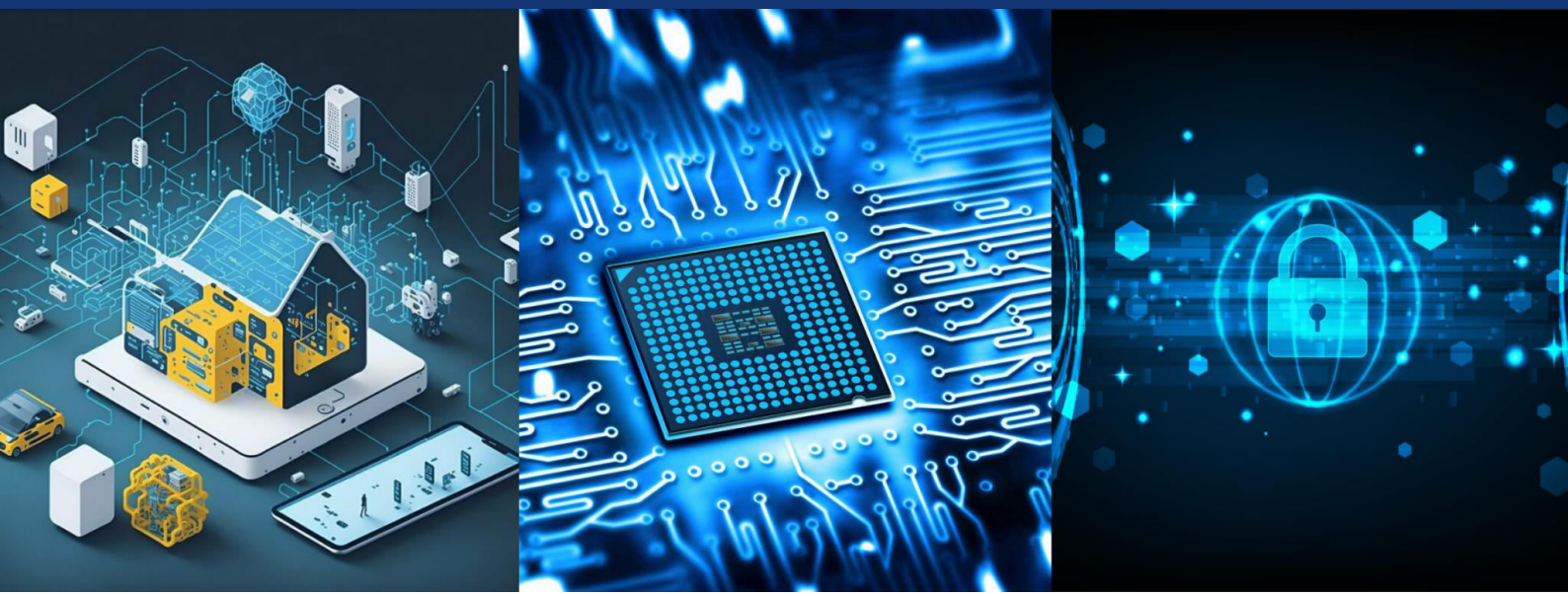
IJIRCCE

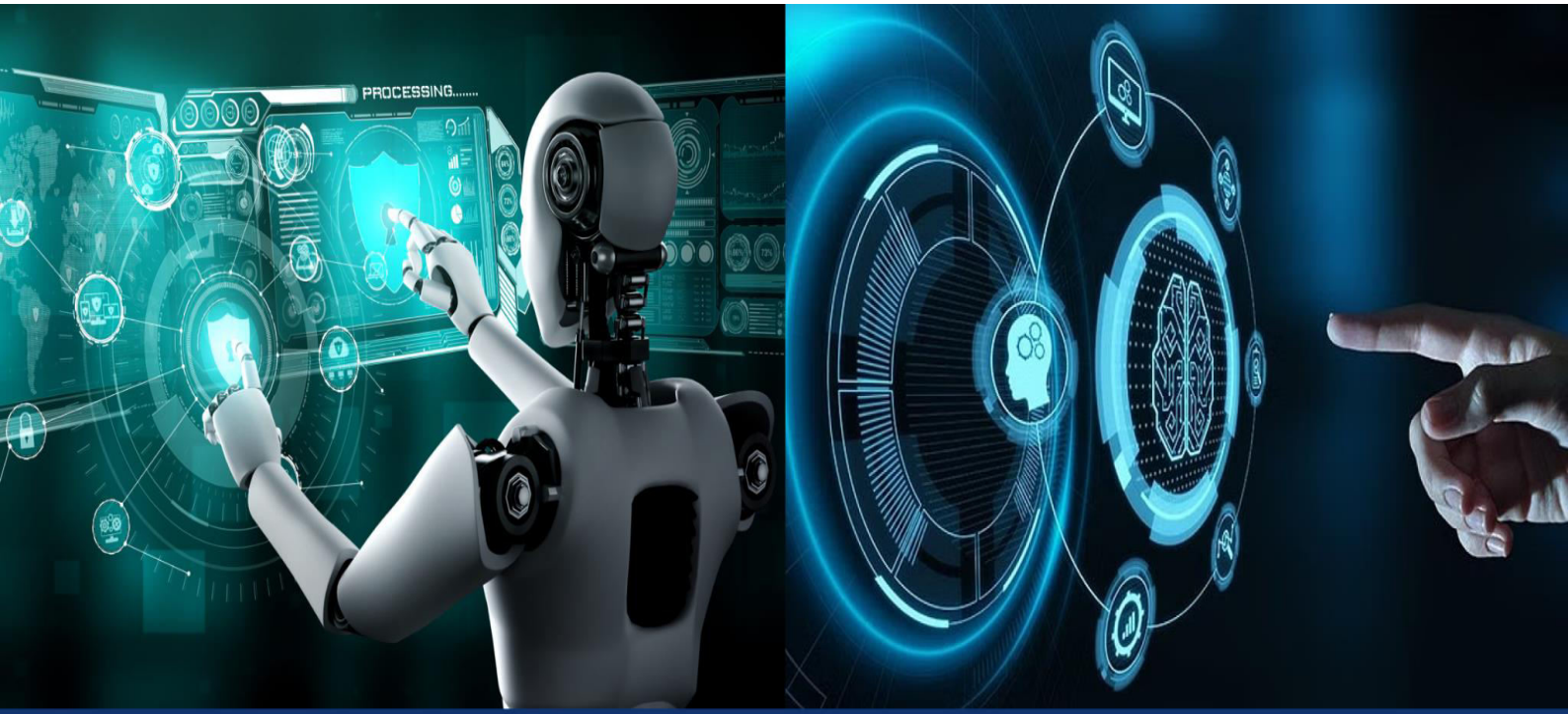
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 10, October 2026

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410029

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Deepfake Video Detection System using Spatio-Temporal Features

Chaitra A Jogihalli[1], Jyothi T H[1], Kavana N[1], Nanda J Kademani[1], Er. Sanjeev Jigali[2], Er. Gagandeep B M[2]

B.E. Students (7th Semester), Department of Computer Science and Engineering, STJ Institute of Technology, Ranebennur, Karnataka, India[1]

Assistant Professors, Department of Computer Science and Engineering, STJ Institute of Technology, Ranebennur, Karnataka, India[2]

ABSTRACT: Deepfake technology uses deep learning to replace or alter facial identities in images and videos, creating realistic content that can be misused for misinformation, impersonation, fraud and privacy violations. This paper presents a deepfake video detection system that combines spatial and temporal information. Five uniformly spaced frames are sampled from each video, the largest face is detected, cropped and resized to 160×160 pixels, and MobileNetV3 extracts spatial features, which are cached to reduce repeated computation. A Gated Recurrent Unit (GRU) network learns temporal relationships across the sampled frames, and a gated fusion mechanism combines both feature types for real/fake classification. The system is built on the FaceForensics++ dataset and compares spatial-only, temporal GRU, feature-concatenation and gated-fusion models using accuracy and precision, with leave-one-manipulation-out testing for generalization and Grad-CAM for visual interpretability. A Streamlit interface lets users upload a video, view the prediction with a confidence score, and download a JSON analysis report.

