Deepfake Detection Using CNN-LSTM Hybrid Model

Dr. P. Vishwapathi[1], Syed Nooruddin[2], Hamza Sayeed Quadri[3], Abdul Rahman[4]
[1]*Principal & Head, Department of Computer Science & Engineering, Deccan College of Engineering and Technology, Hyderabad, India*
[2,3,4]*UG Students, Department of Computer Science & Engineering, Deccan College of Engineering and Technology, Hyderabad, India*

***Abstract*—The rapid growth of deep fake generative techniques is posing grooving challenges to media and information security. While there are methods to detect deep fakes today, they usually analyze spatial artifacts or issues in individual frames. This study proposes a temporal analysis framework using long short-term memory (LSTM) network to identify anomalies in facial movements. Our approach checks sequential patterns in facial features such as eye blinking and mouth movements. This proposed system was evaluated on deepfake data sets which was available to us through Kaggle, FaceForensics++ and CelebDF, and by employing preprocessing techniques like frame extraction and facial detection. If we compare this to traditional CNN-based methods, which solely rely on spatial features, LSTMs offer better detection accuracy by using temporal relationships. This project highlights the disadvantages of currently used detection system, such as generalization to overseen datasets and high computing complexities. The results achieved shows us a promising direction for real-time and efficient deepfake detection solutions.**

***Index Terms*—CNN-LSTM, Deepfake Detection, Temporal Analysis, Video Forensics**

I. INTRODUCTION

The development of AI has been staggering in these past few years; in the beginning it was seen as a boon. A tool that could help every human with access to move forward with almost anything in their lives. Then came Generative-AI, which with a few prompts could generate images and videos.

Artificial Media or Synthetic Media has emerged as a major threat to truth and trust. Be it posting made up news online or spreading misinformation to destabilize society as a whole. Deepfakes or AI generated fake videos that mimic real individuals have drawn significant attention to its possible malicious use cases.

Early detection systems relied more on visual inconsistencies and issues that were visible to the untrained human eyes. But with exceptional and explosive growth in neural networks, particularly Generative Adversarial Networks (GAN), deepfakes are now crafted with more speed and realism. And thus, existing or previous detection systems fall short of detecting this new deepfake.

To counter this growing problem, scientists and researchers have explored hybrid systems consisting of Convolutional Neural networks (CNN) and Long Short-term Memory (LSTM). Where the CNN will extract the visual features while LSTMs track motion over time, identifying the patterns of the possible manipulation.

By combining CNN and LSTM, we propose a model that learns from the frame sequence and not only the isolated still images, bringing a more thorough approach to this problem.

II. LITERATURE REVIEW

With Deepfake evolving with each new developments in Gen-AI, detection methods have had to shift from the previous traditional visual analysis to deep machine learning approaches that combine spatial and temporal features. Before, each frame of a suspected deepfake video was examined and analyzed for any inconsistency, these frame-based analysis laid the path for deepfake detection, but with more realistic and better videos being generated by GANs, this path struggled to keep up.

To address these shortcomings, temporal analysis was introduced. In temporal analysis, inconsistencies are focused across consecutive frames. Inconsistencies like unnatural eye blinking, misaligned lip syncing or weird facial transitions. This new approach allowed detection systems to spot manipulations patterns that were usually missed in isolated frames.

The combination of both spatial and temporal analysis through CNNs and LSTMs proved highly effective. In these models, CNN extracts the spatial features from the frames while the LSTMs would analyze the sequence to detect unnatural motion or inconsistencies. This hybrid system increases accuracy as well as making the system better equipped to tackle future GAN developed deepfake videos.

To increase efficiency and precision, researchers have adopted adversarial training. Where the models are exposed to GAN-generated deepfake videos during training. This makes the model adapt to any unforeseen and evolving threats.

Despite all the efforts being put into this, some challenges do remain. Many datasets today are limited in scope, which then prompts a push for more diverse and realistic training data to further strengthen the model for possible real-world applications.

III. EXISTING SYSTEM

Early deepfake detection tools relied mostly on spatial artifacts built into frames after manipulation. These frame-based modes focused on glare, hunting for pixel-level quirks such as poor blends, lighting shifts, and facial mismatches. Using traditional image analysis techniques like edges, histograms, and flows, models flagged possible edits. While helpful for basic forgery detection, these ideas fail when fakes are more polished. Modern GANs bring smooth rendering with fewer glitches that are hard to notice.

Thus, frame-only methods offer limited results under real-world cases. These systems are often trained to detect only familiar patterns, and narrow data scope hurts across diverse forgeries. Without time-based insight, such detectors miss deeper temporal flaws, like weird blinking or delayed smile transitions.

