A CNN-LSTM BASED DEEP LEARNING MODEL FOR DETECTING MANIPULATED VIDEOS IN DIGITAL MEDIA

[1]K Suma, [2]P Sravanthi, [3] L Bhavani,[4] M Sai Charan , [5]M Abdul Auqeet Furqan,
[1]AssistantProfessor, [2345]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally

sumakadari_cse@siddhartha.co.in, 24TQ1A05H2@siddhartha.co.in, 24TQ1A05D1@siddhartha.co.in, 24TQ1A05H7@siddhartha.co.in, 24TQ1A05F2@siddhartha.co.in

Abstract

Deepfake videos generated using advanced artificial intelligence techniques pose serious threats to digital security, privacy, and information authenticity. The increasing misuse of manipulated videos in social media, digital fraud, misinformation, and cybercrime has created a strong need for reliable deepfake detection systems. In this work, a hybrid CNN-LSTM-based deepfake detection framework is proposed to identify manipulated videos by analyzing both spatial and temporal inconsistencies. The system begins with video preprocessing, frame extraction, and normalization, followed by spatial feature extraction using convolutional neural networks to capture facial texture artifacts, blending errors, and visual distortions.

The extracted frame-level features are then passed to a Long Short-Term Memory (LSTM) network to learn temporal dependencies such as lip-sync mismatches, abnormal facial movements, and frame flickering. The fused spatiotemporal features are classified using dense and Softmax layers to predict whether the video is real or fake. The proposed system improves detection accuracy while maintaining computational efficiency through key frame selection and preprocessing. The model also provides user-friesult visualization, making it suitable for applications such as social media verification, digital forensics, cybersecurity, and multimedia content authentication.

Keywords: Deepfake Detection, Hybrid CNN-LSTM, Video Forgery Detection, Spatial Feature Extraction, Temporal Sequence Analysis, Convolutional Neural Network, Long Short-Term Memory, Multimedia Forensics, Fake Video Classification, Artificial Intelligence, Digital Media Authentication, Video Manipulation Detection.

I. Introduction

Deepfake technology has emerged as one of the most significant challenges in the field of digital media security. With the rapid advancement of artificial intelligence and deep learning techniques, highly realistic fake videos can now be generated with minimal effort. These manipulated videos can alter facial expressions, lip movements, voice synchronization, and even identity, making them difficult to distinguish from genuine content. The widespread availability of deepfake generation tools has increased the risk of misinformation, cyber fraud, privacy violations, and reputational damage, creating an urgent need for reliable deepfake detection systems.

The growing influence of social media and digital communication platforms has further amplified the impact of deepfake videos. Manipulated media can spread rapidly across online platforms, misleading users and causing social, political, and economic consequences. Deepfakes can be misused in areas such as fake news dissemination, identity theft, online scams, and digital blackmail. As a result, ensuring the authenticity of multimedia content has become essential for maintaining public trust, digital safety, and information integrity.

Traditional methods for detecting manipulated videos often rely on manual inspection or basic image processing techniques. However, these approaches are no longer sufficient to identify advanced deepfake content, especially when videos are created using sophisticated generative adversarial networks (GANs) and face-swapping models. Modern deepfake videos contain subtle spatial and temporal inconsistencies that require intelligent systems capable of learning complex visual and sequential patterns. Therefore, deep learning-based detection methods have become a promising solution for addressing this challenge.

II. Literature Survey

1] Zhang et al. – Temporal Feature Prediction in Audio–Visual Deepfake Detection : This study used the FakeAVCeleb dataset and proposed a dual-stream temporal prediction model with contrastive learning for audio-video forgery detection. It achieved 84.33% accuracy and 89.91% AUC. This work is related to our idea as it highlights the importance of temporal inconsistencies across modalities, which supports our focus on robust spatiotemporal deepfake analysis.

[2] Gu et al. – Deepfake Video Detection via Predictive Representation Learning : Using FaceForensics++, DFDC, and Celeb-DF, the authors proposed Latent Pattern Sensing with CNN, ConvGRU, and self-distillation. The model achieved 99.94% AUC on FaceForensics++. This research is relevant to our work because it proves predictive spatiotemporal feature learning can significantly improve deepfake detection robustness.

[3] Zhang et al. – Temporal Feature Prediction in Audio–Visual Deepfake Detection : This work introduced bimodal temporal feature prediction using dual-stream audio-video learning on FakeAVCeleb. By aligning temporal audio and visual cues with contrastive loss, it improved detection performance. The study relates to our idea by emphasizing multimodal temporal consistency, which can inspire our system to capture subtle forged patterns across video sequences.

[4] Almutairi et al. – Real-Time Advanced Computational Intelligence for Deep Fake Video Detection : The authors proposed Deep Fake Network (DFN) using MobileNet blocks and XGBoost classifier on the DFDC dataset. The model achieved 93.28% accuracy with low computational cost. This work is related to our idea because it demonstrates that lightweight real-time architectures can provide efficient and deployable deepfake detection solutions.

