2026 Volume: 02 Issue: 06 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-11

DEEPFAKE AUDIO AND VIDEO DETECTION USING SIGNAL PROCESSING AND MACHINE LEARNING

***Sanika Yadav, Resham Powar, Dr. Manisha Bhanuse**

D.Y. Patil College of Engineering & Technology.

Article Received: 26 April 2026
Article Revised: 16 May 2026
Published on: 05 June 2026

***Corresponding Author: Sanika Yadav**
D.Y. Patil College of Engineering & Technology.

DOI: https://doi-doi.org/101555/ijrpa.4742

ABSTRACT

The emergence of deepfake technology has made the creation of realistic synthetic videos and audio easier than ever, creating challenges for digital trust and information authenticity. This work presents a multimodal detection approach that combines signal processing methods with machine learning techniques to identify manipulated media content. Video and audio features are extracted using techniques such as FFT, DCT, MFCC and STFT to capture hidden artifacts introduced during media generation. These features are then analysed by machine learning models to determine whether the content is genuine or fabricated. By jointly examining visual and audio information, the proposed framework aims to improve detection performance and support applications in cyber security, digital forensics and media verification.

***INDEX TERMS*:** Deepfake Detection, Artificial Intelligence (AI), Machine Learning, Signal Processing, Deep Learning, Video Forensics, Audio Forensics, Mel-Frequency Cepstral Coefficients (MFCC), Short-Time Fourier Transform (STFT), Multimedia Security, Cybersecurity, Digital Forensics.

INTRODUCTION

Deepfake technology is an advanced application of artificial intelligence (AI) that creates realistic fake videos and audio by manipulating original media content. Using deep learning techniques, deepfakes can imitate a person's face, voice, and expressions, making the generated content appear authentic. While this technology has useful applications in entertainment, education, and media production, its misuse has raised serious concerns regarding misinformation, identity theft, cybercrime, and privacy violations.

The widespread use of social media and digital communication platforms has increased the risk of deepfake content being used to spread false information and influence public opinion. As deepfakes become more realistic, detecting manipulated media through human observation alone becomes extremely difficult. Therefore, there is a growing need for automated and reliable detection systems.

This research focuses on detecting deepfake video and audio using signal processing and machine learning techniques. Methods such as Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), Mel-Frequency Cepstral Coefficients (MFCC), Spectrogram Analysis, and Short-Time Fourier Transform (STFT) are used to extract important features from multimedia content. These features are then analysed using machine learning models to classify media as genuine or manipulated. The proposed approach aims to improve detection accuracy and contribute to cybersecurity, digital forensics, and media authenticity.

Literature Review

Several researchers have investigated deepfake detection using machine learning, deep learning, and signal processing techniques. Their studies provide valuable insights into identifying manipulated video and audio content.

Ayan Sar et al. (2025) proposed a unified multimodal framework for real-time deepfake detection by combining video and audio analysis. The authors employed CNN, LSTM, and DenseNet121 models to learn discriminative features from multimedia data. Experimental results on Celeb-DF and FaceForensics++ datasets demonstrated high detection accuracy for both video and audio content. However, the framework required extensive training data and significant computational resources.

Rami Mubarak et al. (2023) presented a comprehensive survey of deepfake generation and detection techniques. Their work analysed the impact of deepfakes on cybersecurity, privacy, and digital trust while reviewing methods based on GANs, CNNs, and transformer architectures. The study highlighted existing challenges in detecting highly realistic synthetic media and emphasized the need for more robust detection mechanisms.

Kavya Verma et al. (2025) focused on deepfake audio detection using advanced deep learning models and MFCC-based feature extraction. The proposed approach was evaluated using the ASVspoof2019 dataset and achieved promising detection performance. The study demonstrated the effectiveness of audio feature analysis for identifying synthetic speech but required high computational power for implementation.

