*As per UGC guidelines an electronic bar code is provided to secure your paper*

*International Journal for Modern Trends in Science and Technology*
*Volume 12, Issue 04*, pages 40-44.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue04.html
DOI: https://doi.org/10.5281/zenodo.19321956

# Human Deepfake Video Detection using Automated Deep Learning Models

**K.S.Bhanu Rekha, Gadamsetty Sai Lokesh, Kollapalli Kalyan, Janjarapu Varshithkar, Alahari Govinda Sai Vigneswar, Shaik Sharu**

*Department of CSE-AI, PBR Visvodaya Institute of Technology and Science, Kavali, A.P, India*

**To Cite this Article**
K.S.Bhanu Rekha, Gadamsetty Sai Lokesh, Kollapalli Kalyan, Janjarapu Varshithkar, Alahari Govinda Sai Vigneswar & Shaik Sharu (2026). Human Deepfake Video Detection using Automated Deep Learning Models. International Journal for Modern Trends in Science and Technology, 12(04), 40-44. https://doi.org/10.5281/zenodo.19321956

**Article Info**
Received: 28 February 2026; Revised: 18 March 2026; Accepted: 22 March 2026.



***KEYWORDS***

*Deepfake Detection, CNN, LSTM, Video Forensics, Artificial Intelligence, Computer Vision*

**ABSTRACT**

*The widespread availability of powerful generative models has made it possible to create highly convincing forged videos in which human facial appearance and expressions are synthetically altered. Such manipulated media threatens digital trust, privacy, and the reliability of online information. This work presents an automated deep learning framework for identifying forged human videos through joint spatial and temporal analysis. In the proposed approach, a Convolutional Neural Network is employed to learn discriminative facial representations from individual frames, while a Long Short-Term Memory network models motion patterns across frame sequences. The system focuses on facial region extraction, landmark behavior, texture inconsistencies, and irregular motion cues to distinguish authentic content from manipulated samples. Experimental evaluation demonstrates that the hybrid architecture achieves better detection performance than methods that rely only on frame- level analysis, while remaining suitable for practical deployment.*

## 1.INTRODUCTION

Recent progress in deep neural networks has enabled the synthesis of highly realistic videos in which a person's identity, expressions, or speech can be artificially modified. Although this technology has beneficial applications in media production and virtual content creation, it can also be misused for spreading false information, impersonation, and financial fraud. Compared with manipulated images, detecting forged videos is more complex because it requires the analysis of both visual appearance and motion continuity. Visual artifacts may appear as blending errors or abnormal textures, whereas temporal abnormalities can be observed in eye blinking patterns, head movement dynamics,or the continuous human lip synchronization.To address these challenges, this work

introduces a unified spatial– temporal learning framework that combines CNN- based feature extraction with LSTM-based sequence modeling to improve robustness against modern deepfake generation techniques.

## NEED FOR STUDY

The rapid growth of social media platforms has significantly increased the circulation of synthetic videos that are difficult to verify manually. Easy access to face-swapping and reenactment tools allows the creation of realistic forged content without specialized expertise. This creates serious concerns related to cybersecurity, identity protection, and the authenticity of digital evidence. Conventional verification techniques cannot cope with the scale and complexity of such data. Therefore, an intelligent and automated detection mechanism that can learn both appearance-based and motion-based inconsistencies is essential for real-world applications.

## EXISTING SYSTEM

Many current deepfake detection approaches operate at the frame level by applying convolutional networks to classify individual images extracted from videos. These methods are effective in identifying visual artifacts such as boundary distortions, inconsistent illumination, and abnormal textures introduced during manipulation.Some techniques also employ frequency-domain analysis or transfer learning using pre-trained models.However, analyzing frames independently ignores the temporal relationship between consecutive frames. Since forged videos are generated as continuous sequences, the absence of motion modeling often leads to misclassification. In addition, models trained on a single dataset may fail when tested on content produced using different generation methods or varying compression levels.

