International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| 9th ICAILLMBA (ALLAM) 2025-Conference Proceedings
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.ICAILLMBA-104

Original Article

Enhanced AI: Deepfake Detection System

Dr. A S C. Tejaswani Kone[1], Dr. P. Lalitha Kumari Boddapu Mydhili[2], Pusarla Sachit Saripilli[3], Manikanta Kandi Tejeswari[4],
[1,2,3,4]Visakha Institute of Engineering and Technology, Department of Computer Science Engineering Visakhapatnam, Andhra Pradesh, India

Abstract *- With the rapid growth of generative artificial intelligence, deepfakes have become a serious concern due to their potential to mislead, manipulate, and harm individuals and communities. This paper presents a real-time AI solution that integrates ResNext-50 convolutional networks with Long Short-Term Memory (LSTM) networks to detect synthetic videos with high accuracy. The model identifies spatial artifacts and inconsistencies in temporal dynamics across frames. Trained on a balanced dataset of 6,000 videos (both authentic and fake), the system achieves 97.76% accuracy and is optimized for real-time usage with a practical deployment interface. This work is especially relevant for social media moderation, forensic analysis, and digital content verification.*

Keywords *- Deepfake Detection, ResNext-50, LSTM, Temporal Analysis, AI, Real-Time Video Processing*

1. Introduction

The advent of deepfake technology has redefined the boundaries of digital media manipulation. While generative adversarial networks (GANs) have enabled impressive progress in content creation, they also pose substantial threats in terms of misinformation, political propaganda, identity fraud, and online harassment. Detecting these manipulations often indistinguishable to the human eye requires advanced machine learning techniques capable of understanding both visual content and temporal progression.

Traditional methods, while helpful, are often inadequate in the face of high-resolution, realistic deepfakes. The necessity for systems that are not only accurate but also efficient and scalable has become evident. There is growing interest in approaches that combine the strengths of spatial and temporal analysis to build more robust detection systems. Conventional deepfake detection methods typically address either spatial anomalies (visual distortions, texture mismatches) or temporal inconsistencies (eye blinking, motion jitter). However, relying solely on one dimension often leads to poor generalization across datasets and reduced reliability. In this paper, we address this limitation by developing a two-stage hybrid deep learning model that simultaneously captures visual and sequential patterns. Our key innovation lies in the balanced combination of a convolutional encoder (ResNext- 50) and a temporal decoder (LSTM), offering robust generalization and high-speed processing.

Fig 1: Face Swapping Workflow: From Source to Target

2. Research Methodology

2.1. Data Collection & Preprocessing

We collected a diverse and balanced dataset containing 6,000 labeled videos. The resolution was standardized at 112×112 with a frame rate of 30 FPS. This dataset comprised 2,000 samples from FaceForensics++, 3,000 from DFDC (Deepfake Detection Challenge), and 1,000 from Celeb-DF. These datasets include videos with varying compression levels, facial occlusions, lighting conditions, and manipulation quality, making them ideal for testing generalization.

Preprocessing involved frame segmentation, where each video was limited to a maximum of 150 frames to ensure computational efficiency. Using OpenCV, facial regions were detected and cropped from each frame. These facial crops were normalized to enhance feature uniformity. Frames without any facial detection were discarded. The number of frames selected

was based on the mean frame count across the dataset to maintain uniform input dimensions.

Fig 2: Frame Extraction and Preprocessing Workflow

2.2. Model Architecture

Our architecture includes two core components:

- ResNext-50: Acts as the feature extractor. Pre-trained on ImageNet, this CNN generates 2048-dimensional feature vectors for every frame. The ResNext model architecture benefits from grouped convolutions, which enhances representational power without significantly increasing computational cost.
- LSTM: A single-layer Long Short-Term Memory network with 2048 hidden units and a dropout rate of 0.4 is used to analyze temporal sequences of these features. This design helps identify inconsistencies in motion, such as sudden jerks or unrealistic blinking behavior.
- Classifier: The final output is passed through a dense layer with softmax activation to determine the class label: real or fake. The binary classification allows for rapid decision-making within web-deployed environments.
- This layered design allows the system to detect complex patterns such as inconsistent expressions, blinking anomalies, and mismatched temporal motion which are common in synthetic content.