KEYWORDS: Deepfake detection, MobileNetV3, Gated Recurrent Unit, spatio-temporal features, gated fusion, Grad-CAM, FaceForensics++.

I. INTRODUCTION

Advances in Artificial Intelligence, Machine Learning and Deep Learning have enabled highly realistic synthetic multimedia. Deepfakes, which replace or alter faces in images and videos, have useful applications in entertainment, filmmaking, education and digital media, but they can also be misused for spreading misinformation, identity impersonation, fraud and privacy violations. These risks make automated systems that distinguish authentic media from manipulated content increasingly important.

Detection is difficult because modern manipulations look visually convincing. However, manipulated content may still contain unnatural facial textures, inconsistencies around facial boundaries, or irregularities across consecutive frames. Spatial analysis examines visual information within individual frames, whereas temporal analysis examines how features change across a sequence of frames. Methods relying only on single frames may miss inconsistencies that appear over time, while temporal analysis alone may not capture fine-grained artifacts within facial regions. The problem addressed here is how to develop and evaluate a detection approach that combines spatial facial features with temporal information from multiple video frames.

II. RELATED WORK

Lu et al. [1] proposed a long-distance attention approach in which frames are face-cropped (Dlib or MTCNN), normalized and passed to an Xception-based extractor; the features are split into spatial and temporal attention modules, fused, and classified with a fully connected layer trained using binary cross-entropy on FaceForensics++ and Celeb-DF.

Kumar et al. [2] presented a hybrid system combining EfficientNet and ResNet50, with Haar Cascade face detection, frame-level predictions aggregated into a video decision, and a Flask-based web application for uploading images and

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410029

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

videos and viewing results. Wodajo and Atnafu [3] introduced a Convolutional Vision Transformer in which a CNN extracts local features and a Vision Transformer learns global relationships using attention; trained on the DFDC dataset with binary cross-entropy, it reached about 91.5% accuracy.

III. LIMITATIONS OF EXISTING SYSTEMS

Existing deepfake detectors based on traditional machine learning and deep learning analyse facial textures, pixel-level inconsistencies and unnatural boundaries. Their main limitations are: (1) many analyse frames independently, giving limited temporal analysis; (2) subtle facial artifacts such as blending errors, inconsistent lighting and minor distortions are hard to detect; (3) spatial and temporal information are often not effectively integrated; (4) complex architectures demand high computational resources; (5) generalization to unseen manipulation techniques is poor, and compression or recording conditions can degrade performance; (6) predictions are difficult to interpret because many models output only a label and confidence score; and (7) processing every frame of a video is costly in time and memory. These limitations motivate an efficient system that fuses spatial and temporal information and supports visual interpretation.

IV. OBJECTIVES AND ADVANTAGES OF THE PROPOSED SYSTEM

The objectives of this work are to: (a) develop a deep learning model that classifies videos as real or fake using FaceForensics++; (b) extract spatial features from facial images using MobileNetV3; (c) analyse temporal relationships between frames using a GRU; (d) combine spatial and temporal features through feature concatenation and gated fusion and compare their performance; (e) evaluate the model using accuracy and precision, together with leave-one-manipulation-out testing; and (f) improve interpretability using Grad-CAM.

