

# Multi-Stream Deepfake Video Detection Using Heartbeat Signals and Lip-Sync Verification

Neha Khannum A<sup>1</sup>, Saniya D<sup>2</sup>, Sneha K S<sup>3</sup>, Syeda Asra<sup>4</sup>, Vinutha D<sup>5</sup>

<sup>1,2,3,4</sup> Department of Information Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere Karnataka, India.

<sup>5</sup> Assistant Professor, Department of Information Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere Karnataka, India.

**To Cite this Article:** Neha Khannum A<sup>1</sup>, Saniya D<sup>2</sup>, Sneha K S<sup>3</sup>, Syeda Asra<sup>4</sup>, Vinutha D<sup>5</sup>, "Multi-Stream Deepfake Video Detection Using Heartbeat Signals and Lip-Sync Verification", International Journal of Scientific Research in Engineering & Technology, Volume 06, Issue 05, September-October 2026, PP:181-186.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by-nc-nd/4.0/); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Abstract:** The rapid development of Artificial Intelligence (AI), deep learning, and generative models has enabled the creation of highly realistic manipulated videos known as deepfakes. These videos can alter facial identities, expressions, and lip movements, creating challenges for digital media authenticity and security. Conventional deepfake detection methods primarily rely on frame-level visual analysis, which may be insufficient when manipulation artifacts are subtle or videos undergo compression and post-processing. This paper proposes a multi-stream deepfake video detection system integrating Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM), remote Photoplethysmography (rPPG), and lip-sync verification. The CNN-LSTM stream extracts spatial facial features and examines temporal inconsistencies across consecutive video frames. The rPPG stream analyzes subtle facial color variations to estimate physiological signals associated with blood-volume changes. The lip-sync verification stream examines the synchronization between audio and mouth movements. The outputs of these streams are intended to be integrated to support deepfake classification. By combining visual, temporal, physiological, and audio-visual evidence, the proposed system aims to provide a more comprehensive approach to deepfake video detection.

**Keywords:** Deepfake Detection; CNN-LSTM; Remote Photoplethysmography (rPPG); Lip-Sync Verification; Video Forgery Detection.

## I. INTRODUCTION

The rapid development of Artificial Intelligence (AI) and deep learning has enabled the creation of highly realistic manipulated videos known as deepfakes. These videos can alter facial identity, expressions, and lip movements, creating risks such as misinformation, identity impersonation, financial fraud, and digital evidence manipulation. Consequently, reliable deepfake detection has become an important research area in computer vision and digital forensics [1].

Conventional detection methods primarily rely on visual features extracted from individual video frames. However, subtle manipulation artifacts, video compression, and post-processing can make detection challenging. To address these limitations, this paper proposes a multi-stream deepfake video detection system integrating Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM), remote Photoplethysmography (rPPG), and audio-visual lip-sync verification. The system examines spatial and temporal facial features, physiological signals associated with facial blood-volume variations, and synchronization between speech and mouth movements. These complementary features are intended to support the classification of videos as real or potentially manipulated.

### A. Objectives

The objectives of this research are: (1) To extract spatial facial features and identify temporal inconsistencies across video frames using CNN-LSTM. (2) To analyze facial colour variations using rPPG as a complementary physiological feature. (3) To examine synchronization between spoken audio and mouth movements. (4) To combine the outputs of the different analysis streams for deepfake classification. (5) To generate a classification result and confidence score to support video-authenticity assessment.

### B. Scope

The scope includes video preprocessing, face detection, visual-temporal analysis, rPPG signal processing, audio extraction, lip-sync verification, and fusion of the resulting features or scores. The system focuses on identifying potential manipulation using complementary forms of evidence. Its performance may be affected by video quality, lighting, facial movement, compression, and manipulation techniques. The system is intended to assist authenticity assessment and does not guarantee detection of every deepfake or provide definitive legal or forensic proof.

## II. MATERIAL AND METHODS

### A. Existing System

Existing deepfake detection systems primarily rely on visual analysis, deep learning models, and audio-visual consistency checks to identify manipulated videos. Convolutional Neural Networks (CNNs) are commonly used to extract facial features and identify potential manipulation artifacts, including irregular facial textures, inconsistent boundaries, and abnormal skin patterns [1]. Some approaches employ temporal models to examine facial movements and inconsistencies across consecutive video frames.

However, visual-based methods may face difficulties when manipulation artifacts are subtle or videos undergo compression and post-processing. Temporal inconsistencies may also be difficult to identify in highly realistic manipulated videos. In addition, conventional single-stream approaches may not fully exploit complementary physiological and audio-visual information.