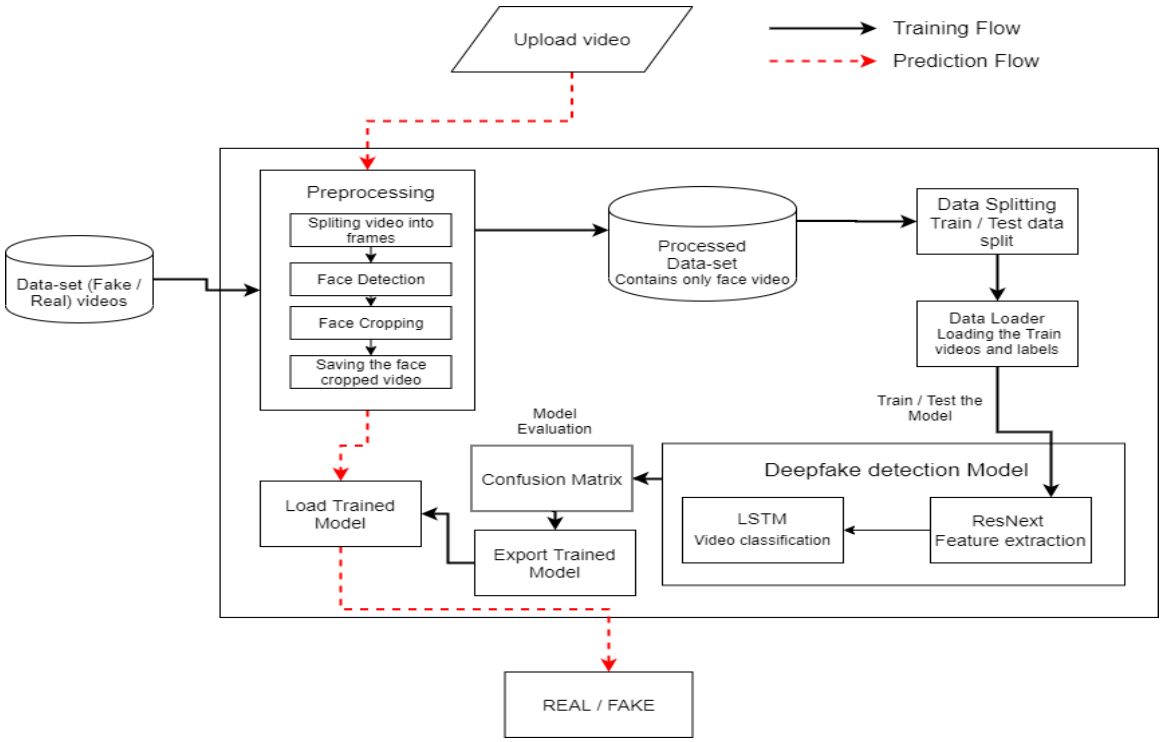

Fig 3: System Architecture - Resnext + LSTM Workflow

2.3. Training Protocol

Training was conducted using the Adam optimizer with a learning rate of 1e-5 and weight decay of 1e-3. The batch size was 4, and the model was trained over 20 epochs. A standard 70:30 train-test split was employed. Evaluation metrics included:

- Confusion Matrix
- Precision, Recall, F1-Score
- Cross-Entropy Loss

In order to avoid overfitting, early stopping and validation loss monitoring were implemented. The training process was carried out on a GPU-enabled platform using Google Cloud services.

