JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An AI-Powered Approach for Detecting and Preventing Facial Swap Manipulations in Videos

Jiya Paliwal, Kanika Garg Muskan Joshi, Dharmveer Singh Jadeja
UG Scholar, Associate Professor, UG Scholar, UG Scholar
Computer Science and Engineering(Artificial Intelligence),
Geetanjali Institute of Technical Studies, Dabok, Udaipur, Rajasthan

Abstract : The increasing advancement of generation of deepfake techniques - especially manipulations involving face-swapping has brought up major concerns related to integrity of online media, data privacy and societal trust. The computer generated videos, created using advanced models can easily replace an individual face with another often fool regular detection tools because changes in lighting, skin tone, facial expressions are so small and hard to notice. Although many AI-based methods have been developed to spot deep fake, most current models still struggle because they only look at single images , don't consider changes over time or require too much computing power. This research proposes a hybrid deepfake detection framework that leverages the strengths of Convolutional Neural Networks (CNNs) for robust spatial feature extraction and Vision Transformers (ViTs) for capturing temporal and contextual relationships across video frames. The CNN part looks for small changes and edits in the face, while the Vision Transformer looks at a series of frames to catch unusual expressions , movements and facial tone. Together, this combination aims to overcome the challenges posed by diverse and highly realistic face-swap techniques. The system is trained and tested on known datasets like FaceForensics++ and DFDC-Preview, providing a complete way to detect face-swap deep fake. By improving on current methods and looking at both the details in each frame and changes over time, this study helps create a stronger and more flexible deepfake detection system that can handle new and growing threats in visual content.

I.INTRODUCTION

Proposed Solution is a state-of-the-art AI-driven system for addressing the rising threats of sophisticated face-swapping deep fakes. As most detection methods fail due to their reliance on single-frame-based image analysis, Proposed Solution addresses small lighting, skin tone, and facial expression variations with the power of CNNs and Vision Transformers to provide more accurate and reliable deep fake detection. The framework utilizes Convolutional Neural Networks (CNNs) to extract fine-grained spatial features from face regions, and Vision Transformers (ViTs) to analyse time-domain patterns and context information throughout video sequences—accurately identifying anomalous facial movements and abnormalities.

Proposed Solution is trained and evaluated based on highly reputable benchmark datasets such as FaceForensics++ and DFDC-Preview, providing robust performance, flexibility, and high accuracy in the detection of diverse varieties of realistic deepfake material. By examining each frame in detail and observing patterns within the video, Proposed Solution provides a comprehensive and scalable solution to modern deep fake detection. This research presents the system design, implementation, and evaluation—augmenting endeavours toward the protection of credible, reliable, and authentic digital content.

II. LITERATURE REVIEW

The rising sophistication of deepfake technology—especially face-swapping—has generated pressing questions regarding digital media authenticity, individual privacy, and public trust. To counter these, researchers and developers have been investigating diverse techniques to identify tampered content. Nevertheless, numerous systems lack adequacy in responding to real-world intricacies, including minute facial modifications, temporal discrepancies, and effective real-time processing. For a deeper appreciation of the situation and determination of research gaps, some existing models, methods, and academic literature were examined. Most current detection tools are based on static image evaluation, which restrains them to detect temporal anomalies and motion-inconsistencies in deep fake videos. Most of these approaches tend not to account for transitions between frames, long-term face behaviour, or minor manipulations in lighting or expression—characteristics essential in the detection of high-quality face-swaps.

Conventional CNN-based models, like FaceForensics++ and DeepFaceLab, are mostly concerned with spatial aspects of a single frame. While they provide reasonable performance for visual artifact detection, they do not have temporal modelling capabilities and tend to be vulnerable to noise, compression, or resolution variations in videos. More recent methods have incorporated Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models to handle variations based on time. These

structures, however, have difficulty with scalability and are computationally expensive. Additionally, they generally do not generalize as well across different types of datasets and manipulations.

Transformers in computer vision, like ViT (Vision Transformer), have demonstrated encouraging performances in learning long-range dependencies on image sequences. Although research involving the combination of ViT with CNNs is on the rise, few centre on deep fake detection, and even fewer report their performance on real-world face-swap operations. Face X-ray and Deep Fake Detection Challenge (DFDC) submissions have advanced research by providing benchmark datasets and motivating well-designed models. Despite this, several participating models remain black-box in character, non-interpretable, or are not designed for real-time detection and deployment. Some of these academic analyses delve into the social and ethical implications of deep fakes, especially in terms of policy and regulation rather than technical countermeasures. Such analyses are crucial but don't directly assist in enhancing detection accuracy or model robustness.

