iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IV Month of publication: April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69233

Deepfake Video Face Detection

Prof. Amol Gosavi[1], Rahul A. Joshi[2], Sejal D. Shingote[3], Roshani K. Pawar[4], Harshil U. Kansagra[5]
[1]Professor, [2, 3, 4, 5]Student, Department of Computer Engineering, MET's Institute of Engineering, Nashik, Maharashtra, India

Abstract: The emergence of deepfake technology, which relies on generative adversarial networks (GANs), has raised substantial concerns in the realm of digital media. This technology enables the manipulation of facial features in videos, leading to potential misuse for spreading false information, misrepresentation, and identity theft. As a result, there is a pressing need to establish robust methods for detecting deepfakes effectively. Detecting deepfake videos is particularly difficult due to their increasingly realistic appearance and the sophisticated techniques involved in their creation. This research introduces an innovative approach to deepfake detection that leverages advanced deep learning methodologies. Specifically, the study employs Convolutional Neural Networks (CNNs) in combination with Recurrent Neural Networks (RNNs), with a particular focus on Long Short-Term Memory (LSTM) networks, to enhance the identification process for deepfake content. The proposed model is trained on comprehensive datasets, including FaceForensics++ and the Deepfake Detection Challenge (DFDC). To bolster detection accuracy, the methodology includes a pre-processing pipeline that not only reduces the frame rates of video inputs but also isolates and focuses on facial regions using Haar Cascade classifiers. This dual approach of analyzing both spatial and temporal inconsistencies within video frames contributes significantly to the overall effectiveness of deepfake detection. Through rigorous testing, the proposed method has demonstrated a high level of accuracy in distinguishing between authentic and manipulated videos, showcasing its potential as a reliable solution in the ongoing fight against digital media fraud. It is crucial for researchers and practitioners in the field of video forensics and digital media security to further explore and refine such advanced detection techniques.

Keywords: deepfake detection, CNN, LSTM, FaceForensics++, GANs, video forensics, digital media security.

I. INTRODUCTION

The emergence of artificial intelligence has presented significant challenges in the form of deepfake videos across the digital landscape. Deepfake technology leverages the power of Generative Adversarial Networks (GANs), which are sophisticated algorithms designed to generate hyper-realistic videos. These videos can involve swapping one person's face with that of another, resulting in content that is often highly deceptive and difficult to distinguish from reality. While there are genuine applications for deepfake technology in industries such as filmmaking and gaming, its potential for misuse raises alarming concerns regarding misinformation, cyber fraud, and the potential harm to individuals' reputations. The ability to easily manipulate video content makes it imperative to create effective automated tools for detecting deepfakes. Current detection methodologies predominantly focus on frame-by-frame analysis or the identification of inconsistencies in the time sequence of a video. However, with advancements in deepfake algorithms, it has become increasingly clear that these technologies can produce videos with an astoundingly low level of detectable artifacts. This evolution necessitates a more comprehensive approach to detection that combines both spatial feature extraction and temporal analysis. In light of these challenges, our study seeks to develop an innovative deep learning model specifically designed to address the nuances of deepfake detection. This model will harness the capabilities of Convolutional Neural Networks (CNNs) for extracting spatial features from the frames and utilize Long Short-Term Memory (LSTM) networks to analyze the sequences over time. By enabling the model to identify inconsistencies and irregularities that hint at potential deepfake manipulation, we aim to contribute valuable advancements to the burgeoning field of digital media integrity and security. Through this integrated approach, we hope to enhance the reliability of deepfake detection methods and mitigate the risks associated with deceptive digital content.

II. OBJECTIVES

1) Design and implement a deep learning model that can effectively differentiate between real and manipulated videos using CNNs, RNNs/LSTMs.
2) Identify inconsistencies in facial features, expressions, and frame transitions to improve the model's accuracy and robustness against various deepfake generation techniques.
3) Integrate the trained model into a web-based or mobile application for real-time detection, ensuring accessibility for security agencies, media platforms, and general users.