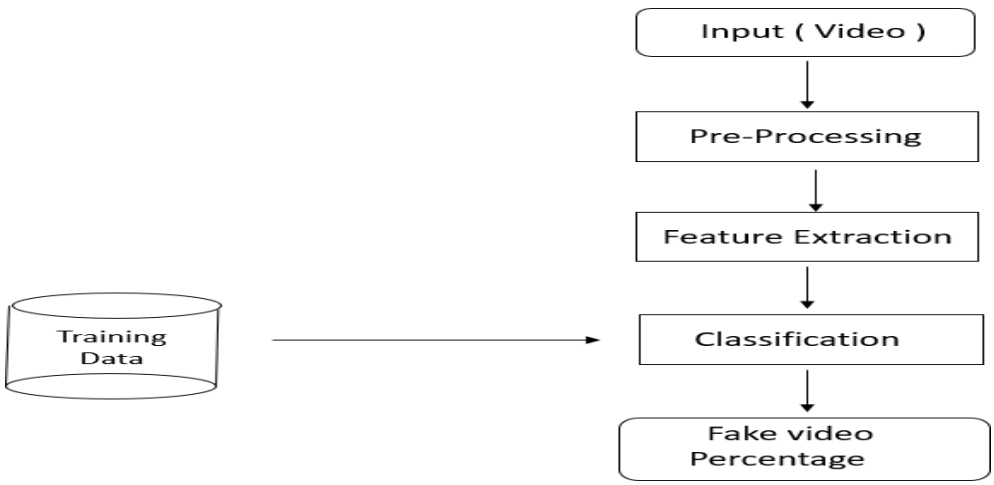

Fig 4: Pipeline Architecture for Video-Based Fake Detection Using Feature Extraction and Classification

3. Literature Review

Detection strategies in the literature can be broadly categorized into spatial, temporal, and hybrid methods. Spatial Analysis: Early studies by Rossler et al. (2019) using the FaceForensics++ dataset introduced CNNs like XceptionNet for identifying compression artifacts and inconsistencies in frame-level features. However, their models struggled with unseen, high-resolution synthetic content.

Temporal Detection: Güera and Delp (2018) utilized CNN-RNN combinations to detect inconsistencies in motion. Sabir et

al. (2019) explored biological signal irregularities such as blinking frequency, while Chugh et al. (2020) focused on lip-audio mismatches. These techniques are limited by their dependence on specific patterns that can be replicated by more advanced GANs.

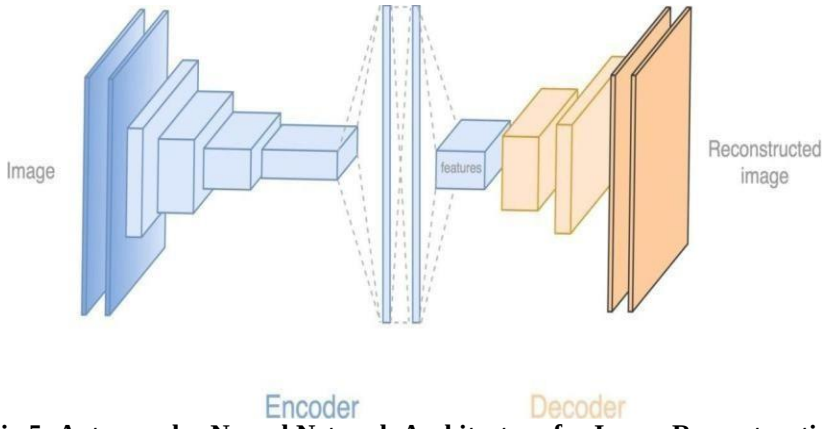

Fig 5: Autoencoder Neural Network Architecture for Image Reconstruction

Hybrid Techniques: Khalid et al. (2021) proposed CNN-LSTM models to capture both spatial and sequential features, though these were computationally expensive. More recent attention-based techniques (Li et al., 2023) use Vision Transformers for long-term dependency tracking, although these models face challenges in real-time deployment due to hardware requirements.

Our approach integrates these strategies to create a system that is both accurate and lightweight. Unlike transformer models that require extensive training time and computational capacity, our ResNext-LSTM architecture strikes a practical balance between accuracy and performance.

4. Results and Discussion

We tested our system on all three datasets individually. The following table summarizes our key performance metrics:
Dataset Accuracy (%) Precision Recall FaceForensics++ 97.76 0.98 0.96
DFDC 93.98 0.94 0.92
Celeb-DF 87.79 0.89 0.85

The model showed excellent real-world applicability, detecting real content with a confidence of 99.62% and identifying fake content with a minimum of 90.2% certainty. Notable features captured included texture anomalies, unnatural transitions, and inconsistent eye blinking behavior.