The proposed approach offers several advantages: temporal feature learning that can expose inconsistencies across frames; spatial feature extraction with a lightweight network; spatio-temporal fusion that considers both kinds of information during classification; potential computational efficiency through uniform frame sampling and feature caching, which makes development practical on computers with limited hardware; an explicit evaluation of generalization to unseen manipulations; and visual interpretability through Grad-CAM heatmaps.

V. PROPOSED METHODOLOGY

The proposed system (Fig. 1) processes an uploaded video through the following stages, following a waterfall development model (requirement analysis, design, implementation, testing, deployment and maintenance).

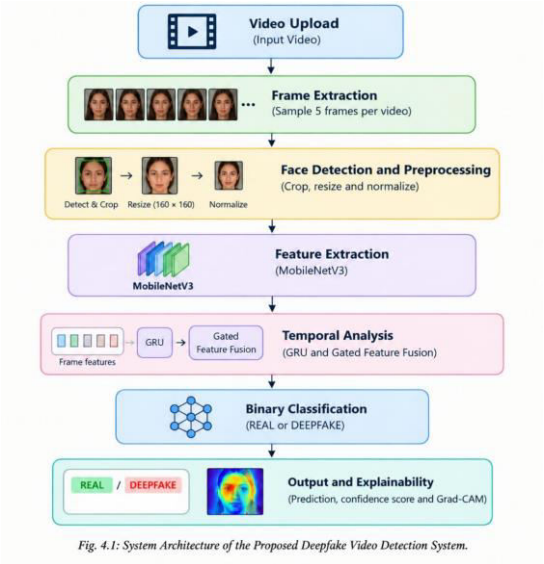


Fig. 1: System architecture of the proposed deepfake video detection system

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410029

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Dataset and Frame Sampling

The FaceForensics++ dataset, with separate training, validation and testing sets, is used. Five uniformly spaced frames are extracted from each video so that different portions of the video are represented while the number of processed frames stays small.

B. Face Detection and Preprocessing

Faces are detected in each frame using OpenCV, the largest face is cropped, resized to 160×160 pixels and normalized. Frames with no detectable face are reported to the user rather than processed.

C. Spatial Feature Extraction

MobileNetV3-Large extracts spatial features such as facial structure, skin texture, boundaries and visual patterns that may indicate manipulation. The resulting feature vectors are cached and reused, avoiding repeated feature extraction during experimentation.

D. Temporal Analysis and Gated Fusion

The feature vectors of the five frames are arranged in temporal order and given to a GRU network, which learns changes in facial appearance and expression across frames. A gated fusion mechanism then combines the spatial and temporal representations. Four models are compared: spatial-only, temporal GRU, feature-concatenation, and gated fusion.

E. Classification, Evaluation and Interpretability

A classifier labels the video as REAL or DEEPFAKE with a confidence score. Performance is evaluated using accuracy and precision (recall, F1-score and ROC-AUC can be computed from test predictions), and leave-one-manipulation-out testing, where one manipulation type is withheld from training, assesses generalization to unseen deepfakes. Grad-CAM heatmaps highlight the facial regions that influenced each prediction.

VI. SYSTEM DESIGN

The system design is documented using data flow, class, sequence, use case and activity diagrams. The Level 0 data flow diagram treats the system as a single process: the user uploads a video and receives a prediction, a confidence score and a JSON analysis report. The Level 1 diagram (Fig. 2) details six processes: upload and validate the file, preprocess the media, extract features, classify the media, display the prediction, and generate the JSON report.