Gunwoo Lee et al. (2025) introduced a dual-channel deepfake audio detection framework that analyzed both direct and reverberant speech signals. The method utilized MFCC, LFCC, and DPTNet-based architectures to improve detection reliability. Experimental evaluation showed improved performance metrics; however, the model architecture was complex and computationally demanding.

Paidimalla Naga Raju et al. (2024) investigated deepfake detection using AI-based signal processing techniques. Their work emphasized spectral analysis and machine learning methods for identifying hidden artifacts introduced during media manipulation. The study demonstrated that signal-level analysis can effectively detect fake content, although performance may decrease against highly advanced GAN-generated deepfakes.

Siddhant Mankar et al. (2025) proposed an audio deepfake detection system that combined signal processing techniques such as MFCC, FFT, and spectral analysis with CNN and SVM classifiers. The research achieved satisfactory detection accuracy and highlighted the benefits of integrating machine learning with signal processing methods.

Shivam Sanjay Parab et al. (2025) conducted a comprehensive review of machine learning and deep learning approaches for deepfake audio detection. The study compared CNN, RNN, LSTM, and Transformer-based models and discussed their strengths and limitations. The authors concluded that although existing models provide strong detection capabilities, challenges related to generalization and dataset dependency still remain.

From the reviewed literature, it is observed that most existing approaches achieve high detection accuracy but often require large datasets, complex architectures, and substantial computational resources. Therefore, a multimodal framework that combines signal processing and machine learning techniques can provide an efficient and reliable solution for deepfake video and audio detection.

Proposed Methodology

This research presents a multimodal framework for detecting deepfake video and audio by integrating signal processing techniques with machine learning algorithms. The proposed methodology is designed to identify subtle artifacts and inconsistencies introduced during the generation of synthetic media. The overall workflow consists of data acquisition, preprocessing, feature extraction, multimodal fusion, classification, and performance assessment.

1 Multimedia Data Acquisition

To ensure comprehensive evaluation, benchmark datasets containing both authentic and manipulated media samples are utilized. Video datasets such as FaceForensics++, Celeb-DF, and Deepfake Detection Challenge (DFDC), along with the ASVspoof audio dataset, provide diverse examples of deepfake content generated using different techniques. These datasets serve as the foundation for model training and validation.

2 Preprocessing and Data Enhancement

Before analysis, the collected media undergoes preprocessing to improve data quality and reduce irrelevant variations. Video sequences are converted into individual frames and standardized to a uniform format. Similarly, audio recordings are normalized and filtered to suppress background noise. This stage ensures that the extracted features accurately represent the underlying characteristics of the media.

3 Signal-Based Feature Analysis

The core of the proposed framework lies in identifying manipulation traces hidden within video and audio signals. For video analysis, frequency-domain representations are obtained using Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT). In addition, facial landmark movements, blinking behaviour, and temporal consistency across consecutive frames are examined to reveal abnormal visual patterns. For audio analysis, acoustic features are extracted using Mel-Frequency Cepstral Coefficients (MFCC), Spectrogram Analysis, and Short-Time Fourier Transform (STFT). These techniques capture spectral and temporal information that may indicate the presence of synthesized or altered speech.

4 Multimodal Feature Integration

Features obtained from both visual and audio domains are combined to create a comprehensive representation of the input media. This fusion strategy allows the system to exploit complementary information from different modalities, thereby improving its capability to detect sophisticated deepfake manipulations that may not be evident when analysing a single source.

5 Machine Learning-Based Classification

The integrated feature set is supplied to machine learning classifiers for decision-making. Algorithms such as Support Vector Machine (SVM), Random Forest (RF), and XGBoost are employed to learn discriminative patterns associated with genuine and manipulated content. Through supervised learning, the models establish classification boundaries that enable accurate deepfake identification.

6 Performance Assessment

The effectiveness of the proposed framework is evaluated using standard performance metrics, including Accuracy, Precision, Recall, F1-Score, and Receiver Operating Characteristic (ROC) analysis. Comparative evaluation across different datasets is conducted to assess the robustness and generalization capability of the system. By combining signal-level analysis with machine learning-based decision-making, the proposed methodology aims to provide a reliable, interpretable, and scalable solution for deepfake video and audio detection. The multimodal design enhances detection performance while supporting applications in cybersecurity, digital forensics, and multimedia authentication.