In conclusion, this Approach overcomes the shortcomings of current deepfake detection techniques with the combination of spatial and temporal feature extraction based on a hybrid architecture—incorporating CNNs for face analysis at detailed levels and Vision Transformers for discovering sequence-based patterns—providing an accurate, efficient, and scalable tool for identifying low-level face-swapping attacks from various datasets such as FaceForensics++ and DFDC-Preview.

III. SYSTEM MODEL

Proposed Solution is an integrated deepfake detection framework designed to identify altered facial videos—especially face-swaps—by leveraging the combined capabilities of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). This model aims to overcome the drawbacks of traditional detection mechanisms by examining spatial features and temporal inconsistency among video frames. Architecture and development are described under the subsequent submodules: Data Exploration, Algorithms & Techniques, and Model Evaluation & Validation.

A. DATA EXPLORATION

It's important to know the type of deep fake videos, how they can be manipulated, and how varied a dataset might be before planning the detection pipeline.

1. Data Sources: We train and evaluate our model on standard benchmark datasets like FaceForensics++, DFDC Preview, and Celeb-DF. These datasets typically include both real and deepfake videos, usually in common formats like.mp4 or.avi. Frame-level annotations are sometimes provided, giving rich labels that indicate what portions of the video are edited and what are original.

2. Data Preprocessing: These annotations can be particularly helpful for fine-tuning detection accuracy. To pre-process the Video data for analysis, each video is first decomposed into separate frames using OpenCV. Face detection and cropping are performed using MTCNN to separate meaningful facial areas. The resulting frames are resized, normalized, and segmented into sequences in order to account for temporal dynamics. In order to increase the strength of the model, various methods of augmentation are used, including horizontal flipping, noise addition, and brightness adjustment—mimicking real-world variation and making the system more capable of detecting deepfakes across a variety of situations

3. Exploratory Data Analysis (EDA): In exploratory data analysis, distribution between real and fake videos are analysed carefully to understand how balance the database is. Other assessments are targeted towards frame coherence, facial alignment consistency, and the existence of deepfake artifacts that could be indicative of manipulation. To facilitate these observations, visual dashboards are generated through software such as Matplotlib and Seaborn to facilitate the presentation of frame quality patterns, video length patterns, and general organization of the dataset in an understandable format.

B. ALGORITHMS & TECHNIQUES

The power of the system comes from its two-line design that coordinates spatial and temporal analysis to achieve strong in-Depth detections. In spatial analysis, the Neural Network (CNN) is used as a rescaling or image to extract high loyal facial Features from a single frame of a video. These include face edges, compositional anomalies and facial feature inconsistencies at pixel level, which are signs of tampering.

Temporal pattern recognition is label with vision transformer (VIT), such as transformers, which examines the temporal patterns across video frames. It identifies some subtle deviations as unnatural eyelids that flash or are high mouthpieces that do not take into account the static frame analysis.

The hybrid model pipeline merges both CNN and ViT output through a merger layer that blends spatial and temporary functions. The resulting joint representation is sent to a final classification team that predicts a real or false label, with a confidence point.

On the back, the models are distributed on cloud infrastructure such as AWS EC2 and S3 to ensure scalability and availability. API is an interface for simple video uploads and real-time Deepfake score.

C. MODEL EVALUATION & VALIDATION

Robustness, accuracy, and adaptability are important in deepfake detection. It is tested using several standard metrics to measure its performance and reliability.

1. Accuracy Metrics: The system evaluates performance using multiple classification measurements including accuracy, precision, recall and F1score, which helps to assess how well the model distinguishes between real and fake videos. In addition, ROC (receiver operating characteristics - area under the basket) is used to measure class insulation, which provides insight into the model's ability to distinguish between classes in different thresholds. An confusion matrix is also used to analyse false positivity and the number of false negative, which gives a more detailed approach to predicated errors.

2. Performance Testing: The system undergoes a complete performance test to ensure efficiency and reliability. This involves measuring the speed of the estimate for both the frame and per video to evaluate how quickly the model process processes. The GPU and CPU resource profiles are performed to understand the calculation load and adapt resource use. In addition, a strong test is done by analysing the system's ability to maintain accuracy while handling compressed or low quality videos, and ensures frequent performance in different real world scenarios.