These limitations motivate the development of a multi-stream detection system that combines visual and temporal analysis with remote Photoplethysmography (rPPG) and lip-sync verification to examine different forms of evidence when assessing video authenticity.

### B. Proposed System

The proposed system employs a multi-stream deepfake video detection approach that combines visual-temporal analysis, remote Photoplethysmography (rPPG), and audio-visual lip-sync verification. The objective is to examine complementary features of a video to identify potential manipulation.

Initially, the input video is validated and processed to extract selected frames and facial regions. The CNN component extracts spatial features from facial frames, while the LSTM component analyzes temporal relationships between consecutive frames to identify potential inconsistencies.

The rPPG stream examines subtle variations in facial skin colour over time to derive physiological features associated with blood-volume changes. Since these signals can be affected by lighting, movement, and video quality, they are treated as complementary evidence rather than definitive indicators of authenticity [2].

The lip-sync verification stream analyzes the correspondence between speech audio and visible mouth movements. The resulting audio-visual consistency information is combined with the visual-temporal and physiological outputs through a feature- or score-level fusion mechanism [3].

Finally, a classification module uses the combined information to produce a prediction indicating whether the input video is classified as real or potentially manipulated. A confidence score may also be provided to support interpretation of the prediction. The overall approach aims to reduce dependence on individual visual artifacts by considering multiple forms of evidence.

### C. System Architecture

The proposed system follows a multi-stream architecture for analyzing video content using visual, temporal, physiological, and audio-visual information. The workflow begins with video input and preprocessing, followed by face detection and feature extraction. The processed data is then analyzed by the corresponding detection streams.

Initially, the input video is validated, and selected frames are extracted. Facial regions are detected and processed for visual analysis. The CNN extracts spatial features from the facial frames, while the LSTM examines temporal relationships across consecutive frames.

In parallel, the rPPG module analyzes temporal variations in facial colour to obtain physiological features associated with blood-volume changes. The audio-visual module extracts speech audio and mouth-region information to assess synchronization between spoken content and visible lip movements.

The outputs of the analysis streams are combined using a feature- or score-level fusion mechanism. The fused information is passed to a classification module, which generates a prediction indicating whether the video is real or potentially manipulated. The system may also display a confidence score alongside the prediction.

**The overall workflow consists of the following stages:**

**Video Input → Video Preprocessing → Face Detection → CNN-LSTM Analysis, rPPG Analysis, and Lip-Sync Verification → Feature Fusion → Classification → Result Display**

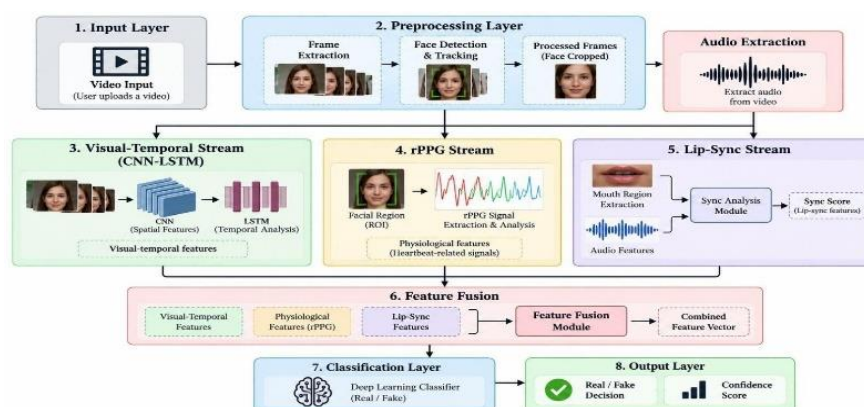


Fig: System Architecture

#### D. Technologies Used

The proposed deepfake video detection system uses computer vision, deep learning, signal processing, and audio analysis technologies to process video content and examine potential manipulation. The technologies and their intended purposes are summarized in Table I.

| Technology         | Purpose                                                               |
|--------------------|-----------------------------------------------------------------------|
| Python             | Main programming language for implementing the system.                |
| OpenCV             | Video reading, frame extraction, and image processing.                |
| TensorFlow/PyTorch | Development and execution of deep learning models.                    |
| NumPy              | Numerical operations and array processing.                            |
| SciPy              | Signal processing and filtering of physiological signals.             |
| Librosa            | Audio processing and feature extraction.                              |
| FFmpeg             | Audio and video extraction and conversion.                            |
| MediaPipe/MTCNN    | Face detection and facial-region localization.                        |
| Scikit-learn       | Machine learning utilities and evaluation metrics.                    |
| Flask/Streamlit    | Development of the user interface for video input and result display. |
| Matplotlib         | Visualization of signals and analysis results.                        |

Table I. Technologies Used

#### E. Implemented Modules

The proposed deepfake video detection system is organized into functional modules that process video data and examine different forms of evidence for potential manipulation.