III.LITERATURE REVIEW

1) A comprehensive review of 112 methods for detecting deepfakes categorized existing techniques into four main groups: deep learning, machine learning, statistical approaches, and blockchain-based methods. The analysis revealed that deep learning techniques, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), excelled in identifying deepfake videos. Furthermore, a detailed framework was proposed to categorize these detection techniques, highlighting the advantages and limitations of each. A significant drawback noted was the absence of a standardized evaluation framework, resulting in discrepancies in datasets, assessment metrics, and the availability of original videos.
2) Another investigation centered on the use of transformer-based architectures for detecting deepfake videos by examining spatial-temporal inconsistencies across multiple frames. In contrast to traditional CNN models, transformers are capable of capturing long-range dependencies and subtle facial distortions, which greatly enhance detection accuracy. The results indicated that transformer-based methods outperformed conventional approaches; however, challenges persisted, particularly regarding their limited integration with existing detection methods. Additionally, the variability in performance across different datasets emphasized the need for ongoing optimization and better generalization.
3) A separate study examined the integration of multi-view learning techniques with adversarial training to improve deepfake detection capabilities. By allowing the model to analyze videos from various perspectives, multi-view learning enhanced detection accuracy across a range of manipulation types. Adversarial training further bolstered the model's resilience against evolving deepfake techniques, contributing to a more robust detection system. Nonetheless, this approach required additional optimization for real-time applications and faced scalability challenges when applied to large video datasets, complicating practical implementation
4) Another proposed method utilized multi-scale residual networks combined with attention mechanisms for deepfake detection. The residual networks facilitated superior feature extraction across multiple levels, while the attention mechanisms concentrated on key facial regions that were most impacted by manipulations. This strategy resulted in improved detection accuracy by emphasizing the most significant features of altered videos. However, this method showed limitations in terms of robustness against adversarial deepfake techniques, and further enhancements were necessary to improve processing efficiency for real-time use.
5) This study introduces a novel approach combining multi-scale convolutional neural networks (CNNs) with vision transformers to enhance deepfake detection. The multi-scale convolution captures features at various resolutions, while the vision transformer models long-range dependencies, resulting in improved detection performance.
6) This paper explores the detection of deepfake videos by leveraging spatiotemporal inconsistencies generated during the deepfake creation process. The proposed method integrates spatial and temporal features through an interactive fusion mechanism, effectively identifying manipulated content.
7) This research introduces Adversarial Feature Similarity Learning (AFSL), a method that enhances the robustness of deepfake detection models against adversarial attacks. By optimizing the similarity between adversarially perturbed and unperturbed examples, the approach aims to distinguish real from fake instances more effectively.

IV.PROPOSED SYSTEM

The proposed Deepfake Video Detection System is an advanced, AI-based framework designed to identify and address the growing threat of deepfake videos. This system utilizes computer vision and deep learning techniques to analyze video frames, detect inconsistencies, and assess the authenticity of facial content. The architecture includes a pre-trained deep learning model, such as XceptionNet, which has been extensively trained on both real and manipulated videos to accurately distinguish between genuine and fake facial expressions.

The detection process consists of several key steps: video pre-processing, frame extraction, facial region identification, and deep feature analysis. This comprehensive approach ensures a robust and efficient method for identifying deepfakes. The system aims to enhance media authentication, making it a crucial tool for digital forensics, cybersecurity, and misinformation prevention. By automating the detection process, it reduces the need for manual inspection, offering a scalable and efficient solution for various applications, including social media moderation, law enforcement, and content verification platforms.