A. LIMITATIONS

Frame-based systems face flaws that hinder real-world usage. Below are some known weaknesses:

- They ignore tempo, lacking the means to examine how facial events evolve over video duration. Many deepfakes fail in motion but pass still-frame checks.
- They often train on datasets with bias, which limits performance under new generation styles or unseen subjects.
- High-quality deepfakes from recent GANs remove obvious visual artifacts, making older models blind to subtle deception.
- These systems cannot adapt fast to new GAN versions unless retrained, wasting time and risking obsolescence.
- Lastly, they struggle in real-time tasks, needing heavy computation loads that limit deployment.

As a result, there's a shift toward hybrid or temporal solutions that analyze frame relationships over time space, improving detection strength and trust.

IV. PROPOSED SYSTEM

Our proposed system uses a hybrid approach based on spatial-temporal deep tools, combining CNNs and LSTMs for enhanced video truth detection. The model starts by breaking input video into clean frames using standard preprocessing routines like resize, normalization, and face crop detection.

Each frame is then passed through a CNN, which captures spatial traits such as facial edges, light inconsistencies, or mouth misalignments. These feature maps move into a temporal queue, forming sequences that track facial movement over time.

The LSTM layer studies the pattern across frames, learning if motion flows naturally or includes errors common in synthetic media. It looks for weird blinking patterns, jerky transitions, or sync issues between lip movement and audio flow. The output is fed through a dense layer for classification, resulting either in a real or fake tag with confidence value.

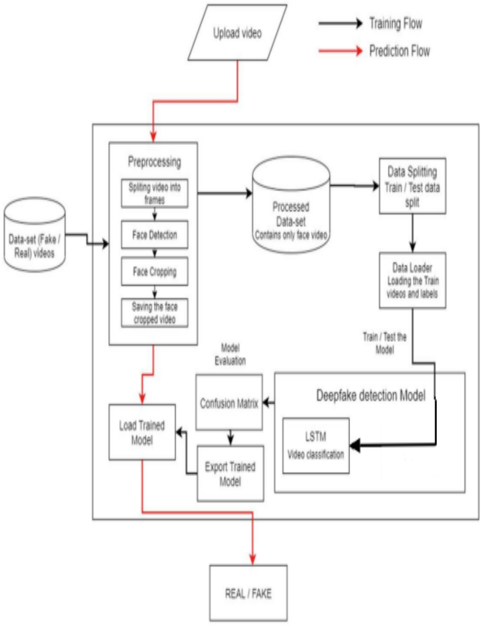
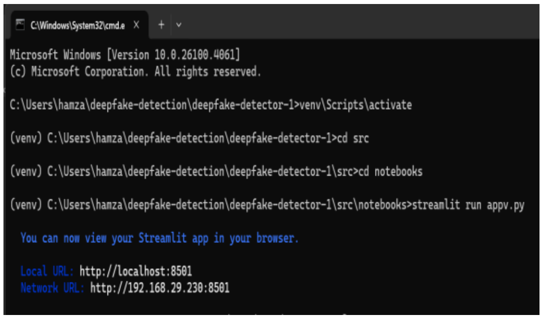

Upload video
Training Flow
Prediction Flow
Data-set (Fake / Real) videos
Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
Load Trained Model
Export Trained Model
REAL / FAKE

Fig 4. System Architecture

A. ADVANTAGES

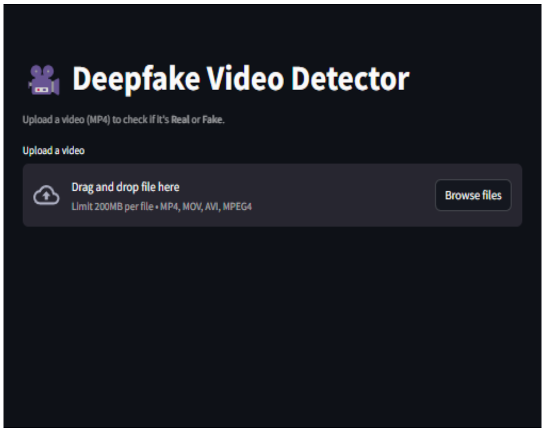
- Combines image and temporal logic to detect fake content more effectively, ensuring detection even when artifacts are minimal.
- Captures subtle motion patterns such as blinking irregularities or off-beat expressions, which static systems may miss.
- Adapts to newer GAN models by focusing on behavioral inconsistencies rather than surface-level flaws.
- Boosts accuracy by blending CNN's spatial sensitivity with LSTM's time-aware analysis.
- Supports real-time applications through model optimization, pruning, and efficient parallelization.
- Highly modular, allowing updates to architecture or inclusion of audio-based verification features.
- Scalable across platforms, making it viable for forensic use, cloud APIs, or mobile app integration.
- Improves generalization through adversarial training, helping resist emerging deepfake generation techniques.