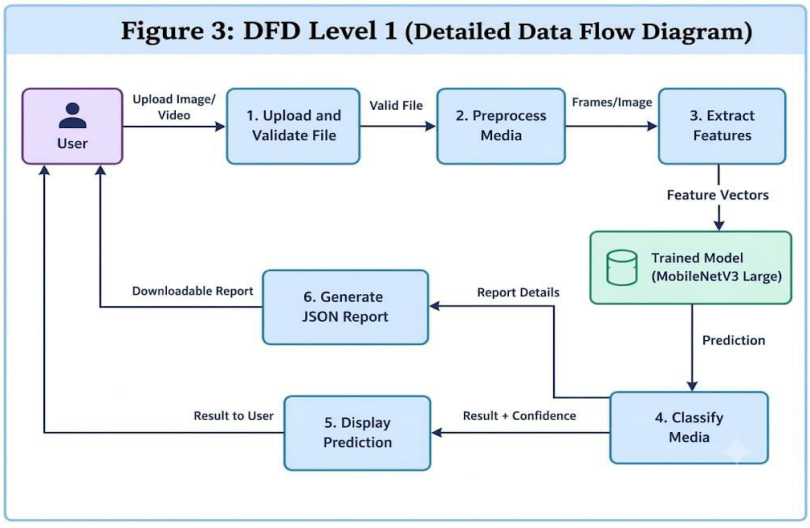


Fig. 2: Level 1 data flow diagram

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410029

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The class design consists of User, StreamlitApp, FileHandler, Preprocessing, FeatureExtractor, MobileNetV3Large, DeepfakeClassifier and ReportGenerator classes. As the sequence diagram (Fig. 3) shows, the web application passes the uploaded video to the preprocessing module; the processed frames go to feature extraction, then to the temporal analysis and fusion module, then to the classifier; Grad-CAM generates the visual explanation, and the output module returns the prediction, confidence score and explanation to the user. If no face is detected, the activity workflow ends with a "No Face Detected" message.

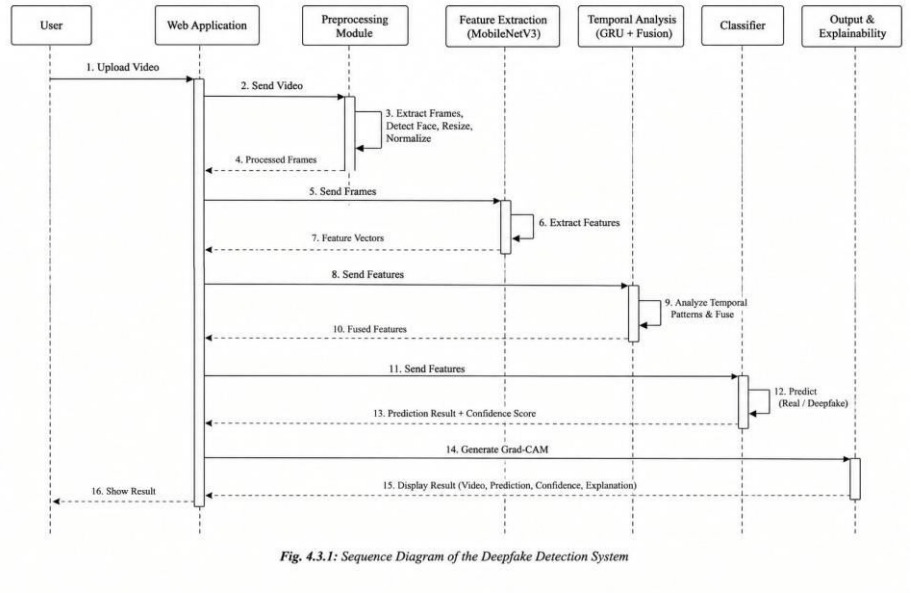


Fig. 3: Sequence diagram of the system

VII. SYSTEM REQUIREMENTS AND IMPLEMENTATION

The system is implemented in Python using PyTorch for the MobileNetV3 and GRU models, OpenCV for video reading, frame extraction, face detection and resizing, and Pillow for image loading. Grad-CAM provides visual explanations, development uses Visual Studio Code, and a Streamlit web interface allows video upload, result display and JSON report download. The recommended development system is a Windows laptop with an Intel Core i5 (multi-core) processor, at least 8 GB RAM and sufficient storage for code, libraries, extracted frames and model files.