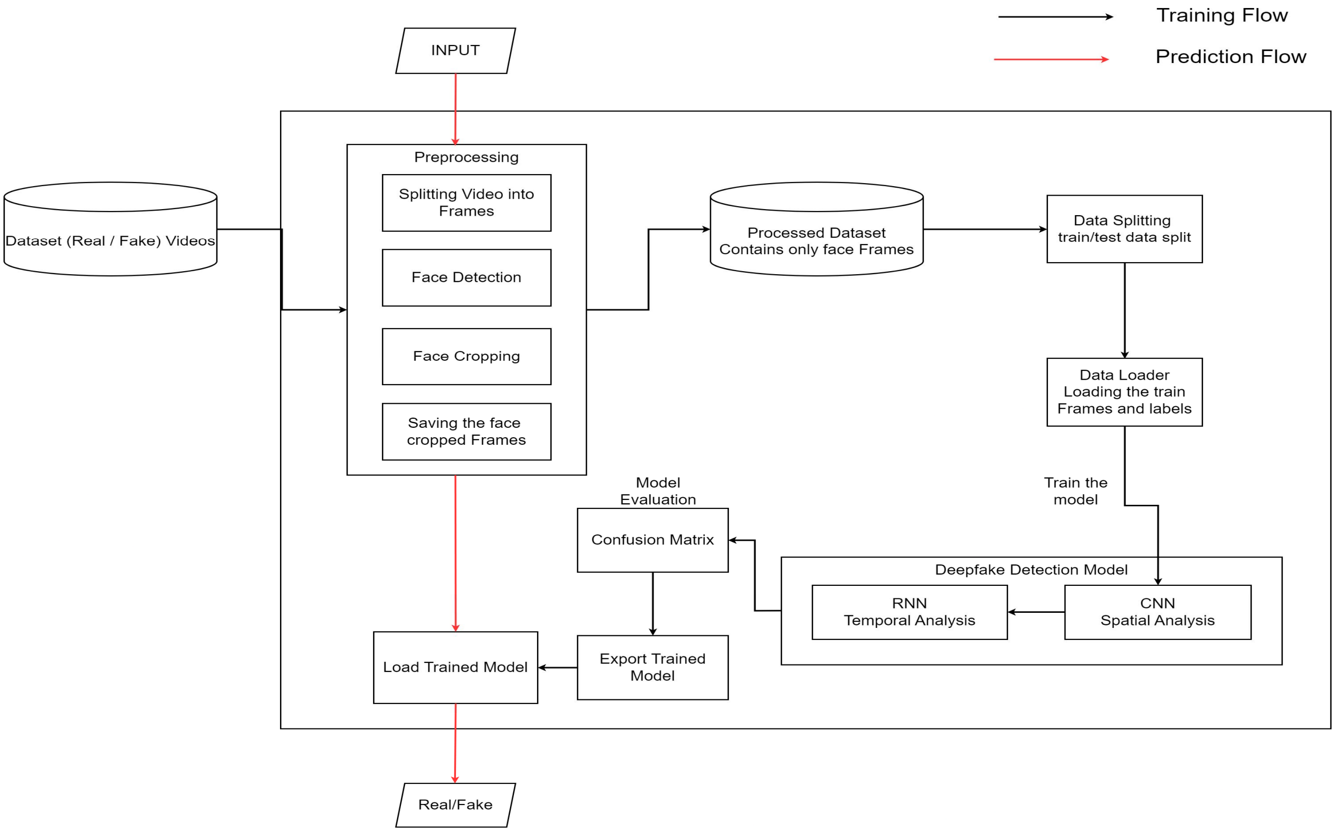


Fig. 1 Proposed System of Deepfake Video Face Detection

V. WORKING

The proposed deepfake detection system is designed to analyze video inputs and classify them as either real or fake by utilizing deep learning techniques, spatial-temporal feature extraction, and hybrid neural network models. The working of the system follows a multi-stage pipeline that ensures accurate detection of face-swapped manipulations.

A. Data Preprocessing and Frame

Extraction Before analyzing a video for deepfake detection, it undergoes a preprocessing stage to optimize its quality. Since videos consist of a sequence of frames, processing each frame individually can be computationally intensive. Therefore, we implement the following steps: -

1) Frame Rate Reduction: Videos from various sources often have different frame rates (e.g., 60 fps, 30 fps). Processing every frame is inefficient, so the system reduces the frame rate from 60 fps to 30 fps without losing essential temporal information. This ensures that only relevant frames are analyzed while maintaining computational efficiency.
2) **Face Detection and Cropping: Deepfake manipulations mainly affect facial regions, so the system first extracts the face from each frame. A Haar Cascade Classifier is used for face detection, allowing the model to focus exclusively on relevant facial features rather than background details. Once detected, the face is cropped and resized to a fixed dimension (e.g., 299x299 pixels) to ensure uniform input across all video frames. –
3) Normalization and Data Augmentation: To enhance the model's performance, pixel values are normalized to a fixed range (0 to 1) before being input into the model. Data augmentation techniques, such as rotation, flipping, and brightness adjustments, are applied to increase the system's robustness against variations in video input.

B. Feature Extraction Using Convolutional Neural Networks (CNNs)

The next stage involves extracting critical spatial features from individual frames using a deep CNN-based model called XceptionNet.

1) Why XceptionNet?

XceptionNet is a modified version of InceptionNet, designed to capture finer spatial details. It employs depthwise separable convolutions, improving computational efficiency while extracting complex patterns. This model is particularly effective at identifying inconsistencies in texture, lighting, and blending artifacts, which are common in deepfake videos.

