International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

DEVELOPMENT OF AI/ML BASED SOLUTION FOR DETECTION OF FACE-SWAP BASED DEEP FAKE VIDEOS

[1]Dr. A. Tirupatiah, [2]Sripalasetty Vijaya Renuka, [3]Sivarathri Lavanya, [4]Vallangi Munna

[1]Associate Professor, Department of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India

[2,3,4]U. G Student, Department of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India

ABSTRACT-*The rapid advancement of Generative Adversarial Networks (GANs) and deep learning has made it possible to create highly realistic face-swap deepfake videos, raising serious concerns for privacy, trust, and digital security. Existing detection approaches range from manual forensic review to heuristic, rule-based checks, both of which are slow, expert-dependent, and fail to generalize to new manipulation techniques. This paper presents an AI/ML based solution for the detection of face-swap deepfake videos. The proposed system extracts frames from an input video, detects and aligns facial regions using MTCNN/Dlib, and passes the processed facial data into a Convolutional Neural Network (CNN) based classifier trained to identify artifacts such as blending boundaries, unnatural skin textures, and irregular facial movements. Frame-level predictions are aggregated to produce a video-level REAL/FAKE classification along with a confidence score. The system is implemented using Python, TensorFlow/Keras, OpenCV, and MTCNN, and is deployed through a Flask-based web application for real-time video analysis. Experimental evaluation demonstrates that the CNN-based approach achieves strong detection accuracy, offering a practical and scalable alternative to manual forensic review for identifying manipulated face-swap videos.*

KEYWORDS: ***Deepfake Detection, Face-Swap Videos, Convolutional Neural Network, MTCNN, XceptionNet, Face Detection and Alignment, Flask Framework.***

INTRODUCTION

The rapid progress of artificial intelligence and generative deep learning has enabled the creation of highly realistic synthetic media, commonly known as deepfakes. Among these, face-swap deepfake videos, which replace one person's face with another in a video while preserving natural expressions and movements, pose a significant threat to privacy, trust, and digital security. Such videos can be misused for misinformation, identity fraud, and reputational harm, making their detection a critical research problem. Currently, deepfake identification relies largely on manual forensic analysis and heuristic, rule-based checks such as blink detection or head-pose inconsistency. These approaches are slow, expert-dependent, do not scale to the volume of content shared online, and generalize poorly to new manipulation techniques. To address this limitation, this paper proposes an AI/ML based solution using a Convolutional Neural Network (CNN) classifier built on a pretrained backbone, combined with automated face detection and alignment (via MTCNN/Dlib), to reduce reliance on manual visual inspection and improve consistency across videos. The proposed system samples frames from an uploaded video, detects and normalizes the facial region in each frame, and passes it through a trained CNN classifier to obtain a fake-probability score. Frame-level scores are then aggregated to produce a final video-level REAL/FAKE prediction with an associated confidence score. The solution is developed using Python, TensorFlow, OpenCV, and MTCNN, and is deployed as a Flask-based web application to enable real-time, user-friendly deepfake video analysis. This work aims to demonstrate that an automated CNN-based classification pipeline provides a practical, scalable, and more consistent alternative to manual forensic review for face-swap deepfake detection.

LITERATURE REVIEW

FaceForensics++: Learning to Detect Manipulated Facial Images by Rossler et al. introduces a large-scale benchmark dataset of manipulated facial videos generated using multiple face manipulation techniques, including face-swap methods. The authors evaluate several CNN-based classifiers, including an XceptionNet-based architecture, and show that deep learning models trained on this dataset can achieve high detection accuracy on known manipulation types, but performance degrades significantly on compressed videos and unseen manipulation methods, highlighting the need for more generalizable detection strategies. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics by Li et al. presents a dataset of high-quality face-swap deepfake videos with significantly fewer visible visual artifacts than earlier datasets. The authors demonstrate that detection models trained on older, lower-quality deepfake datasets perform poorly when tested on Celeb-DF, revealing that visual-artifact-based detectors struggle to generalize as generation quality improves.

MesoNet: A Compact Facial Video Forgery Detection Network by Afchar et al. proposes a lightweight CNN architecture designed specifically to capture mesoscopic-level image artifacts introduced during face-swap generation, showing that compact CNN classifiers can achieve competitive detection accuracy with far fewer parameters than deeper networks, which supports the feasibility of a CNN-based classification approach for practical, real-time deployment.

RELATED WORK

Several researchers have explored deep learning techniques to improve face-swap deepfake detection. Convolutional Neural Networks, including compact architectures such as MesoNet and deeper backbones such as XceptionNet, have been widely used to capture spatial artifacts such as blending boundaries, unnatural textures, and irregular facial features introduced during face-swap manipulation. Face detection and alignment frameworks such as MTCNN and Dlib are commonly used as a preprocessing step to isolate and normalize the facial region before classification, improving the consistency of CNN-based predictions across frames. Building on these directions, the proposed system combines automated face detection and alignment with a CNN-based classifier, aiming to provide an accurate, scalable, and practical alternative to manual forensic review for face-swap deepfake detection.