1. **Video Input and Preprocessing:** Accepts the input video, validates the file, extracts selected frames, and prepares them for subsequent analysis.
2. **Face Detection:** Identifies facial regions in the selected frames and extracts the relevant areas for further processing.
3. **CNN-Based Visual Feature Extraction:** Extracts spatial facial features, including texture patterns, facial boundaries, and potential visual inconsistencies.
4. **LSTM-Based Temporal Analysis:** Processes sequences of extracted features to examine temporal relationships and inconsistencies across consecutive frames.
5. **rPPG Analysis:** Examines subtle facial colour variations over time to derive physiological features associated with blood-volume changes. Signal quality may be affected by lighting, motion, and video compression.
6. **Audio Extraction and Lip-Sync Verification:** Extracts the audio track and analyzes its correspondence with visible mouth movements to identify potential synchronization inconsistencies.
7. **Feature Fusion and Classification:** Combines the outputs of the analysis streams and uses the resulting information to classify the video as real or potentially manipulated. A confidence score may also be generated to communicate the model's prediction.

#### F. Testing Methodology

Testing is performed to evaluate the functional correctness and reliability of the proposed multi-stream deepfake video detection system. The testing methodology comprises the following stages.

1. **Unit Testing:** Individual components, including video preprocessing, face detection, CNN-based feature extraction, LSTM temporal analysis, rPPG signal processing, and audio extraction, are tested independently to verify their expected functionality.
2. **Integration Testing:** The interaction between connected modules is examined to verify that extracted frames, facial features, physiological signals, and audio-visual information are passed correctly to subsequent processing stages. The integration process also examines feature fusion and classification.
3. **System Testing:** The complete workflow is evaluated from video input and validation through preprocessing, multi-stream analysis, classification, and result display. Test cases include valid video input, invalid file formats, videos without detectable faces, and videos with audio suitable for lip-sync analysis.
4. **Performance Evaluation:** If a labeled test dataset and corresponding predictions are available, the system can be evaluated using accuracy, precision, recall, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). Processing time may also be measured to assess computational efficiency.

The testing process is intended to identify functional errors, integration issues, and limitations in video processing and classification. The reliability of the resulting predictions depends on the quality and diversity of the evaluation data, the recording conditions, and the manipulation techniques represented in the dataset.

### III. RESULTS AND DISCUSSION

#### A. Functional Testing

Functional testing was conducted to examine the operation of the developed deepfake detection application. The tests covered video uploading, video preprocessing, face detection, rPPG signal visualization, lip-motion analysis, and final

| Test ID | Test Case             | Input                  | Expected Result          | Output                                    |
|---------|-----------------------|------------------------|--------------------------|-------------------------------------------|
| TC01    | Upload valid video    | MP4                    | Video accepted           | Video uploaded successfully               |
| TC02    | Upload invalid file   | TXT                    | Error displayed          | Invalid file format message               |
| TC03    | Detect face           | Face video             | Face detected            | Facial region detected                    |
| TC04    | No face               | Empty/background video | Warning displayed        | No face detected warning                  |
| TC05    | Frame extraction      | Valid video            | Frames generated         | Video frames extracted                    |
| TC06    | CNN processing        | Face frames            | Features generated       | Spatial features generated                |
| TC07    | LSTM processing       | Frame sequence         | Temporal score generated | Temporal features/score generated         |
| TC08    | rPPG processing       | Facial video           | Signal generated         | rPPG signal generated                     |
| TC09    | Audio extraction      | Video with audio       | Audio extracted          | Audio track extracted                     |
| TC10    | Lip-sync processing   | Talking video          | Sync score generated     | Lip-sync score generated                  |
| TC11    | Feature fusion        | Multiple scores        | Combined score generated | Fused detection score generated           |
| TC12    | Real video detection  | Genuine video          | Real prediction          | Real + confidence score                   |
| TC13    | Fake video detection  | Deepfake video         | Fake prediction          | Fake+confidence score                     |
| TC14    | Result display        | Processed video        | Result displayed         | Prediction and analysis details displayed |
| TC15    | Long video processing | 5+ minute video        | System handles video     | Video processed successfully              |

Table II: Testing Of Test-Cases

## B. Results Summary

The application produced classification outputs for the processed videos and displayed supporting analysis information. The dashboard recorded 17 predictions, comprising 12 Real and 5 Fake classifications. The example Real prediction shown in Fig. 4 displays a confidence score of 98.1%, while the example Fake prediction is reported with a confidence score of 86.8%. These values describe individual application outputs and should not be interpreted as overall detection accuracy.