2) How CNN Extracts Features?

Each frame is processed through multiple convolutional layers, where filters detect edges, facial structures, and intricate details. Activation functions (ReLU, Softmax) help in determining whether a pixel region corresponds to a natural or manipulated facial texture. The CNN identifies spatial-level discrepancies such as blurred edges, unnatural skin tones, asymmetrical facial features, and inconsistent shadows, all of which may indicate potential deepfake manipulation.

C. Temporal Feature Analysis Using Recurrent Neural Networks (RNNs) and LSTMs

Unlike traditional CNN models that focus solely on individual frames, our system incorporates RNNs and LSTMs to analyze the temporal relationships between frames.

1) Why Temporal Analysis is Important?

Deepfake videos often lack consistency across frames, resulting in unnatural facial expressions, jittering effects, and inconsistent head movements. While a single frame may appear real, anomalies can become evident when examining multiple frames (e.g., irregularities in eye blinking or unnatural lip movements). RNNs and LSTMs help capture these time-based inconsistencies.

2) How LSTMs Work in Our System?

The sequence of extracted spatial features from the CNN is passed to an LSTM-based RNN model. LSTMs maintain memory across frames, allowing them to track movements and detect unnatural transitions over time. The system analyzes facial expressions, lip-sync accuracy, and blinking patterns to determine whether the video is genuine or manipulated.

D. Classification and Deepfake Detection

Once the CNN (spatial analysis) and LSTM (temporal analysis) models extract features, the system classifies the video as real or fake.

1) Classification Using Fully Connected Layers

The extracted spatial and temporal features are fed into a fully connected dense layer, which aggregates learned patterns and makes a final classification. The final layer uses a Softmax activation function to generate a probability score indicating whether the video is real or fake. If the probability of being a deepfake exceeds a certain threshold (e.g., 0.5 or 50%), the video is classified as fake; otherwise, it is classified as real.

2) Performance Optimization Using Transfer Learning

Instead of training the model from scratch, the system leverages pre-trained CNN architectures like XceptionNet, which have already learned complex feature representations from large datasets. This approach improves model accuracy while reducing the need for extensive training data.

E. Web-Based Deployment for Real-Time Detection

After classification, the system is integrated into a Flask-based web application with an intuitive frontend built using Jinja templates.

1) User Interaction Flow

a) Video Upload: Users upload a video file to the web application.
b) Processing: The backend processes the video by extracting frames and passing them through the detection pipeline.
c) Result Generation: The system classifies the video and displays the result (e.g., "This video is a deepfake" or "This video is authentic").
d) Visualization: A probability score is presented, providing users with clarity on the classification outcome.

VI. RESULT

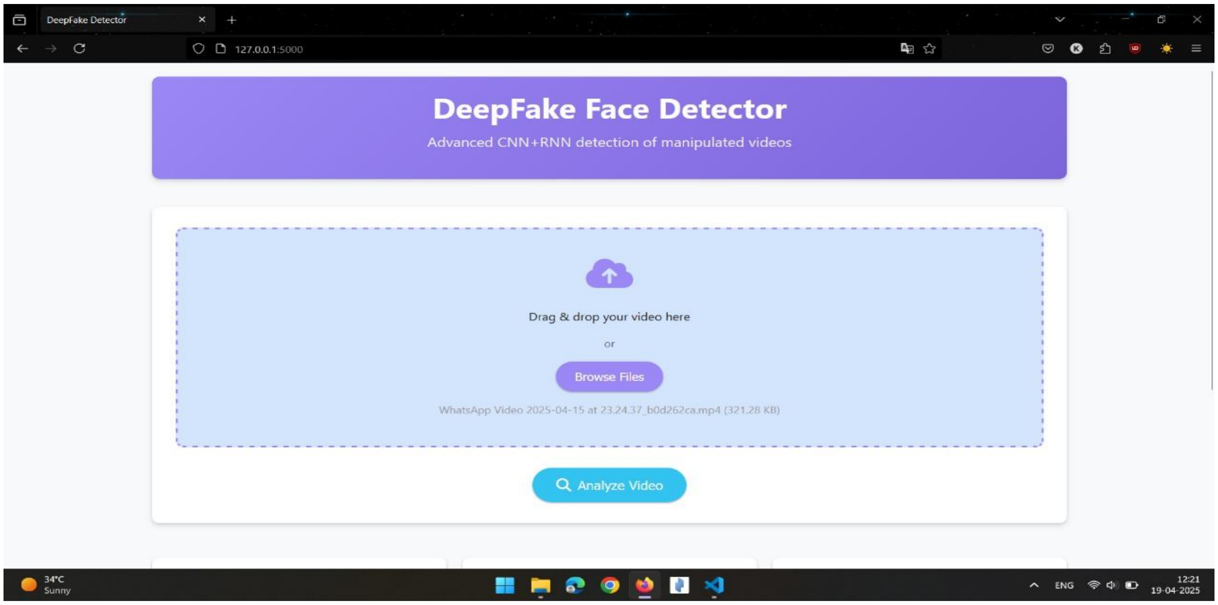
The proposed deepfake detection system was evaluated on various video samples containing both real and manipulated content. The system processes the input video, extracts facial regions, and applies deep learning models to classify individual frames as either "real" or "fake." The final classification is determined based on an aggregated confidence score computed across all frames.