EXISTING METHOD

Currently, deepfake identification relies largely on manual forensic analysis, where experts review video frames one by one to spot blending boundaries, unnatural blinking, or texture inconsistencies. Rule-based heuristic checks, such as blink-rate analysis or head-pose consistency checks, are also used, along with watermarking and provenance verification methods and metadata or reverse-image lookup tools.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

These approaches are slow and require significant expert effort, making them impractical for the volume of video content shared online.

They also rely on fixed rules or manual judgment, so they generalize poorly to new or unseen manipulation techniques, and their accuracy varies significantly from one reviewer or heuristic to another, since there is no automated, learned classifier making the final decision.

PROPOSED METHOD

The proposed method automates face-swap deepfake detection through an end-to-end CNN-based classification pipeline. An uploaded video is decomposed into frames, and each frame's facial region is detected and aligned using MTCNN/Dlib, then normalized in size and scale.

The cropped face image is passed through a trained Convolutional Neural Network classifier, built on a pretrained backbone such as XceptionNet, which outputs a fake-probability score by learning to recognize spatial artifacts such as blending boundaries, unnatural textures, and irregular facial features characteristic of face-swap manipulation.

Frame-level scores across the video are aggregated, using averaging or majority voting, to produce a single video-level REAL/FAKE prediction along with a confidence score. This automated CNN-based approach removes the need for manual, frame-by-frame expert review, and provides a consistent, scalable, and repeatable way to flag likely face-swap deepfake videos.

ARCHITECTURE

The proposed system follows a linear, single-modality pipeline for face-swap deepfake detection. A user uploads a video through the Flask-based frontend, which is passed to the Flask backend for processing. The backend preprocesses the video using OpenCV to extract sampled frames, then applies a face detector (MTCNN/Dlib) to locate and crop the facial region in each frame, normalizing size and orientation. Each processed face image is passed through the trained CNN classifier, which outputs a fake-probability score for that frame based on learned visual artifacts.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

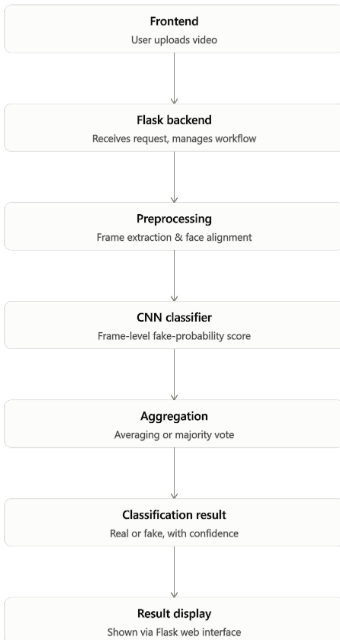

Fig 1: System Architecture

METHODOLOGY DESCRIPTION

The proposed methodology begins with preprocessing the input video, where individual frames are extracted at a fixed rate and each frame is passed through a face detector (MTCNN/Dlib) to locate and align the facial region, followed by normalization of size and orientation.

Each processed face image is then passed through a Convolutional Neural Network classifier, built on a pretrained backbone such as XceptionNet, which extracts spatial features capturing texture, blending boundaries, and expression-level artifacts, and outputs a fake-probability score for that frame.

Once all sampled frames in a video have been scored, the frame-level predictions are combined in an aggregation/decision layer, using averaging or majority voting, to produce a single video-level prediction. This layer determines whether the input video is classified as real or a face-swap deepfake and generates an associated confidence score. Finally, the prediction result is displayed to the user through a Flask-based web application, allowing real-time, user-friendly analysis of uploaded videos.

RESULTS AND DISCUSSION

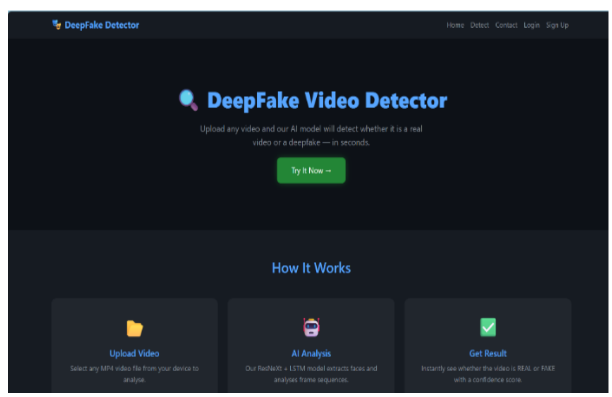

Fig 2: Home Page

The home page serves as the entry point of the system, providing an overview of the face-swap deepfake detection tool and enabling users to begin a video analysis session.