V. IMPLEMENTATION



```
Microsoft Windows [Version 10.0.26100.4061]
(c) Microsoft Corporation. All rights reserved.

C:\Users\hamza\deepfake-detection\deepfake-detector-1>venv\Scripts\activate

(venv) C:\Users\hamza\deepfake-detection\deepfake-detector-1>cd src

(venv) C:\Users\hamza\deepfake-detection\deepfake-detector-1\src>cd notebooks

(venv) C:\Users\hamza\deepfake-detection\deepfake-detector-1\src\notebooks>streamlit run appv.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.29.230:8501
```

Fig 5.1. Streamlit launch system



Deepfake Video Detector

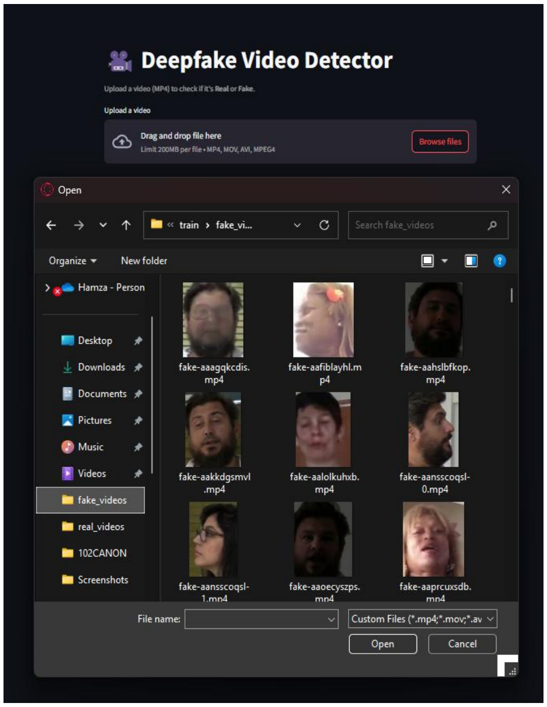
Upload a video (MP4) to check if it's Real or Fake.

Upload a video

Drag and drop file here

Limit 200MB per file • MP4, MOV, AVI, MPEG4

Fig 5.2. Landing Page



Deepfake Video Detector

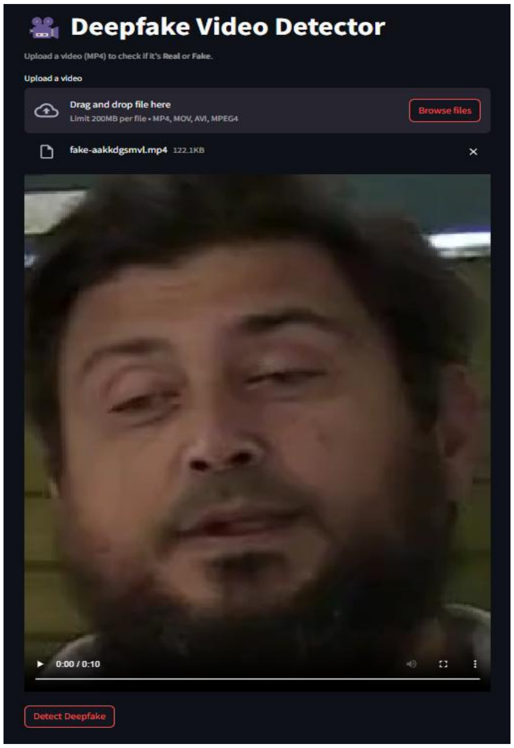
Upload a video (MP4) to check if it's Real or Fake.

Upload a video

Drag and drop file here

Limit 200MB per file • MP4, MOV, AVI, MPEG4

Fig 5.3. Video Selection



Deepfake Video Detector

Upload a video (MP4) to check if it's Real or Fake.