A. User Interface for Deepfake Detection

1) Video Upload Interface

As shown in Figure 2, the initial screen allows users to upload a video for analysis. The interface includes:

- A clear heading for branding the system
- A file input field where users can browse and upload videos in supported formats (e.g., .mp4, .avi)
- A submit button to initiate the analysis process



DeepFake Detector

127.0.0.1:5000

DeepFake Face Detector

Advanced CNN+RNN detection of manipulated videos

Drag & drop your video here

or

Browse Files

WhatsApp Video 2025-04-15 at 23.24.37_b0d262ca.mp4 (321.28 KB)

Analyze Video

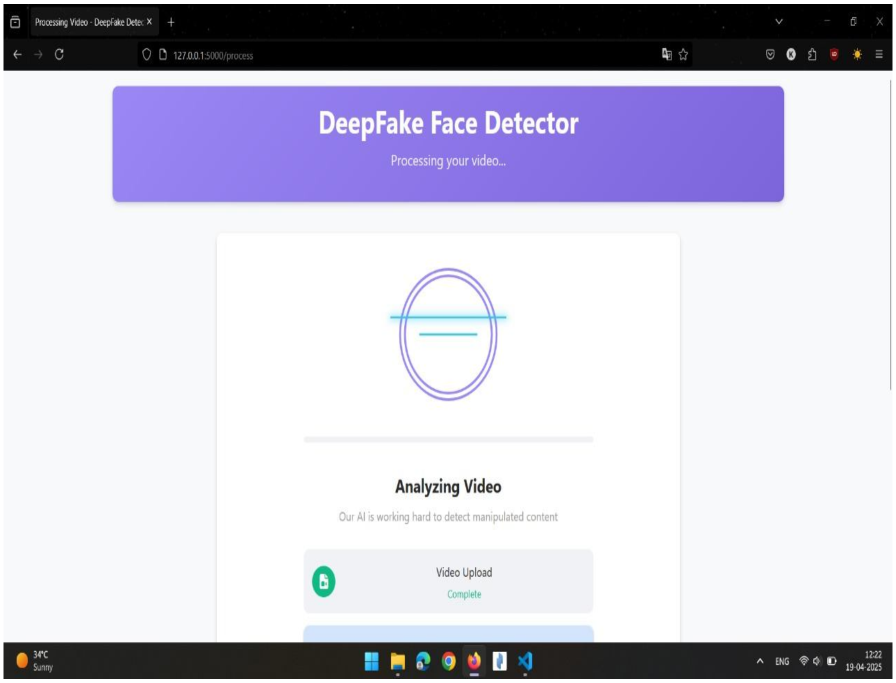
34°C Sunny

ENG 12:21 19-04-2025

Fig. 2 Video Upload Interface for Deepfake Detection System

2) Analysis Page

Once the video is uploaded, the system redirects the user to an analysis page. The backend performs frame extraction and applies a pretrained model (e.g., XceptionNet) to evaluate the authenticity of the video. As illustrated in Figure 3 & 4, these pages shows:



Processing Video - DeepFake Detec

127.0.0.1:5000/process

DeepFake Face Detector

Processing your video...

