International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Deep Fake Video Detection using Convolutional Neural Networks

**Yogvedant Patel[1], Twisha Salde[2], Bhargav Shingala[3], Prof. Anas Dange[4]**

*[1]Yogvedant Patel, Artificial Intelligence & Machine Learning, UCoE [2]Twisha Salde, Artificial Intelligence & Machine Learning, UCoE [3]Bhargav Shingala, Artificial Intelligence & Machine Learning, UCoE [4]Prof. Anas Dange, UCoE*

---

**Abstract -** In recent months, the rise of free, deep learning- based software tools has significantly simplified the creation of highly realistic face-swapped videos, commonly known as "Deep Fakes" (DFs). While video manipulation through visual effects has been practiced for decades, recent advancements in deep learning have drastically enhanced the realism and accessibility of such synthetic content. Creating DFs using AI tools has become relatively straightforward; however, detecting these manipulations remains a major challenge. This is due to the complexity involved in training algorithms to recognize subtle and often imperceptible signs of tampering. In this work, we present a deep learning-based approach that combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to detect Deep Fake videos. Our system employs CNNs to extract spatial features at the frame level, which are then passed to an RNN that captures temporal inconsistencies between frames introduced during Deep Fake generation. We evaluate our method on a large dataset of manipulated videos and demonstrate that our approach achieves competitive results using a relatively simple architecture.

***Key Words***: Deep fake video detection, Convolutional neural networks(CNN), Recurrent neural networks(RNN)

## 1. INTRODUCTION

The proliferation of high-resolution smartphone cameras and the global accessibility of fast internet connections have significantly contributed to the surge in social media usage and digital video sharing. Concurrently, advances in computational power have propelled the capabilities of deep learning to levels that were previously considered unattainable. One of the most notable—and concerning— developments arising from this progress is the advent of "Deep Fakes" (DFs), synthetic media generated using deep generative models such as Generative Adversarial Networks (GANs). These models can produce highly realistic video and audio content that closely mimics real individuals, thereby raising serious concerns about misinformation, privacy, and digital security.

Deep Fake videos have become increasingly prevalent on social media platforms, where they are often used to spread disinformation, manipulate public perception, and cause reputational harm. Given the ease with which such content can be created and disseminated, there is a growing imperative to develop robust and reliable deep fake detection methods. Effective detection strategies are essential not only to preserve the integrity of digital media but also to mitigate the societal and ethical risks posed by the unchecked spread of synthetic content.

A thorough understanding of the underlying mechanisms of Deep Fake generation is crucial for developing effective detection approaches. In typical GAN-based synthesis, a target video and a source image are used to generate a manipulated video in which the source face is mapped onto the target. This process involves splitting the video into frames, synthesizing new face images using autoencoders, and reconstructing the manipulated video. Due to practical constraints such as limited computational resources and processing time, the generated

face images are often of fixed resolution and require affine warping to align with the facial geometry of the target. This transformation introduces resolution inconsistencies and visual artifacts, particularly around the boundaries of the manipulated region.

In this study, we propose a novel deep learning-based approach to detect Deep Fake videos by identifying these artifacts. Our method employs a ResNeXt-based Convolutional Neural Network (CNN) to extract spatial features from individual video frames and a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) to capture temporal inconsistencies across frames. By simulating resolution inconsistencies during training, our approach enhances the model's ability to detect the subtle artifacts introduced by GAN-based synthesis. Experimental results on a large dataset of manipulated videos demonstrate the effectiveness and competitiveness of our method in accurately distinguishing real videos from AI-generated Deep Fakes.

## 2. LITERATURE SURVEY

The rapid growth and misuse of DeepFake (DF) videos pose a significant threat to democratic institutions, legal systems, and public trust. This has led to an urgent demand for effective methods of DeepFake analysis, detection, and mitigation. Several recent works have proposed novel approaches to tackle this challenge, each focusing on specific features or signals indicative of synthetic content.

Li et al. [1] in *Exposing DeepFake Videos by Detecting Face Warping Artifacts* introduced a method that leverages inconsistencies caused by affine transformations in GAN- generated faces. Their approach employs a Convolutional Neural Network (CNN) to compare the warped face regions

with the surrounding context, thereby identifying resolution- based artifacts. This work is grounded in the observation that current DeepFake generation techniques typically produce images at a fixed resolution, requiring spatial transformations that introduce detectable anomalies.