3. Cross-Dataset Generalization: To assess the normalization functions of the model, the training is performed on a dataset, such as FaceForensics++, while testing is tested on different datasets such as test celeb-DF and YouTube videos in the real world. This approach helps to evaluate how well the model adapters for different manipulative techniques and data distribution ensure that it effectively performs in a wide range of deepfake style and sources

4. User Feedback: The system includes a response to confirmation verification of the model for journalists or media organizations, so that they can review and confirm the accuracy of Detection Results. In order to increase openness, clarification equipment such as Grad-CAM (Gradient-Weighted class activation mapping) is integrated, enabling views on areas of the video frames that affect model's decision. This helps users understand the reasoning of models of models, promote trust and clarity in the results of the system.

D. SUMMARY: FEASIBILITY & VIABILITY

Starting from the initial idea and prototyping, this Approach demonstrates strong potential for application in the real world:

1. Feasibility Aspect: The system benefits from the reference dataset that is available publicly in public, and provides a strong basis for model development. Pretrained CNN and transformer models are used to reduce calculation costs and improve efficiency. In addition, cloud services and Open Source tools facilitate rapid development, which enables rapid distribution and repetition of the system.

2. Challenges faced: One of the main challenges is necessary to train hybrid models, especially by combining CNN and transformer architecture. In addition, the lack of annotation timely data has difficulty training the system to identify Deepfake over time, while Deepfakes adds the real variation, such as compression, video quality and lighting status, and adds complexity to the process of detection.

3. Solution: For their resolution, learning and pretrained models are shared in order to lower training time as well as computation requirements. Variability is provided for a single frame, and artificial data instances are employed for enhancing the capability of the model to cover multiple situations. Apart from it, the system has been architected in the modular part to permit CNN or VIT (Vision Transformer) parts for effortless modification without completely restructuring the whole model, facilitate adaptability as well as scalability.

Step	Module	Details
1.	Video Upload	User uploads a video via UI.
2.	Frame Extraction	Extract video frames at fixed intervals using OpenCV.
3.	Face Detection	Detect and crop faces using MTCNN.
4.	Preprocessing	Resize faces, normalize pixel values, apply augmentations.
5.	CNN Module (Spatial)	A CNN analyses frame-wise textures, edges, lighting.
6.	ViT Module (Temporal)	Vision Transformer captures patterns across sequences of frames.

7.	Feature Fusion	Output embeddings from CNN & ViT merged into a combined vector.
8.	Classification	Fully connected layer with Softmax or Sigmoid gives "Real" or "Fake". score
9.	Post Processing	Show prediction with a confidence level.
10.	Results	Display Visual dashboard with video label and probability.