Analyzing Video

Our AI is working hard to detect manipulated content

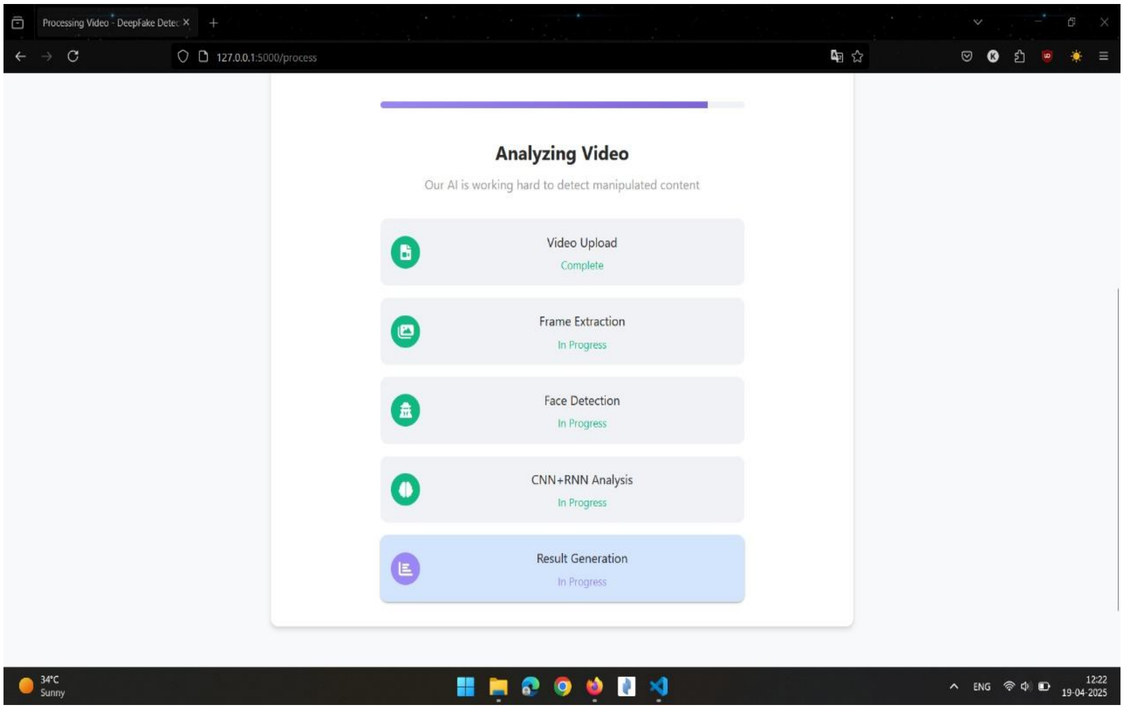
Video Upload

Complete

34°C Sunny

ENG 12:22 19-04-2025

Fig. 3 Video Analyze Page for Deepfake Detection System



Analyzing Video

Our AI is working hard to detect manipulated content

Video Upload
Complete

Frame Extraction
In Progress

Face Detection
In Progress

CNN+RNN Analysis
In Progress

Result Generation
In Progress

Fig. 4 Video Analyze Page for Deepfake Detection System

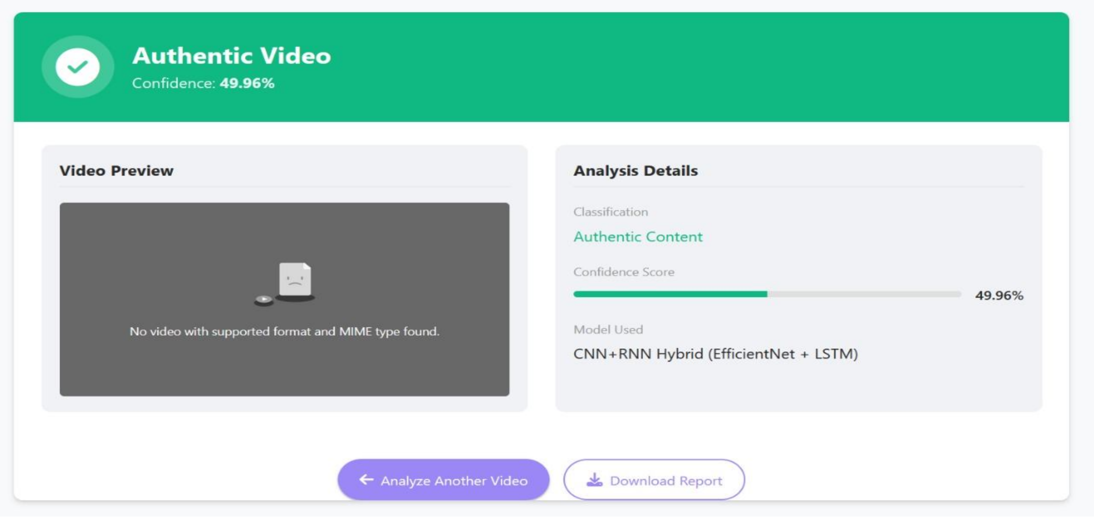
- A confirmation message indicating the analysis status
- The predicted result (Real or Fake)
- A confidence score (e.g.97.3%)
- The filename of the uploaded video