Disadvantages

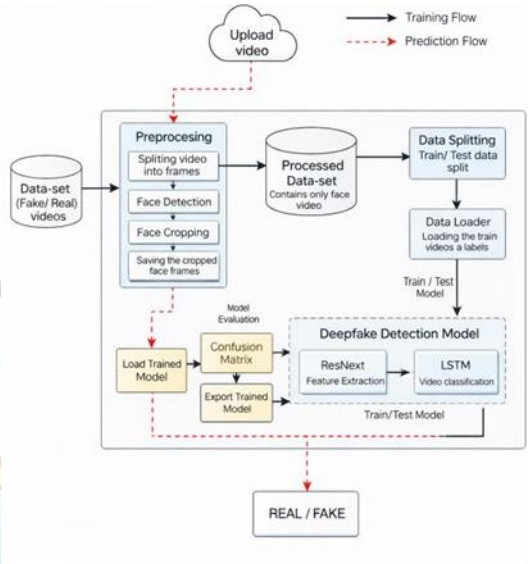
- Lack of Temporal Analysis:

Most traditional methods analyze videos frame by frame and fail to capture temporal inconsistencies such as unnatural blinking, irregular head movements, or lip synchronization errors..

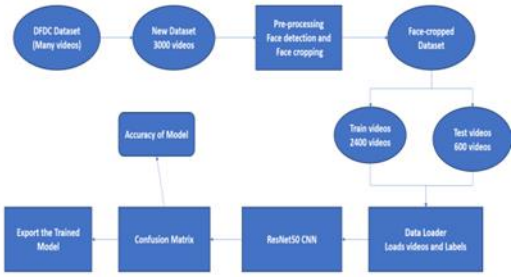
- Some detection models require high computational resources, making real-time deployment difficult..
- Sensitivity to Compression and Noise: Existing systems often struggle with low-resolution or heavily compressed videos commonly found on social media platforms.
- Limited Robustness Against Advanced GANs:

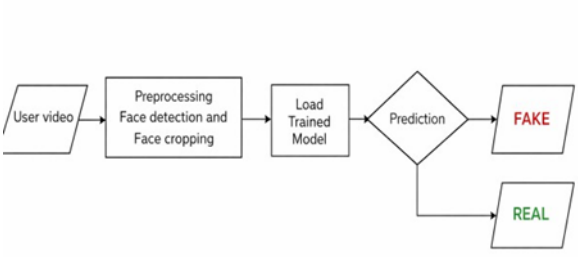
Modern deepfake generation models produce highly realistic outputs, reducing the effectiveness of traditional artifact-based detection methods.

## SYSTEM ARCHITECTURE




DATAFLOW DIAGRAM OF PROPOSED SYSTEM




## MODULES

1. Dataset Collection Module

This module is responsible for collecting real and fake video datasets used for training and testing the model. Publicly available deepfake datasets are utilized to

ensure diversity in facial expressions,lighting conditions, and manipulation techniques. The dataset is labeled as real or fake for supervised learning.

2. Preprocessing Module

The preprocessing module prepares raw videos for model training. It includes:

- Video frame extraction
- Face detection using computer vision techniques
- Face cropping and alignment
- Frame resizing and normalization

This step ensures that only relevant facial regions are processed, reducing noise and improving model performance.

3. Data Splitting Module

The processed dataset is divided into training and testing sets. Typically, a major portion of the data is used for training, while the remaining portion is reserved for performance evaluation. This separation ensures unbiased model validation.

4. Feature Extraction Module

In this module, a Convolutional Neural Network (CNN) is used to extract spatial features from facial frames. The CNN captures important visual patterns such as texture inconsistencies, blending artifacts, and abnormal facial details introduced during manipulation.

5. Temporal Modeling Module

To analyze sequential dependencies across frames, a Long Short-Term Memory (LSTM) network is used. This module captures temporal inconsistencies such as unnatural blinking, irregular head movements, and lip synchronization errors.

6. Classification Module

The extracted spatial and temporal features are passed to a fully connected classification layer. The model outputs a binary prediction indicating whether the input video is real or fake.

7. Model Evaluation Module

This module evaluates model performance using metrics.