Figure 1: Model Pipeline

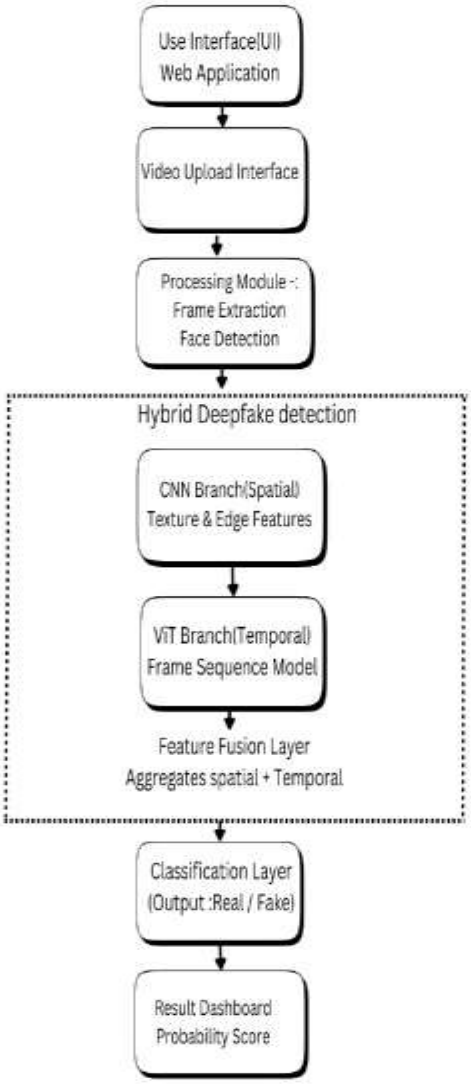


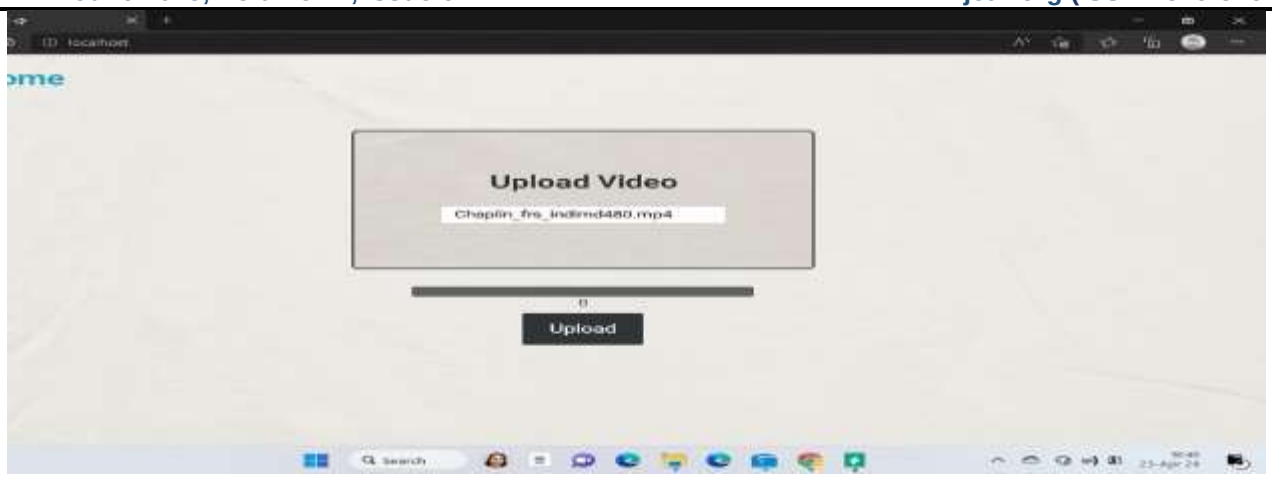
Figure 2: Model Flow

IV. PROPOSED SOLUTION

This project is a deep learning-based web app built to detect deep fake videos, especially face-swap ones. It uses tools like OpenCV, TensorFlow, Flask, and MediaPipe to analyse faces and decide if a video is real or fake. With a simple and easy-to-use interface, the platform helps protect digital content by checking its authenticity and stopping the spread of fake media.

1. User Interface: The user interface for solution system as proposed enables users to upload MP4 videos conveniently using basic web page. Uploading first verifies the system whether the file is in the right format and if it is in the required specifications. Upon a valid file, the system proceeds to scan the video frame. A progress indicator notifies the user of the upload and detection status. Once the analysis is finished, the outcome is shown, which tells whether the video is real or not.

1.1 Backend System: On Backend, the entire process is handled by a bottle server, uploads the file, monitors frame recovery and forward data to the model. MediaPipe and OpenCV are used by the system for facial detection, and it detects every frame for faces. It scans for forgery indicators, including irregular face movement. A trained model produced using TensorFlow/Keras provides the final output, which identify if the video is original or edited.



2. How It Works (Workflow): The workflow begins when the user uploads a video, which is then divided into individual frames of the backend system. Each frame is treated to identify faces, and the system ensures that all faces are analysed properly. Then an intensive teaching model is used to confirm whether the video is edited or manipulated. When the analysis is completed, the result is displayed for the user and it is also stored for future references. If necessary, the user can download the result and log for further use.

3. Advantages & Contribution: This system provides many valuable benefits in different fields. At the social level, it helps to raise awareness of the dangers of Deepfake, prevent the spread of false videos by examining in advance, and encourages people to share material responsibly. When it comes to security and faith, the deepfakes used in fraud can assist law enforcement police or cyber security teams to confirm digital evidence. It also helps build trust in the authenticity of online videos. From an economic point of view, it saves time and money that will otherwise be spent manually on reviewing the video. This brand also helps prevent the reputation and abuse of digital materials and can be used in areas such as HR, legal work or media for material verification.