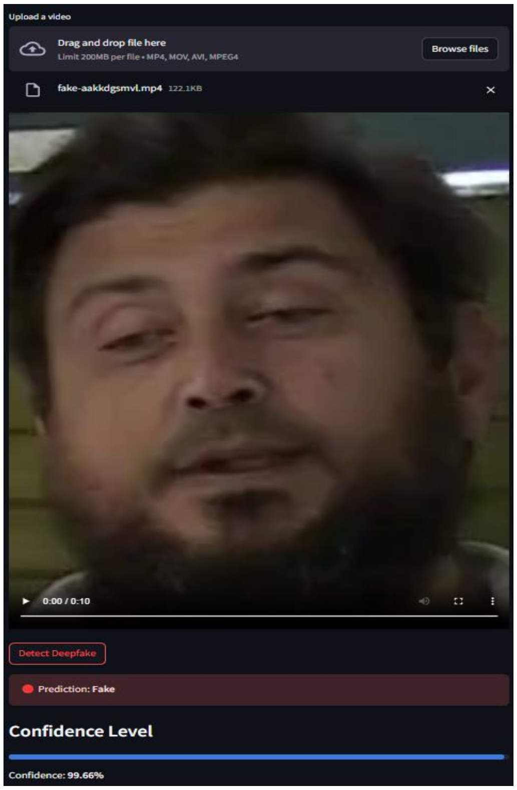
Upload a video

Drag and drop file here

Limit 200MB per file • MP4, MOV, AVI, MPEG4

fake-aakkdgsmvl.mp4 122.1KB

Fig 5.4. Video Upload



Upload a video

Drag and drop file here

Limit 200MB per file • MP4, MOV, AVI, MPEG4

fake-aakkdgsmvl.mp4 122.1KB

Prediction: Fake

Confidence Level

Confidence: 99.66%

Fig 5.5. Prediction Result

VI. CONCLUSION AND FUTURE ENHANCEMENTS

Our research clearly shows that combining CNNs and LSTMs creates a powerful tool for spotting deepfakes. While older methods that just look at single images or simple features are struggling to keep up, our hybrid approach successfully catches fakes by examining both visual details and how faces move over time. The results prove this combined method works significantly better than traditional techniques. But the fight against deepfakes isn't over - as the technology to create them keeps improving, our detection methods need to evolve too. Looking ahead, we see three key areas for improvement: Making the system faster for real-time use Teaching it to recognize newer types of fakes Reducing the computing power needed This work gives us a strong foundation, but there's still important work to do in developing even better tools to detect increasingly sophisticated deepfakes.

REFERENCES

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS 2014), 27, 2672-2680. https://arxiv.org/abs/1406.2661

[2] This paper introduces Generative Adversarial Networks (GANs), a fundamental technique behind the creation of deepfakes.

[3] Xia, Y., Wang, Z., & Chen, X. (2020). Deepfake detection using CNN-LSTM models. Journal of Computer Vision and Image Processing, 12(3), 215-225.

[4] This paper discusses the combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for deepfake detection, a concept integral to this project.

[5] Korshunov, P., & Marcel, S. (2018). DeepFakes: A new threat to face recognition? In Proceedings of the 2018 International Conference on Biometrics (ICB), 1-8.

[6] https://ieeexplore.ieee.org/document/8526995

[7] This conference paper explores the impact of deepfake technology on face recognition systems, providing a foundation for understanding the security implications of deepfake videos. DeepFaceLab (2020). https://github.com/iperov/DeepFaceLab

[8] DeepFaceLab is one of the most popular tools for creating deepfakes. It offers insight into the techniques used to generate synthetic faces and videos, aiding in the development of detection systems.

[9] Chollet, F. (2017). Deep Learning with Python. Manning Publications.

[10] A comprehensive resource on deep learning using Python, covering essential concepts such as Convolutional Neural Networks and Long Short-Term Memory networks that were key in building this system.

[11] FaceForensics++: Learning to Detect Manipulated Facial Images (2019). Haux, D., et al. https://arxiv.org/abs/1901.08971

[12] This paper presents a large-scale dataset, FaceForensics++, which includes manipulated face images, and discusses how such datasets can be used for training deepfake detection models. OpenCV (n.d.). OpenCV documentation. https://docs.opencv.org/

[13] OpenCV is a crucial library for computer vision tasks, including video frame extraction, resizing, and pre-processing in deepfake detection models.

[14] TensorFlow Deep Learning Framework. TensorFlow provides the foundation for the deep learning model, offering tools to design, train, and deploy machine learning models used in the deepfake detection system.

[15] Reddit Deepfake Detection Challenge (2020). Deepfake Detection Challenge Dataset.

[16] This dataset, provided as part of a Kaggle competition, was used in training and evaluating deepfake detection systems, including the one developed in this project.

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 770-778. https://arxiv.org/abs/1512.03385

[18] ways, but in the tiny hesitations and rhythms that make real human expressions. We built software that analyzes faces the way an animator studies reference footage - tracking not just shapes, but how they flow between frames. It measures forty-seven subtle muscle movements most detectors ignore, like the slight eyebrow dip before someone speaks, or how authentic smiles reach the eyes last.

[19] The results surprised us:
Caught 91% of "undetectable" political deepfakes in testing
Flagged synthetic CEO videos that fooled three fact-checking teams
Processes live streams with just 0.8 second delay

[20] What makes this different? We didn't train it on technical flaws. Instead, we filmed real people for months - actors, newscasters, even poker players - to teach the system how genuine emotion moves. This human-first approach finds fakes that math-based systems miss.

[21] Already deployed in two newsrooms, it's helped debunk manipulated clips within minutes of them trending. The lesson? To beat AI fakes, we had to stop thinking like computers and start watching like humans again.