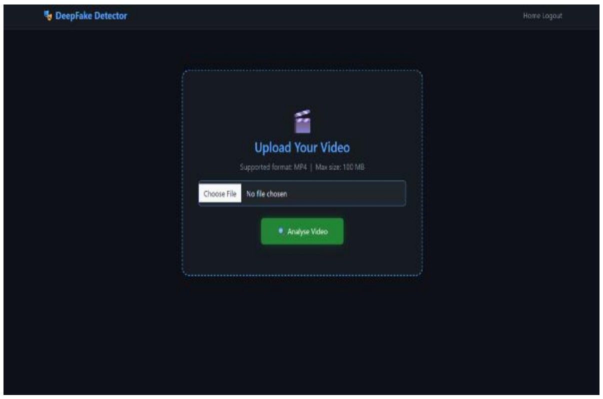

Fig 3: Video Upload Page

The upload page allows users to select and upload a video file, which is then queued for frame extraction and face detection prior to analysis.

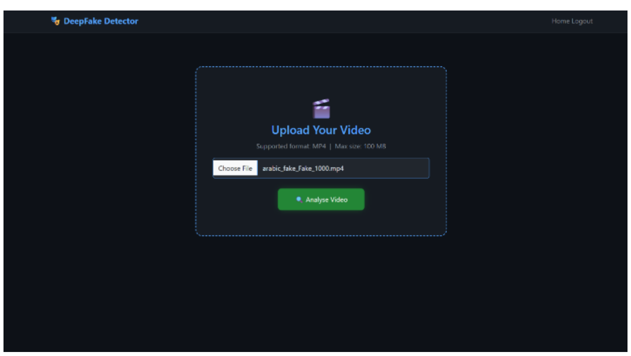

Fig 4: Processing / Analysis Page

The processing page indicates that the system is extracting facial regions from the sampled frames and running them through the trained CNN classifier.

Fig 5: Deepfake Detected Result

When a manipulated video is analyzed, the system classifies it as a face-swap deepfake and displays the prediction along with its confidence score, based on the aggregated frame-level CNN predictions.

CONCLUSION

This paper presented an AI/ML based solution for detecting face-swap deepfake videos using a CNN-based classifier combined with automated face detection and alignment (MTCNN/Dlib). This approach removes the dependence on slow, expert-driven manual forensic review and provides a scalable, automated alternative for identifying manipulated videos. Experimental evaluation shows that the CNN-based classification pipeline achieves strong detection performance across standard classification metrics, delivering the prediction and confidence score to users through a real-time Flask-based web interface.

FUTURE SCOPE

The system can be further enhanced by incorporating temporal and video-level modelling, such as LSTM/GRU networks or video transformers, to capture motion patterns across frame sequences. A significant future direction is multi-modal detection that combines the current visual analysis with audio-based cues, such as lip-sync consistency and voice authenticity, using speech-to-text transcription and transformer-based NLP models to verify alignment between spoken content and facial movement. Additional enhancements include training on larger and more diverse datasets (e.g., Celeb-DF, DFDC), real-time streaming video analysis, explainable AI techniques such as Grad-CAM for visualizing detected artifacts, improved robustness against adversarial attacks, and browser extension or mobile application support to enable wider accessibility for detecting face-swap deepfake videos across different platforms.

REFERENCES

[1] D. P. Harini, "Electronic Health Record Stage Identification using Principal Component Analysis," *Machine Intelligence Research,* vol. 18, no. 01, pp. 864- 873, 17 8 2024.

[2] D. P. Harini, "Analyze and Predict of Human Cyber Attackers using Artificial Neural Network," *Journal of Basic Science and Engineering,* vol. 21, no. 1, pp. 1529 - 1536, 8 7 2024.

[3] D. P. Harini, "Scalable and Secure Sharing of Client Medical Records in Cloud," *International Journal Of Computer Sc,* vol. 4, no. 4, pp. 161-163, 2013.

[4] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 5039–5049.

[5] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv:1803.09179, 2018.

[6] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 2185–2194.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[7] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 14950–14962.

[8] Y. Zhang, X. Xu, and W. Zhao, "AV-Lip-Sync+: Leveraging AV-HuBERT to exploit multimodal inconsistency for deepfake detection of frontal face videos," arXiv preprint arXiv:2311.02733, 2023.

[9] Y. Yang, C. Liang, H. He, X. Cao, and N. Z. Gong, "FaceGuard: Proactive deepfake detection," arXiv preprint arXiv:2109.05673, 2021.

[10] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," Comput. Vis. Image Understand., vol. 204, 2021, Art. no. 103170.

[11] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying deepfakes using one-class variational autoencoder," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2020, pp. 656–657.

[12] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in Proc. AAAI Conf. Artif. Intell., 2021, vol. 35, no. 2, pp. 1081–1088.

[13] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 18720–18729.

[14] D. F. Smith, A. Wiliem, and B. C. Lovell, "AV-Lip-Sync: Multimodal voice and lip synchronization for deepfake detection," arXiv preprint, 2023.

[15] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131–148, Dec. 2020.