Datasets used

The effectiveness of a deepfake detection system largely depends on the quality and diversity of the datasets used for training and evaluation. In this research, widely recognized public datasets containing both authentic and manipulated multimedia samples are considered to ensure reliable performance assessment.

1 FaceForensics++

FaceForensics++ is one of the most extensively used datasets for deepfake video analysis. It contains original videos along with manipulated versions generated using various face modification techniques. The dataset provides high-quality samples that enable the detection system to learn visual inconsistencies introduced during facial manipulation.

2 Deepfake Detection Challenge (DFDC)

The Deepfake Detection Challenge (DFDC) dataset was developed to support research in multimedia forensics and deepfake identification. It consists of a large collection of real and synthetic videos created under different conditions. The diversity of subjects, lighting environments, and manipulation methods makes this dataset valuable for evaluating the robustness of detection models.

3 Celeb-DF

Celeb-DF is a benchmark dataset designed to address the limitations of earlier deepfake datasets. It contains high-quality manipulated videos of public personalities with fewer visual artifacts, making detection more challenging. The dataset is widely used to assess the capability of detection systems against realistic deepfake content.

4 ASVspoof

ASVspoof is a standard dataset used for deepfake audio and synthetic speech detection. It includes both genuine and artificially generated voice recordings created using various

speech synthesis and voice conversion techniques. The dataset helps evaluate the effectiveness of audio-based feature extraction and classification methods. The combination of these datasets provides a comprehensive collection of video and audio samples for developing and validating the proposed multimodal deepfake detection framework. Their diversity enables the system to learn a wide range of manipulation patterns and improves its ability to generalize across different deepfake generation techniques.

Feature Extraction Techniques

Feature extraction plays a crucial role in deepfake detection by transforming raw video and audio data into meaningful representations that can be analysed by machine learning models. In the proposed framework, both visual and acoustic features are extracted using signal processing techniques to identify hidden artifacts associated with manipulated media.

1. Fast Fourier Transform (FFT)

Fast Fourier Transform (FFT) is used to convert signals from the time domain to the frequency domain. In deepfake video analysis, FFT helps identify abnormal frequency patterns and distortions introduced during the generation process. These frequency-based characteristics can reveal inconsistencies that are not easily visible in the original video frames.

2. Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT) is employed to analyse the distribution of visual information within video frames. It is effective in capturing compression artifacts and subtle modifications that often occur during deepfake creation. DCT-based features provide valuable information for distinguishing authentic videos from manipulated ones.

3. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a widely used technique for extracting important speech characteristics from audio signals. It represents the spectral properties of speech in a manner similar to human auditory perception. In deepfake audio detection, MFCC features help identify unnatural voice patterns and artifacts introduced by speech synthesis systems.

4. Spectrogram Analysis

A spectrogram provides a visual representation of how signal frequencies change over time. By examining spectral variations, it becomes possible to detect irregular patterns and inconsistencies present in synthetic audio. Spectrogram-based features are particularly useful for identifying manipulated speech recordings.

5. Short-Time Fourier Transform (STFT)

Short-Time Fourier Transform (STFT) is used to analyse non-stationary audio signals by capturing both temporal and frequency information. This technique enables the detection of subtle variations in speech characteristics that may indicate audio manipulation. STFT features contribute significantly to improving deepfake audio detection performance.

6. Facial Landmark and Temporal Features

In addition to frequency-based features, facial landmark tracking is employed to analyse facial movements, eye blinking behaviour, and expression consistency across video frames. Temporal features help identify unnatural transitions and motion irregularities that commonly occur in deepfake videos.

The extracted visual and audio features are subsequently combined to form a comprehensive feature representation. This multimodal feature set provides rich information for machine learning classifiers, enabling accurate and reliable identification of deepfake content.