These values represent outputs displayed by the application during testing. They indicate that the application generated predictions for the tested inputs but do not establish its overall detection accuracy or generalization performance.

## C. Observed Application Functionality

The following functionalities were observed in the developed application:

1. **Video Upload:** The interface allows users to upload video files for analysis.
2. **Facial Analysis:** The application displays detected facial regions for further processing.
3. **rPPG Visualization:** A heartbeat-related signal is displayed as a waveform for physiological analysis.
4. **Lip-Motion Analysis:** The application presents visual information related to mouth movements and lip-motion analysis.
5. **Classification Output:** The system displays a Real or Fake prediction accompanied by a confidence score.
6. **Dashboard:** The application maintains a summary of predictions and displays previously processed video results.

The observed outputs demonstrate the application's video-analysis workflow and result-display functionality. However, further evaluation using a labelled test dataset and standard performance metrics is required to establish the reliability of its classifications.

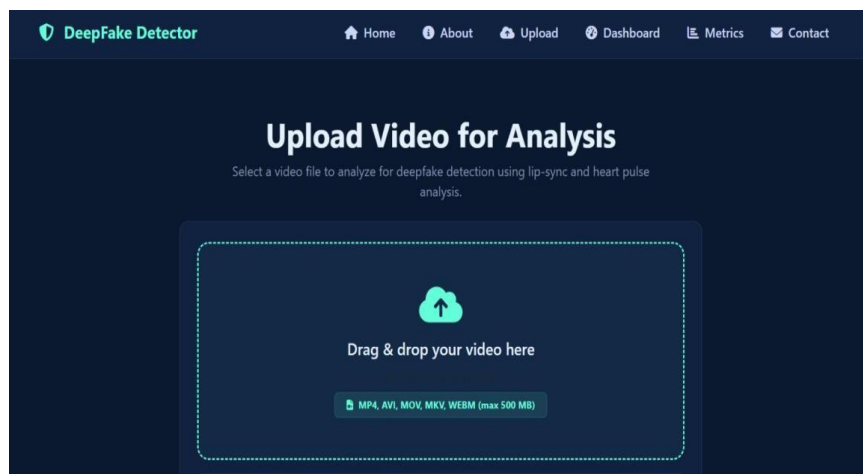


Fig. 1. Video upload interface for deepfake analysis.

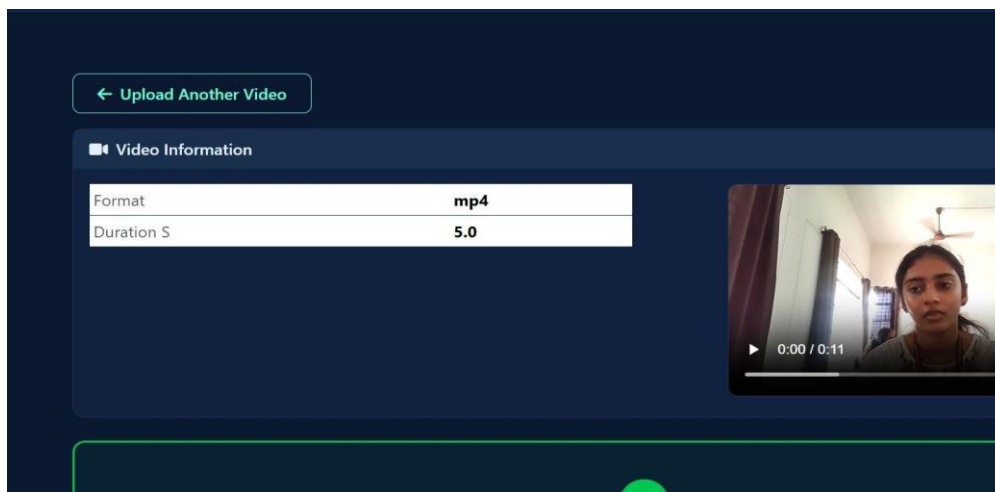


Fig. 2. Example Fake classification output with confidence score.



Fig. 3. rPPG pulse-signal visualization from the input video.