B. Detection Results with Confidence Scores

After processing the uploaded video, the system extracts facial frames, evaluates them using the deepfake detection model, and assigns confidence scores. Figures 3 and 4 show examples of the output results.

1) Case 1: Video Classified as Real

In Figure 5, the system calculates an average confidence score of 0.4996 and classifies the video as real. While some individual frames may have slightly higher fake scores, the majority are consistent with genuine content. The final decision is based on a threshold and the combined analysis of all frame-level predictions.



Authentic Video
Confidence: **49.96%**

Video Preview

No video with supported format and MIME type found.

Analysis Details

Classification
Authentic Content

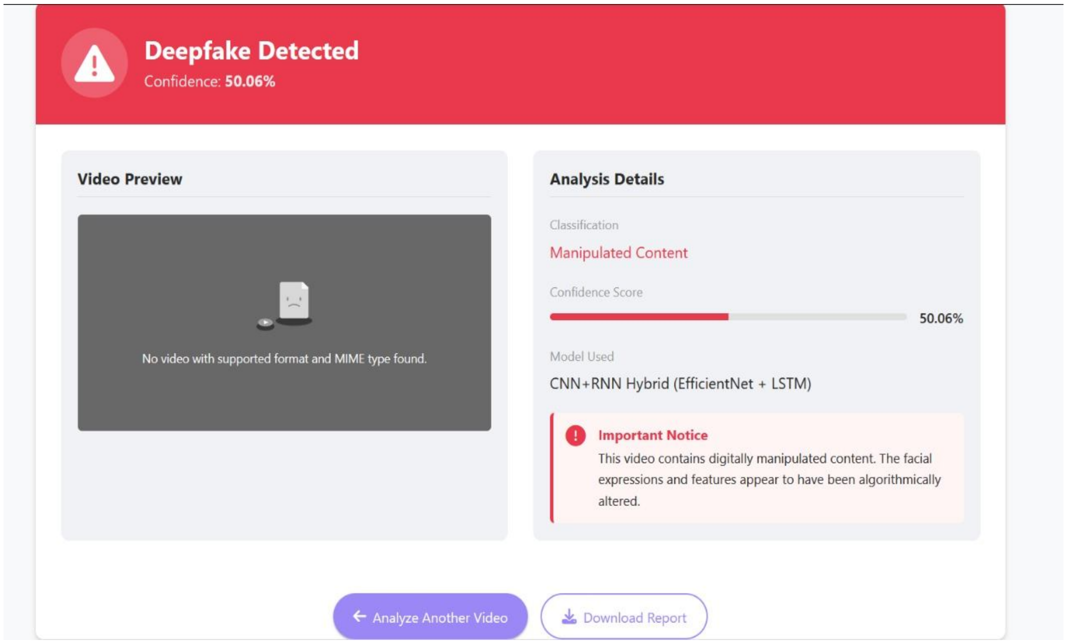
Confidence Score
49.96%

Model Used
CNN+RNN Hybrid (EfficientNet + LSTM)

Fig. 5 Example Output for a Real Video

2) *Case 2: Video Classified as Fake*

In Figure 6, the system calculates an average confidence score of 0.5006 and classifies the video as fake. A greater proportion of frames are detected with fake characteristics, leading to a result above the decision threshold. The system identifies subtle manipulations in facial movements and textures that are common in synthetic videos.



Deepfake Detected
Confidence: **50.06%**

Video Preview

No video with supported format and MIME type found.

Analysis Details

Classification
Manipulated Content

Confidence Score
50.06%

Model Used
CNN+RNN Hybrid (EfficientNet + LSTM)

Important Notice
This video contains digitally manipulated content. The facial expressions and features appear to have been algorithmically altered.