Machine Learning Models

Machine learning algorithms play a vital role in the proposed deepfake detection framework by analysing extracted features and classifying multimedia content as genuine or manipulated. After feature extraction and fusion, the generated feature vectors are used to train supervised learning models capable of recognizing patterns associated with deepfake media.

1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classification algorithm widely used for pattern recognition tasks. It works by identifying an optimal decision boundary that separates different classes within the feature space. In the proposed framework, SVM utilizes the extracted visual and audio features to distinguish between authentic and manipulated content. Its ability to handle high-dimensional data makes it suitable for deepfake detection applications.

2. Random Forest (RF)

Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification performance. Each tree is trained on a different subset of the data, and the final prediction is obtained through majority voting. The model is effective in handling complex feature relationships and reducing overfitting, making it a reliable choice for multimedia classification tasks.

3. Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced boosting algorithm designed to enhance predictive accuracy through iterative learning. It builds a sequence of decision trees where each new tree attempts to correct the errors of previous models. Due to its efficiency and strong generalization capability, XGBoost has shown excellent performance in various classification problems, including deepfake detection.

4. Model Training and Classification

The selected machine learning models are trained using labelled datasets containing both real and manipulated video and audio samples. During training, the models learn distinctive patterns present in the extracted features. Once trained, they classify unseen multimedia content into genuine or deepfake categories based on learned characteristics.

5. Comparative Evaluation

The performance of SVM, Random Forest, and XGBoost is compared using evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. This comparative analysis helps identify the most suitable classifier for achieving reliable and efficient deepfake detection.

By leveraging multiple machine learning algorithms, the proposed framework enhances classification robustness and improves its ability to detect sophisticated deepfake manipulations across diverse multimedia datasets.

RESULTS & DISCUSSIONS

The proposed deepfake detection framework combines signal processing techniques with machine learning algorithms to identify manipulated video and audio content. By integrating information from both visual and acoustic modalities, the system is expected to provide improved detection performance compared to approaches that rely on a single source of information.

The extracted features obtained through FFT, DCT, MFCC, Spectrogram Analysis, and STFT are designed to capture hidden artifacts introduced during the deepfake generation process. Visual features such as facial landmark inconsistencies, abnormal blinking patterns, and temporal irregularities contribute to video analysis, while spectral and temporal audio features assist in identifying synthetic speech characteristics. The combination of these features creates a comprehensive representation of multimedia content.

Machine learning classifiers such as Support Vector Machine (SVM), Random Forest (RF), and XGBoost are expected to effectively learn the distinguishing characteristics of genuine

and manipulated media. The multimodal feature fusion strategy is anticipated to improve classification accuracy and reduce false detection rates by utilizing complementary information from both video and audio domains.

The performance of the proposed framework can be assessed using evaluation metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These metrics provide a comprehensive measure of the system's reliability and effectiveness in identifying deepfake content. Furthermore, comparative analysis among different classifiers can help determine the most suitable model for practical deployment.

Overall, the proposed approach is expected to contribute to digital forensics, cybersecurity, and media authentication by providing a reliable mechanism for detecting deepfake videos and audio. The integration of signal processing and machine learning offers a balanced solution that combines interpretability, computational efficiency, and detection capability.

FUTURE SCOPE

Deepfake generation techniques are continuously evolving, making the development of advanced detection methods an ongoing research challenge. Future work can focus on improving the robustness and adaptability of deepfake detection systems against newly emerging manipulation techniques and AI-generated content.

The proposed framework can be enhanced by incorporating advanced deep learning architectures such as Transformer networks and Vision Transformers (ViT) to improve feature learning and classification performance. The integration of hybrid models that combine machine learning and deep learning approaches may further increase detection accuracy.

Future research can also explore real-time deepfake detection systems capable of monitoring multimedia content on social media platforms and online communication networks. Such systems can help reduce the spread of misinformation and manipulated media before it reaches a large audience.