In another study, *Exposing AI-Created Fake Videos by Detecting Eye Blinking*, Li et al. [2] focused on physiological inconsistencies—specifically, the absence of natural eye blinking in synthetic videos. Since blinking is often underrepresented in datasets used to train generative models, its detection becomes a strong indicator of forgery. Although this method performs well on benchmark datasets for eye-blinking detection, it relies solely on this single physiological cue. The limitation of considering only blinking suggests the need for broader feature integration, such as subtle facial details like wrinkles and dental visibility.

Sabir et al. [3] proposed the use of capsule networks

for detecting manipulated images and videos across a range of scenarios, including replay attacks and computer-generated content. Their method introduces random noise during training to simulate real-world perturbations. However, the use of noisy data during training could reduce model generalization to actual DeepFake content, particularly in real- time applications. This highlights the importance of training on clean, real-world datasets for improved detection reliability.

In *Detection of Synthetic Portrait Videos Using Biological Signals*, Ciftci et al. [5] utilized photoplethysmographic (PPG) signals—biological signals extracted from facial regions—to detect synthetic content. Their method processes authentic and fake video pairs to compute spatial coherence and temporal consistency, ultimately training a CNN and a probabilistic SVM on these features. Although effective, their technique faces challenges in formulating a differentiable loss function that accurately preserves and utilizes biological signals, particularly in the absence of an explicit discriminator.

Another notable work, *FakeCatcher*, aims to detect DeepFake content with high accuracy regardless of the generation method, video content, resolution, or quality. The approach excels at preserving biological signals and exhibits robust performance across diverse scenarios. However, the method faces limitations in implementing a differentiable loss function for signal-based features, which adds complexity to the training pipeline.

In contrast to these approaches, our proposed method integrates multiple indicators of forgery, including frame-level artifacts, temporal inconsistencies, and facial anomalies such as unnatural blinking, wrinkles, and misaligned features. By training on realistic, noise-free datasets and leveraging a combination of ResNeXt-based CNN for

feature extraction and LSTM-based RNN for temporal analysis, our method aims to address the limitations observed in existing detection frameworks and improve generalizability in real-world applications.

## 3. PROPOSED SYSTEM

The While numerous tools exist for the creation of DeepFake (DF) content, the availability of reliable and scalable detection tools remains limited. Our proposed system aims to address this critical gap by providing an accessible, web-based platform that enables users to upload a video and receive aclassification—real or manipulated. The system holds potential for expansion into browser-based plugins or integration with major social platforms such as WhatsApp and Facebook, allowing pre-distribution DF detection at the source.
This system primarily targets all major categories of DF content: replacement DF, retrenchment DF, and interpersonal DF. In addition to delivering accurate results, the system emphasizes key user-centric factors such as security, user-friendliness, accuracy, and reliability.

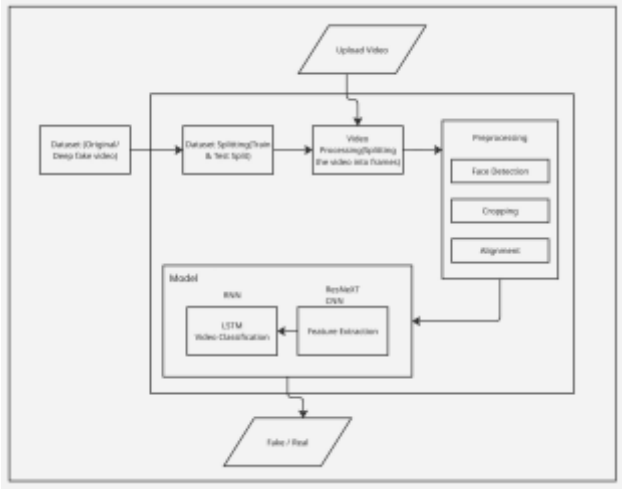


**Fig-1:** System Architecture

A. Dataset and Preparation:
We utilize a mixed dataset comprised of videos from diverse sources, including YouTube, FaceForensics++ [14], and the DeepFake Detection Challenge dataset [13]. Our custom dataset includes an equal distribution—50% authentic and 50% manipulated videos. This dataset is split into 70% training and 30% testing subsets.

B. Preprocessing:
Preprocessing involves the following sequential steps: Splitting videos into individual frames, detecting faces in each frame and cropping accordingly, discarding frames without detectable faces, normalizing the frame count for each video based on the mean video length across the dataset to ensure uniformity. Given the high computational cost of processing long videos (e.g., 300 frames for a 10-second video at 30 fps), the training phase is limited to the first 100 frames of each video to optimize performance.