**PROPOSED SYSTEM**

The proposed framework performs deepfake detection through a sequence of automated stages. Initially, real and manipulated video samples are collected and converted into frame sequences. Face localization and alignment are applied so that the model focuses only on the most informative region. A deep convolutional network is then used to transform each processed frame into a compact feature representation that captures subtle visual irregularities. To incorporate motion information, these feature vectors are arranged as temporal sequences and passed to an LSTM network.This component learns variations in facial dynamics such as eye movement frequency, head pose continuity, and speech-related lip motion. The combined representation is finally fed into a fully connected layer that produces a binary decision indicating whether the input video is authentic or forged.By learning both spatial and temporal characteristics simultaneously, the system improves detection reliability and generalizes better to unseen manipulations.

Advantages

- Combines spatial (CNN) and temporal (LSTM) feature learning
- Improved accuracy for high-quality deepfake detection
- Reduced false positives and false negatives
- Better generalization to unseen deepfake techniques
- Robust performance on compressed and low-resolution videos
- Fully automated and scalable framework
- Suitable for real-time and forensic applications
- Capable of capturing subtle facial motion irregularities
- Adaptable to evolving deepfake generation models through retraining

Hardware Requirements

- Processor: Intel i5 / AMD Ryzen 5 or higher
- RAM: Minimum 8 GB (Recommended: 16 GB)
- Storage: Minimum 30 GB free disk space
- GPU (Optional): NVIDIA GPU with CUDA support (for faster training)
- Display: Standard monitor with minimum 1366×768 resolution

Software Requirements:

- Operating System: Windows /Linux / macOS
- Programming Language: Python 3.x
- Development Environment:VS Code/Jupyter Notebook.

## ALGORITHM

Input:Real and deepfake video datasets containing humans.

Output: Classification result (REAL / FAKE) 1:
Load the real and deepfake video datasets.

2:Import the required libraries including packages like the TensorFlow/Keras, OpenCV, NumPy, and others. 3: Split the dataset into training and testing sets in an 80:20 ratio.

4: Extract frames from each video in the dataset.
5: Perform preprocessing on extracted frames:

- Detect facial regions
- Crop detected faces
- Resize images to required input dimensions
- Normalize pixel values $_{it}$

6: Apply a pre-trained ResNet50 model to extract spatial facial features from each processed frame.

7: Train the deep learning model using the extracted features from the training dataset.

8: Evaluate the trained model using the testing dataset and compute performance metrics such as accuracy and confusion matrix.

9: For a new input video, perform frame extraction and preprocessing, then classify the video as REAL or FAKE using the trained model.

## MATHEMATICAL FORMULATION OF THE PROPOSED SYSTEM

Let the dataset be defined as:

$$D = \{(V_i, y_i)\}^N$$

where:

- $V_i$represents the $i^{th}$video
- $y_i \in \{0,1\}$denotes the label (0 = Real, 1 = Fake)
- $N$is the total number of videos

**1.** *Dataset Splitting* $\hat{y}_i = \{$

The dataset is divided into training and testing sets:

$$D = D_{train} \cup D_{test}$$

such that:

$$| D_{train} | = 0.8N, | D_{test} | = 0.2N$$

**2.** *Frame Extraction*

Each video $V_i$is represented as a sequence of frames:

$$V_i = \{F_{i1}, F_{i2}, \dots, F_{iT}\}$$

where $T$is the number of frames in the video.

3. Preprocessing

Facedetection and croppingproduce processed frames: $Xit$ = Preprocess($Fit$)

Normalization is applied as:

$$X' = \frac{X_{it} - \mu}{\sigma}$$
$$it$$

where $\mu$and $\sigma$denote mean and standard deviation.

**3.** *Spatial Feature Extraction (ResNet50)*

Spatial features are extracted using a CNN model:

$$f_{it} = CNN(X')$$

where $f_{it} \in \mathbb{R}^d$represents the feature vector of dimension $d$.

For a video sequence:

$$F_i = \{f_{i1}, f_{i2}, \dots, f_{iT}\}$$

**4.** *Temporal Modeling (LSTM)*

The sequence of feature vectors is passed to the LSTM:

$$h_t = LSTM(f_{it}, h_{t-1})$$

The final hidden state $h_T$represents the temporal embedding of the video

**5.** *Classification*

The final prediction probability is computed as:

$$\hat{y}_i = \sigma(Wh_T + b)^{i=1}$$

where:

- $W$and $b$are learnable parameters
- $\sigma$is the sigmoid function

Decision rule:

$$1 \quad \text{if } \hat{y}_i > 0.5$$
$$0 \quad \text{otherwise}$$

**6.** *Loss Function*

Binary Cross-Entropy Loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i)]$$

**7.** *Optimization*

Model parameters $\theta$are updated using Adam optimization:

$$\theta_{t+1} = \theta_t - \alpha\nabla\theta\mathcal{L}$$

where is the learning rate.