Another promising direction is the development of cross-dataset and cross-platform detection models that maintain high performance across different types of deepfake generation techniques. Improving model generalization will enable reliable detection even when encountering previously unseen manipulations.

In addition, the use of explainable artificial intelligence (XAI) can provide greater transparency in the decision-making process of detection systems. This can help investigators

and users understand the reasons behind classification results and increase trust in automated detection mechanisms.

Overall, future advancements in signal processing, machine learning, and artificial intelligence are expected to contribute significantly to the creation of more accurate, scalable, and reliable deepfake detection solutions for cybersecurity, digital forensics, and multimedia authentication.

CONCLUSION

The rapid advancement of artificial intelligence has significantly increased the capability to generate highly realistic deepfake videos and audio, creating challenges for digital security, media authenticity, and public trust. As manipulated multimedia content becomes more sophisticated, the need for reliable detection mechanisms continues to grow.

This research presents a multimodal deepfake detection framework that integrates signal processing techniques with machine learning algorithms to identify manipulated content. By analysing both visual and audio information, the proposed approach aims to detect hidden artifacts and inconsistencies introduced during the deepfake generation process. Techniques such as FFT, DCT, MFCC, Spectrogram Analysis, and STFT are utilized to extract meaningful features, while machine learning models are employed for accurate classification.

The combination of video and audio analysis is expected to enhance detection reliability and provide improved performance compared to single-modal approaches. The proposed framework offers a practical and interpretable solution that can support applications in cybersecurity, digital forensics, media verification, and content authentication.

In summary, the integration of signal processing and machine learning provides a promising direction for combating the growing threat of deepfake media. The proposed system contributes toward building a safer digital environment by assisting in the identification of manipulated multimedia content and promoting trust in digital communications.

REFERENCES

1. Ayan Sar, Subhangi Sati, Tanupriya Choudhury, Purvika Joshi, Roohi Sille, K. Srihari, Kritika Bansal (2025). A Unified Neural Framework for Real-Time Deepfake Detection Across Multimedia Modalities to Combat Misleading Content. IEEE
2. Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, Simon Parkinson [2023]. A Survey on Detection & Impacts of Deepfakes in Visuals, Audio and Textual Formats. IEEE

3. ++Kavya Verma, Divyanshi Mittal, Sagnik Samanta, Kabir Gulati, Ojas Kulkarni, Muzaffar Ahmad Dar, C. L. Biji [2025]. Deepfake Audio Detection: A Comparative Study of Advanced Deep Learning Models. IEEE
4. Gunwoo Lee, Jungmin lee, Minkyo Jung, Joseph Lee, Kihun Hong, Souhwan Jung, Yoseob Han [2025]. Dual Channel Deepfake Audio Detection: Leveraging Direct & Reverberant Waveforms. IEEE
5. Paidimalla Naga Raju, Dr Prasad Rayi, Dr Rayi Rajasree Yandra, Rama Subbanna [2024]. Deepfake Detection using AI-Based Signal Processing. JES
6. Siddhant Mankar, Harsh Pore, Prasad Modhave, Om Juvekar, Dr. Geetha Chillarge [2025]. Deepfake Audio detection Using Deep Learning Techniques. IJSDR
7. Shivam Sanjay Parab, Om Chandrakant Parab, Shweta Avadhut Yenaji, Tanvi Prashant Sawant [2025]. Deepfake Audio detection Using Deep Learning and Machine Learning-A Comprehensive Review. IJCRT
8. IEEE Xplore Digital Library. Available: https://ieeexplore.ieee.org
9. Google Scholar. Available: https://scholar.google.com
10. FaceForensics++ Dataset. Available:
11. https://justusthies.github.io/posts/faceforensics++/
12. Deepfake Detection Challenge (DFDC) Dataset. Available:
13. https://ai.meta.com/datasets/dfdc/
14. ASVspoof Challenge Dataset. Available: https://www.asvspoof.org
15. National Institute of Standards and Technology (NIST). Available: https://www.nist.gov
16. Kaggle Datasets. Available: https://www.kaggle.com/datasets