org/abs/2406.06965
5. A Multimodal Framework for DeepFake Detection, Authors: Kashish Gandhi, Prutha Kulkarni, Taran Shah, Piyush Chaudhari, Meera Narvekar and Kranti Ghag Available : https://arxiv.org/abs/2410.03487
6. Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis, Authors: Sagar Nailwal,Nongmeikapam Thoiba Singh,Arbaz Raza,Saksham Singhal. Available:https://ieeexplore.ieee.org/abstract/document/10369155/
7. How Multimodal Biometric Systems are Winning the War on AI-Generated Deepfakes? Author: Arjun Singh. Available : https://blog.mantratec.com/Multimodal-Biometric-Systems-Vs-AI-Deepfakes
8. Rui Wang, Dengpan Ye, Long Tang, Yunming Zhang, Jiacheng Deng (2024) AVT2-DWF: Improving Deepfake Detection with Audio-Visual Fusion and Dynamic Weighting Strategies. Available : https://arxiv.org/abs/2403.14974
9. Zhimao Lai, Yufei Wang, Renhai Feng, Xianglei Hu, Haifeng Xu (2022) Multi-Feature Fusion Based Deepfake Face Forgery Video Detection. Available :https://www.mdpi.com/2079-8954/10/2/31
10. Siddiqui Muhammad Yasir, Hyun Kim (2025), Lightweight Deepfake Detection Based on Multi-Feature Fusion. Available : https://doi.org/10.3390/app15041954
11. Yuncan Sheng, Zhengrui Zou, Zongxuan Yu, Wei Ou (2025), ID-Insensitive Deepfake Detection Model Based on Multi-Attention Mechanism. Available : https://www.nature.com/articles/s41598-025-96254-6
12. Hasam Khalid, Shahroz Tariq, Minha Kim, Simon S. Woo (2021) FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. Available:https://github.com/DASHLab/FakeAVCeleb
13. Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Ser-Nam Lim, YuGang Jiang (2021) M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. Available : https://arxiv.org/abs/2104.09770