Final Output

Output = {REAL,FAKE}

## CONCLUSION

In this work, a hybrid deep learning framework for human deepfake video detection was presented, combining CNN-based spatial feature extraction with LSTM-based temporal sequence modeling. The proposed system effectively captures both facial artifacts and motion inconsistencies, improving detection accuracy compared to frame-level approaches. By incorporating preprocessing techniques such as face detection and normalization, the model focuses on relevant facial regions and reduces noise. Experimental evaluation demonstrates that the spatial–temporal architecture enhances robustness, generalization capability, and reliability in identifying manipulated videos. The proposed system provides a scalable and automated solution suitable for real-world applications in digital forensics, media verification, and cybersecurity. Furthermore, the model architecture can be adapted to evolving deepfake generation techniques through continuous retraining with updated datasets. This adaptability ensures long-term effectiveness in combating emerging multimedia manipulation threats

## FUTURE ENHANCEMENT

Future improvements to the proposed deepfake detection system can focus on enhancing robustness, scalability, and real-world applicability. Transformer-based architectures such as Vision Transformers (ViT) can be integrated to capture long-range spatial–temporal dependencies more effectively. The system can be extended to a multimodal framework by incorporating audio analysis to detect voice cloning and lip–audio synchronization inconsistencies. Additionally, adversarial training techniques can be employed to improve resistance against sophisticated manipulation attacks. Optimizing the model using pruning and quantization methods will enable real- time deployment on edge devices and mobile platforms. Further research can also explore cross- dataset generalization and explainable AI techniques to increase transparency and trust in detection results

## Conflict of interest statement

Authors declare that they do not have any conflict of interest.

## REFERENCES


[1] [Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "SoK Deepfake Detection—A Systematic Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[2] S. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot… for Now," in Proc. IEEE/CVF Conf Computer Vision and Pattern Recognition (CVPR), 2022.

[3] Y. Liu, X. Wang, and Z. Zhao, "Generalizable Deepfake Detection via Self-Supervised Learning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023.

[4] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Multimodal Deepfake Dataset," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), 2022.

[5] Y. Zhou and S.-N. Lim, "Joint Audio-Visual Deepfake Detection Using Multimodal Transformers," IEEE Access, vol. 11, pp. 2023–2024.

[6] J. Chen, X. Tan, and L. Wang, "Vision Transformer-Based Deepfake Video Detection," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2023.

[7] M. Kim, S. Tariq, and S. Woo, "Fusing Spatial and Temporal Features for Robust Deepfake Detection," Pattern Recognition Letters, vol. 168, 2023.

[8] P. Yu, T. Zhao, and Y. Li, "Learning Temporal Inconsistencies for Deepfake Video Detection," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2022.

[9] R. Sun, Y. Li, and S. Lyu, "Detecting GAN Generated Media via Frequency Analysis," IEEE Transactions on Information Forensics and Security, 2022.

[10] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022.

[11] X. Jiang, Y. Zhang, and W. Wang, "Cross Dataset Generalization for Deepfake Detection," IEEE Transactions on Multimedia, 2023.

[12] L. Wang, Z. Chen, and H. Li, "Robust Deepfake Detection via Adversarial Training," in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), 2023.

[13] S. Agarwal and H. Farid, "Protecting World Leaders Against Deepfakes," IEEE Security & Privacy, vol. 21, no. 2, 2023.

[14] T. Jung and K. Kim, "Explainable Deepfake Detection Using Attention Maps," IEEE Access, vol. 12, 2024.

[15] Y. Zhang, F. Ding, and R. Zhao, "Multimodal Transformer Networks for Deepfake Video Detection," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023.Proceedings of the IEEE Intelligent Vehicles Symposium (IV), 2018, pp. 1773– 1780.