Functionally, the system validates the uploaded video format, extracts five frames, detects and crops faces, preprocesses them, extracts spatial features, performs temporal analysis and fusion, classifies the video, and produces Grad-CAM visualizations. Non-functional goals are reasonable processing time, consistent predictions, computational efficiency, usability, compatibility with the Python environment, and modular code organized into video processing, face detection, preprocessing, feature extraction, temporal analysis and classification components.

VIII. TESTING

Testing verified video input handling, frame extraction, face detection and cropping, preprocessing, spatial and temporal feature extraction, classification, performance evaluation and error handling. Table I summarizes the principal test cases; all behaved as expected.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410029

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE I: SUMMARY OF TEST CASES

S.No	Test scenario	Expected result	Status
1	Upload a valid video file	Video accepted for processing	Pass
2	Upload unsupported/corrupted video	Appropriate error message displayed	Pass
3	Extract frames from a valid video	Five uniformly spaced frames extracted	Pass
4	Detect and crop a face	Largest detectable face cropped correctly	Pass
5	Preprocess face images	Resized to 160×160 pixels and normalized	Pass
6	Extract spatial features	MobileNetV3 produces features for frames	Pass
7	Analyse temporal features	GRU processes the frame-feature sequence	Pass
8	Classify a video	Real or fake prediction returned	Pass
9	Video with no detectable face	Handled without crashing; message shown	Pass

IX. RESULTS AND INTERFACE

The application provides a landing page, a video upload interface and an input preview. After confirmation, five frames are extracted, faces are detected and cropped, and the system displays a REAL or DEEPFAKE result with confidence information, together with a downloadable JSON analysis report. Fig. 4 to Fig. 7 show the output screens for a video classified as real and a video classified as a deepfake.