4. Who Can Use It (Target Users): The system is designed for a wide range of users. News agencies can use it to check the video before sending. Schools and colleges can use it to ensure that students are posted in real. Companies can use it to confirm the authenticity of employee or promotional videos. And finally, everyday users can trust whether they watched a video is real or false.

5. Tools & Technologies Utilized: The system is a combination of strong tools and technologies. Flask is used to manage video uploads and run the server to run the application. OpenCV is used to extract video frames and detect faces, while MediaPipe is used to track facial expression for detailed analysis. The deepfake detection model that has been trained and operates using TensorFlow/Keras. The interface of the website is constructed through HTML, CSS, and JavaScript for a seamless user experience. Both NumPy and Pandas are used for data organization and processing. Matplotlib and Seaborn are optionally used to create charts and visualize outcomes clearly.

6. Special Features: The system provides a number of important features that enhance efficiency. It has automatic frame extraction, which bypasses any kind of manual video preparation. The prediction processed the data, and determine if a video is real or fake. The model targets facial areas only to enhance precision. All uploaded files are handled safely to protect user privacy. Going forward, the system hopes to detect not only fake videos but also fake images, voices, and text. It also wants to use tools like blockchain to help check if content is real. Another goal is to team up with media platforms and groups to verify live content as it happens.

7. Future Plans: In the future, the system aims to go beyond video detection and also identify fake images, voices, and text. It add blockchain tagging to help confirm whether digital content is real. Additionally, the system hopes to work with media platforms and organizations to check and verify live content as it's being shared.

V. CONCLUSION

The increasing popularity of deepfake technologies, particularly those that are face-swap based, has created an urgent need for effective and reliable detection mechanisms. We have developed an AI-based system for detecting face-swap-based deep fakes with the aid of deep learning and computer vision techniques in this project. Our approach merges Convolutional Neural Networks (CNNs) for feature extraction and classification with OpenCV and MediaPipe for face tracking, landmark detection, and frame-by-frame processing. Targeting facial swap manipulations—where a subject's face in a video is replaced by another—our system is able to identify subtle inconsistencies in facial structure, expression alteration, and defects in blending that are generally overlooked by standard detection models.

The solution is in a web-based user interface built using HTML, CSS, and Flask, where one uploads video clips and receives predictions in real-time. The UI-based solution makes it usable and accessible, and thus the tool is not just beneficial to researchers, but also to journalists, digital forensic professionals, and social media moderators. Our work also aimed to explore existing deepfake datasets such as FaceForensics++, DFDC, etc., to show training and testing the model on various datasets with varying manipulation styles and video qualities. Utilization of such datasets allowed generalizable and stable learning, but we encountered severe challenges such as class imbalance, low dataset diversity, and evolving manipulation techniques which are causing ongoing trouble for detection mechanisms.

In addition to measuring the performance and correctness of our model, we identified the following existing shortcomings of deep learning-based detection: vulnerability to adversarial examples, overfitting to manipulative methods, and lower dependability in noisy real-world settings. These were also commonly noted issues, as well as suggested directions for future work, which included the incorporation of temporal knowledge (dynamics of frame sequence), transformer-based models, and real-time optimization.
This project not only provides an applicable tool for detecting deep fakes but also provides new insight into the evolving world of synthetic media manipulation. By narrowing our field of study to facial swap-based deep fakes, we have created a foundation upon which more advanced and scalable detection systems can be constructed.

Ultimately, our aim is to promote trust and integrity in digital content by equipping users with tools capable of discerning authenticity. As deepfake technology continues to advance, so must our defence mechanisms—our project represents one step in that direction, encouraging further exploration and innovation in AI-powered media forensics.

VI. REFERENCES

1. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
2. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*.s
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*
5. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition (ResNet). *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
6. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks (MTCNN). *IEEE Signal Processing Letters*.
7. MediaPipe by Google. (2021). MediaPipe: Cross-platform, customizable ML solutions for live and streaming media. https://mediapipe.dev
8. Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
9. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. *Software available from tensorflow.org*.
10. Chollet, F. (2015). Keras: The Python deep learning library. https://keras.io
11. Flask Documentation. (2024). Flask: Web development, one drop at a time. https://flask.palletsprojects.com
12. Seaborn and Matplotlib Documentation. (2024). *Data visualization libraries in Python*. https://seaborn.pydata.org, https://matplotlib.org
13. DeepFaceLab. (2020). DeepFaceLab: The leading software for creating deep fakes. https://github.com/iperov/DeepFaceLab