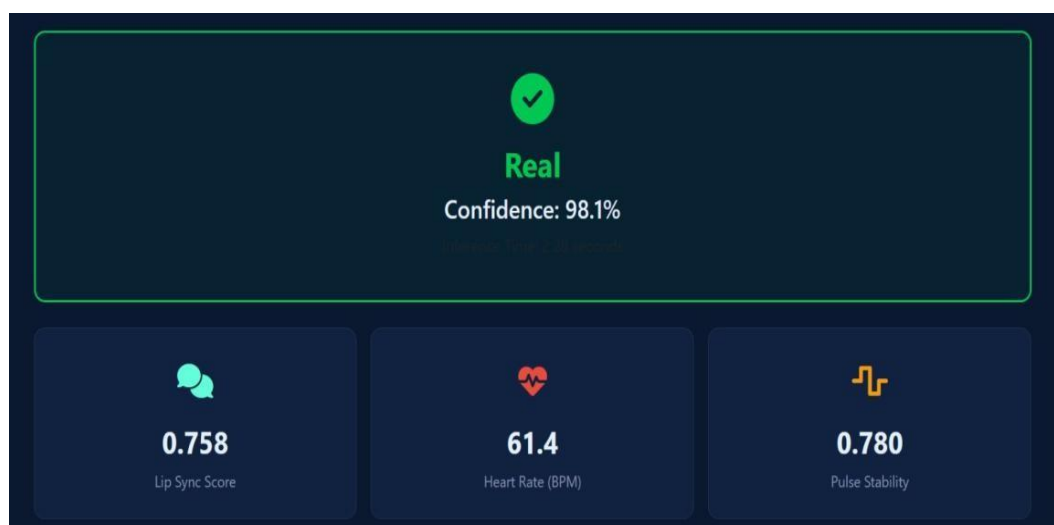


Fig. 4. Example Real classification output with confidence score.

#### D. Discussion

The developed deepfake detection application demonstrates a multi-stream approach to video authenticity assessment by presenting facial analysis, rPPG signal visualization, lip-motion analysis, and final classification outputs. The application provides users with a Real/Fake prediction and an associated confidence score, along with supporting information from the analysis modules.

The displayed results illustrate the application's ability to process video inputs and present different forms of analytical information. Integrating visual, physiological, and audio-visual evidence provides a broader analysis approach than relying

exclusively on facial appearance. However, the displayed predictions alone do not establish that combining these signals improves detection accuracy.

The reliability of rPPG analysis may be affected by lighting conditions, facial movements, video resolution, and compression. Similarly, lip-sync analysis depends on the quality of the audio track and the correspondence between speech and mouth movements. These factors may influence the information available to the classification system.

Although the application displays confidence scores for individual predictions, further quantitative evaluation using a labelled dataset is necessary to determine its overall performance. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC would help assess its ability to distinguish genuine videos from manipulated ones.

Overall, the developed application demonstrates the practical implementation of a multi-stream deepfake analysis workflow. Further evaluation across diverse video sources and manipulation techniques is required to establish its robustness and reliability for digital media authenticity assessment.

### IV. CONCLUSION

This paper presents a multi-stream approach to deepfake video detection that combines CNN-LSTM-based visual-temporal analysis, remote Photoplethysmography (rPPG), and audio-visual lip-sync verification. The approach examines facial features, temporal inconsistencies, heartbeat-related signals, and synchronization between speech and mouth movements to support video authenticity assessment.

The developed application demonstrates video uploading, facial analysis, rPPG signal visualization, lip-motion analysis, and Real/Fake classification with confidence scores. By presenting complementary forms of analytical information, the system provides a broader approach to deepfake detection than visual analysis alone.

However, the reliability of the system may be influenced by video quality, lighting conditions, facial movements, compression, and audio quality. The displayed predictions do not independently establish overall detection accuracy. Further evaluation using labelled datasets and standard performance metrics is required to assess the system's effectiveness and generalization across different deepfake techniques.

Future work may focus on improving feature fusion, strengthening the robustness of physiological and audio-visual analysis, evaluating performance across diverse datasets, and exploring advanced deep learning architectures. These improvements could help develop a more reliable deepfake detection system for digital media authenticity assessment.

### REFERENCES

1. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11. <https://doi.org/10.1109/ICCV.2019.00009>.
2. W. Chen and D. McDuff, "DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks," Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 356–373. [https://doi.org/10.1007/978-3-030-01216-8\\_22](https://doi.org/10.1007/978-3-030-01216-8_22).
3. Y. Zhou and S.-N. Lim, "Joint Audio-Visual Deepfake Detection," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14800–14809. <https://doi.org/10.1109/ICCV48922.2021.01453>.