Fig. 6 Example Output for a Fake Video

C. *Interpretation of Results*

The model provides confidence scores for each detected face, showcasing the reliability of classification at a frame level.

The system is capable of identifying inconsistencies in facial features and expressions across frames, which are crucial for detecting deepfake videos.

Minor variations in confidence scores indicate the presence of real frames within manipulated videos, suggesting the need for a robust temporal analysis.

VII. FUTURE SCOPE

To enhance deepfake detection capabilities, several strategic advancements can be pursued in future research:

1) Improving Real-Time Detection- Fine-tuning models for immediate deepfake detection on platforms like YouTube and various social media channels. Minimizing computational demands to achieve quicker inference times while ensuring accuracy remains high.
2) Integrating Adversarial Defense Mechanisms - Employing adversarial training methods to strengthen the model's resilience against sophisticated deepfakes designed to bypass detection. - Utilizing generative adversarial networks (GANs) for counter-training purposes, enabling the model to keep pace with emerging deepfake tactics.
3) Expanding Training Datasets - Broaden the training sets to encompass high-resolution videos, represent diverse ethnic backgrounds, and include real-world deepfake scenarios for better generalization across various media. - Incorporate synthetic deepfake datasets to prepare the model for a wider array of manipulated video content.
4) Deploying on Social Media and News Platforms - Developing the detection system as a browser extension or plugin for social media that alerts users to possible deepfake videos. - Collaborating with fact-checking organizations and media outlets to integrate deepfake detection capabilities into existing content moderation frameworks.
5) Creating a Mobile Application - Designing a mobile-compatible version of the deepfake detection system that enables users to verify video authenticity directly on their smartphones. - Optimizing the model for use on low-power devices to enhance accessibility across various hardware setups.

By implementing these improvements, the deepfake detection system can transform into a comprehensive, automated real-time framework, offering a robust defense against threats posed by digital manipulation in media and cybersecurity fields.

VIII. CONCLUSION

Deepfake technology has emerged as a significant threat in the digital era, enabling the creation of highly realistic manipulated videos that can be used for misinformation, cybersecurity breaches, and digital fraud. As deepfake techniques continue to evolve, they pose serious risks to public trust and digital security. In response to this growing concern, this research proposes an advanced deepfake detection system that combines Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for analyzing temporal sequences. By integrating both spatial and temporal analysis, the model effectively detects inconsistencies in video frames and unnatural motion patterns, leading to improved accuracy in identifying deepfakes.

To enhance usability, the proposed system is deployed through a Flask-based web interface, allowing seamless interaction for users without technical expertise. This accessibility ensures that journalists, researchers, and the general public can efficiently verify the authenticity of video content. While the model demonstrates high accuracy in detecting various types of deepfakes, it faces challenges in identifying adversarial deepfakes designed to evade detection. Further optimizations and advancements in adversarial training will be required to enhance the system's robustness against sophisticated manipulation techniques.

Overall, this research makes a significant contribution to the field of deepfake detection by introducing a hybrid approach that effectively captures both spatial and temporal features. With ongoing improvements, this system has the potential to play a crucial role in combating the misuse of deepfake technology and preserving the integrity of digital media.

REFERENCES

[1] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, Andrew H. Sung, Deepfake Detection: A Systematic Literature Review, 2020.

[2] M. Wang, Z. Liu, H. Zhang, Video Deepfake Detection Using Transformers and Spatial-Temporal Features, 2023.

[3] A. Patel, R. Sharma, K. Gupta, Features Enhanced Deepfake Detection Using Multi-View Learning and Adversarial Training, 2023.

[4] S. Kim, J. Choi, Y. Jeong, DeepFake Video Detection Using Multi-Scale Residual Networks and Attention Mechanisms, 2023.

[5] Yong Wang, Zhen Cui, Jian Yang, DeepFake Detection with Multi-Scale Convolution and Vision Transformer, Digital Signal Processing, Vol.120, 2023.

[6] Xiaoming Li, Yibing Song, Spatiotemporal Inconsistency Learning and Interactive Fusion for Deepfake Video, ACM Transactions on Multimedia Computing, Communications, and Applications, Volume 19, Issue 1, February 2023

[7] Hao Wang, Jie Zhang, Adversarially Robust Deepfake Detection via Adversarial Feature Similarity Learning, arXiv preprint arXiv:2403.08806, March 2024