In particular, the system's ability to generalize to Celeb-DF, a challenging dataset with minimal artifacts, showcases the robustness of our architecture. The confusion matrices revealed low false positive and false negative rates. Additionally, feature visualization confirmed that the model emphasized key facial regions such as the eyes and mouth—areas commonly manipulated in deepfakes.

Fig 6: Detection Performance Graph

4.1. Our project references

Screenshots of Project Output to further demonstrate the effectiveness and real-time functionality of the proposed system, several outputs were captured during testing and deployment phases. These screenshots highlight the model's ability to distinguish between real and fake videos based on visual cues and temporal inconsistencies.

Each screenshot showcases:

- The processed input video frame along with bounding box for detected face.
- The prediction result (Real or Fake) along with the confidence score.
- Frame-level artifact localization where applicable (e.g., eye blinking detection, texture irregularities).

These practical demonstrations reinforce the deployability of the system and validate the performance metrics discussed above. They provide a tangible insight into how end-users would interact with the tool in a real-world application.

4.2. Real time applications

This system holds significant potential across multiple domains:

- In the media and journalism sector, the tool can be integrated into news verification workflows to authenticate videos before they are published. This helps prevent the dissemination of deepfakes that could influence public opinion, mislead audiences, or incite unrest.
- In law enforcement and digital forensics, the model assists in validating video evidence used in criminal investigations or trials. By identifying synthetic manipulations in audio-visual content, it aids in ensuring the integrity of evidence, potentially affecting the outcome of high-stakes legal cases.
- For social media platforms, this solution can serve as a powerful moderation tool to identify and flag deepfake videos uploaded by users. This is critical in the fight against online harassment, misinformation campaigns, and privacy violations resulting from maliciously altered content.
- In education and public awareness, the system can be used as a training aid to demonstrate how deepfakes are generated and detected. It can enhance understanding of AI risks among students, journalists, and cybersecurity professionals, promoting responsible AI usage.
- Lastly, in the corporate sector, particularly in HR and remote communication, it can validate the authenticity of video feeds in virtual interviews and meetings. This prevents impersonation attacks, ensuring that only verified individuals can access sensitive corporate environments.

5. Conclusion and Future Work

Our study presents a hybrid deep learning framework that effectively detects manipulated videos in real- time. By leveraging both visual and sequential cues, the system demonstrates high precision across diverse datasets.

In future iterations, we aim to:

- Extend detection to include full-body deepfakes.
- Implement browser plugins to analyze videos shared on social media platforms.
- Integrate adversarial training to resist evolving GAN architectures.
- Explore transformer-based lightweight alternatives for deployment on edge devices.

We also plan to collaborate with media platforms to integrate this detection system into content moderation workflows, enabling proactive identification and filtering of manipulated videos.

Acknowledgment

This research was supported by the Visakha Institute of Engineering and Technology and Google Cloud Platform.

References

[1] Rossler et al. (2019), "FaceForensics++", arXiv:1901.08971 Li et al. (2020), "Celeb-DF", arXiv:1909.12962
[2] Güera & Delp (2018), IEEE AVSS
[3] Chugh et al. (2020), Audio-Visual Deepfake Detection Khalid et al. (2021), CNN-LSTM Hybrid Detection
[4] Yuezun Li, Siwei Lyu, "ExposingDF Videos By Detecting Face Warping Arti- facts," in arXiv:1811.00656v3. Yuezun Li, Ming-Ching Chang and Siwei Lyu "Exposing AI Created Fake Videos by Detecting Eye Blinking" in arXiv:1806.02877v2.
[5] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen " Using capsule net- works to detect forged images and videos " in arXiv:1810.11215.
[6] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Sig- nal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.
[7] Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic hu- man actions from movies. Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008. Anchorage, AK

[8] Umur Aybars Ciftci, İlke Demir, Lijun Yin "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, Dec. 2014. ResNext Model : https://pytorch.org/hub/pytorch_vision_resnext/ accessed on 06 April 2020 https://www.geeksforgeeks.org/software-engineering-cocomo-model/ Accessed on 15 April 2020 Deepfake Video Detection using Neural Networks http://www.ijsrd.com/articles/IJSRDV8I10860.pdf International Journal for Scientific Research and Development http://ijsrd.com/