C. Model Architecture:
The model consists of a ResNeXt-50 (32x4d) architecture followed by a Long Short-Term Memory (LSTM) layer. Videos, once preprocessed, are loaded using a custom DataLoader which batches them into training and testing samples. Each batch is fed into the model sequentially.

D. Feature Extraction with ResNeXt:
For efficient and robust spatial feature extraction, we adopt ResNeXt-50, a powerful convolutional neural

network. Rather than building a custom classifier from scratch, we fine- tune the ResNeXt model by adjusting hyperparameters such as learning rate and appending additional layers where necessary. The network extracts 2048-dimensional feature vectors from the final pooling layer of each frame, which are subsequently passed to the LSTM module.

E. Sequence Analysis with LSTM:

To perform temporal analysis, we utilize a 2048-unit LSTM layer with a dropout rate of 0.4. This layer processes the sequential features obtained from ResNeXt to capture inconsistencies over time, such as those introduced during frame reconstruction in DF generation. Temporal relationships are evaluated by comparing frame features across time intervals (e.g., frame at time t versus frame at t - n), enabling the model to effectively learn manipulation patterns.

F. Prediction Phase:

In the prediction stage, a new video is subjected to the same preprocessing pipeline—frame extraction, face detection, and cropping. Unlike training, the cropped frames are directly streamed to the trained model for inference, bypassing local storage to enhance performance. The model outputs a classification indicating whether the video is real or a DeepFake.

## 4. RESULT

The output of the model is going to be whether the video is deepfake or a real video along with the confidence of the model. One example is shown in the figure 2.


**Fig-2:** Expected Results

## 5. CONCLUSIONS

In this study, we presented a deep learning-based framework for the detection and classification of DeepFake videos, focusing on achieving high reliability and accuracy in real- world scenarios. Our approach is grounded in understanding the generative mechanisms of DeepFakes—primarily the useof Generative Adversarial Networks (GANs) and autoencoders—which enable the seamless synthesis of facial imagery and expressions in videos. Inspired by the pipeline used to create DeepFakes, our detection methodology employs a dual-stage neural network architecture comprising ResNeXt CNN for spatial feature extraction at the frame level and a Recurrent Neural Network (RNN) with Long Short- Term Memory (LSTM) units for capturing temporal inconsistencies across video frames.

The system processes each video by decomposing it into individual frames, detecting and cropping facial regions, and analyzing these regions using ResNeXt-50 to extract high-dimensional feature vectors. These vectors are then passed into an LSTM network that models the temporal sequence, enabling the classification of videos as real or fake based on learned inconsistencies and artifacts typically introduced during GAN-based video synthesis. The model not only provides a binary classification but also outputs a confidence score, indicating the certainty of the prediction. Our proposed solution is designed to be scalable and adaptable for integration into various platforms, including web-based systems and browser extensions, thus broadening its practical applicability. We believe that the architecture's ability to leverage both spatial and temporal cues will significantly enhance its performance, especially on real-time data encountered in dynamic environments like social media. In summary, this research contributes a robust and comprehensive system for DeepFake detection that aligns closely with the technological processes behind synthetic media creation. Future work will focus on optimizing the model for faster inference, expanding the dataset for improved generalization, and deploying the solution in real-time
detection environments.

## 3. CONCLUSIONS

The online version of the volume will be available in LNCS Online. Members of institutes subscribing to the Lecture Notes in Computer Science series have access to all the pdfs of all the online publications. Non-subscribers can only read as far as the abstracts. If they try to go beyond this point, they are automatically asked, whether they would like to order the pdf, and are given instructions as to how to do so.

## ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who supported us throughout the development of this research. We are especially thankful to our mentors and faculty members for their continuous guidance, insightful feedback, and encouragement, which played a vital role in shaping the direction of our work. We acknowledge the use of publicly available datasets such as FaceForensics++, YouTube video sources, and the DeepFake Detection Challenge dataset, which were instrumental in training and validating our deep learning models. We also appreciate the open-source tools and libraries in Python, particularly those related to Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory(LSTM) frameworks, which enabled us to experiment with and implement our proposed system effectively.

## REFERENCES

1. Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3.
2. Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.
3. Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.
4. Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.
5. Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.
6. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.
7. David G¨uera and Edward J Delp. Deepfake video detection recurrent neural networks. In AVSS, 2018.
8. https://www.kaggle.com/c/deepfake-detection challenge/data
9. https://github.com/ondyari/FaceForensics