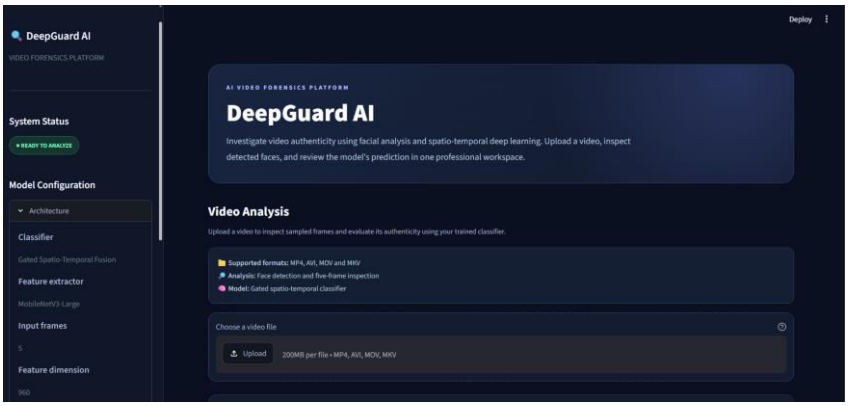


Fig. 4: Landing page of the application

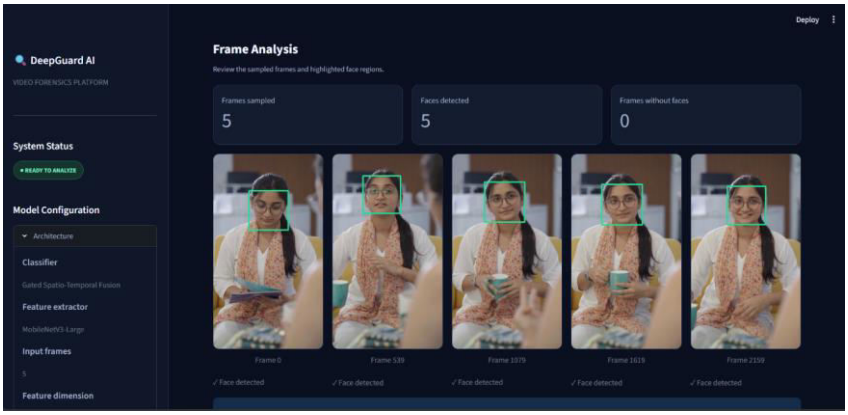


Fig. 5: Face detection in five extracted video frames

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410029

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

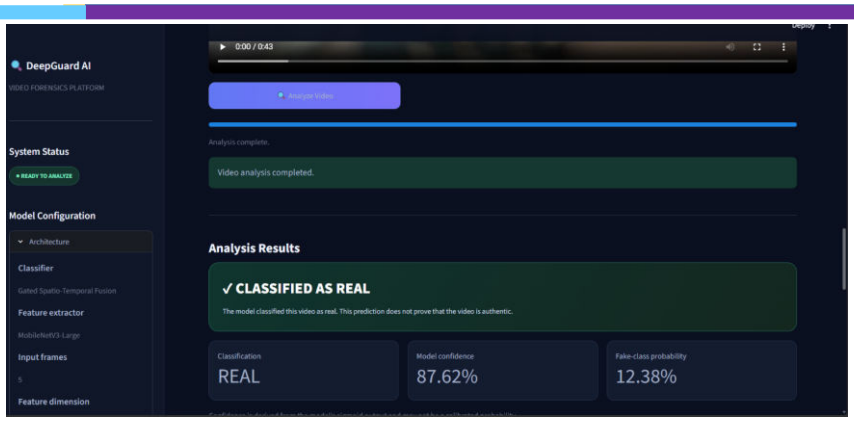


Fig. 6: Detection result for a real video

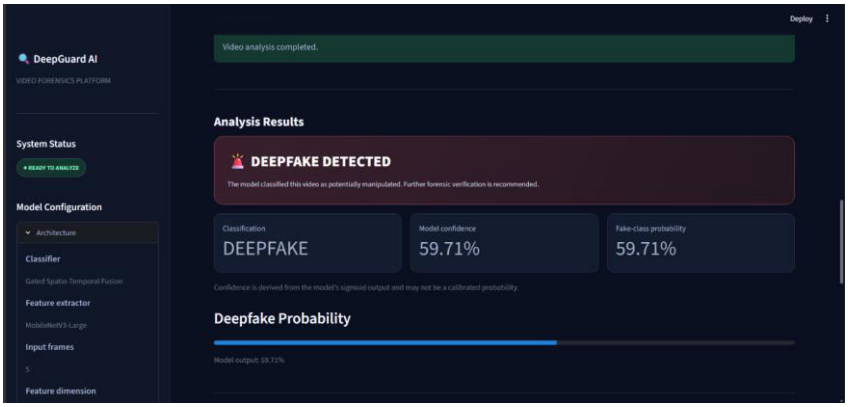


Fig. 7: Detection result for a deepfake video

X. CONCLUSION AND FUTURE SCOPE

This paper presented a deepfake video detection system that uses MobileNetV3 for spatial feature extraction, a GRU for temporal analysis and gated fusion to combine them, with uniform frame sampling and feature caching to keep computation modest and Grad-CAM to make predictions interpretable. Factors such as compression, lighting and unseen manipulation techniques may affect performance. Future work includes testing additional datasets, improving generalization and processing speed, and refining the user interface.

REFERENCES

[1] W. Lu, L. Liu, J. Luo, X. Zhao, Y. Zhou and J. Huang, "Detection of Deepfake Videos Using Long-Distance Attention," IEEE, June 2021.

[2] S. K. K. N., G. V. J. P. Reddy, G. P. B. and K. Naveen, "Deepfake Detection System Using Hybrid Deep Learning for Images and Videos," in Proc. Sixth Int. Conf. on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT), IEEE, 2026.

[3] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," Jimma University and Addis Ababa University, 2021.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details