[5] Hu et al. – FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos : This study proposed FInfer, which predicts future facial frame representations using autoregressive learning to detect temporal inconsistencies in high-quality deepfakes. It showed strong in-dataset and cross-dataset performance. This is relevant to our idea because it validates frame prediction as an effective strategy for detecting subtle manipulations in realistic forged videos.

[6] Rahman et al. – Deepfake Video Prediction Using Attention-Based CNN and MFCC : This study used the FakeAVCeleb dataset and combined Attention-CNN

with MFCC audio features for multimodal deepfake detection. The model achieved 94% accuracy. This work is related to our idea because it proves that combining audio and visual cues improves detection of sophisticated forged videos.

[7] Amerini et al. – Exploiting Prediction Error Inconsistencies through LSTM-based Classifiers : The authors used inter-frame prediction error features with an LSTM classifier to detect forged portrait videos. Their sequence-based method captured temporal inconsistencies effectively and showed promising results. This study supports our idea by showing that temporal artifacts across video frames are useful indicators for deepfake detection.

[8] Ge et al. – Sharp Multiple Instance Learning for DeepFake Video Detection : This work proposed Sharp Multiple Instance Learning (S-MIL) with spatiotemporal encoded instances for partially manipulated deepfake videos. Using the DFDC and FFPMS datasets, it outperformed existing methods. It is relevant to our idea as it demonstrates effective learning of subtle local and temporal forgery clues.

[9] Sharma et al. – Zero-shot Visual Deepfake Detection : This study explored zero-shot deepfake detection using self-supervised learning, transformers, meta-learning, and generative fingerprinting. It focused on detecting unseen manipulations without prior training. This work relates to our idea because it highlights the importance of generalization and adaptability for future-proof deepfake detection systems.

[10] Tolosana et al. – A Survey on Deepfake Video Detection

This survey reviewed deepfake generation methods, datasets, and detection techniques, emphasizing challenges in robustness and real-world deployment. It concluded that existing systems still struggle with generalization. This paper is related to our idea as it identifies research gaps that our proposed method aims to address for more reliable detection.

[11] **Li et al. – Facial Muscle Motions for Detecting Compressed Deepfake Videos :** This study proposed the FAMM framework using facial landmark motion features and Dempster–Shafer fusion for compressed deepfake detection. It showed strong performance on social media compressed videos. This work relates to our idea because it proves facial motion dynamics remain reliable cues even under video compression and real-world transmission noise.

[12] **Reddy et al. – Deep Fake Image and Video Detection using Deep Learning :** The authors used Meso4 for image analysis and LRCN for video sequence modeling on benchmark deepfake datasets. Their method effectively captured spatial and temporal artifacts. This study is relevant to our idea because it validates hybrid deep learning models for accurate detection of manipulated multimedia content.

III. System Analysis

The proposed system focuses on detecting manipulated or fake videos using a hybrid deep learning approach combining CNN and LSTM models. With the rapid growth of digital media, video manipulation techniques such as deepfakes have become more sophisticated and harder to detect. The system analyzes both spatial and temporal features of videos to identify inconsistencies. Convolutional Neural Networks (CNN) are used to extract spatial features from individual frames, while Long Short-Term Memory (LSTM) networks capture temporal dependencies across frames. The system processes video input by breaking it into frames and analyzing patterns over time. It ensures high accuracy by learning complex visual cues and motion irregularities. The model is trained on labeled datasets of real and manipulated videos. It supports

scalability for large video datasets. The system provides automated and efficient detection, reducing reliance on manual verification.

Existing System

Existing systems for detecting manipulated videos mainly rely on manual inspection or basic image processing techniques. Some approaches use simple machine learning models that analyze individual frames without considering temporal relationships. These systems often fail to detect advanced deepfake videos due to lack of contextual understanding. Traditional methods focus only on facial features or pixel-level inconsistencies. They are not capable of capturing motion-based anomalies across frames. Many systems lack automation and require human intervention for verification. They are not scalable for large volumes of video data. The absence of deep learning models limits their accuracy. Existing approaches also struggle with variations in lighting, resolution, and compression. As a result, detection performance is often unreliable.

Disadvantages of Existing System

- Depends on manual inspection
- Low accuracy for advanced deepfakes
- Ignores temporal information in videos
- Cannot detect motion inconsistencies
- Limited scalability
- Not robust to variations in video quality
- High false positive and false negative rates
- Lacks automation
- Inefficient for real-time detection

Proposed System

The proposed system introduces a CNN-LSTM based deep learning model for accurate detection of manipulated videos. It extracts spatial features from video frames using CNN and captures temporal relationships using LSTM. The system first converts videos into sequences of frames. Each frame is processed by the CNN to extract meaningful visual features. These features are then passed to the LSTM network to analyze sequential dependencies. The model learns patterns of both genuine and manipulated videos during training. It is capable of identifying subtle inconsistencies such as unnatural facial movements and temporal artifacts. The system supports automated detection and reduces human effort. It can handle large datasets efficiently. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Overall, it provides a robust and scalable solution.

Advantages of Proposed System

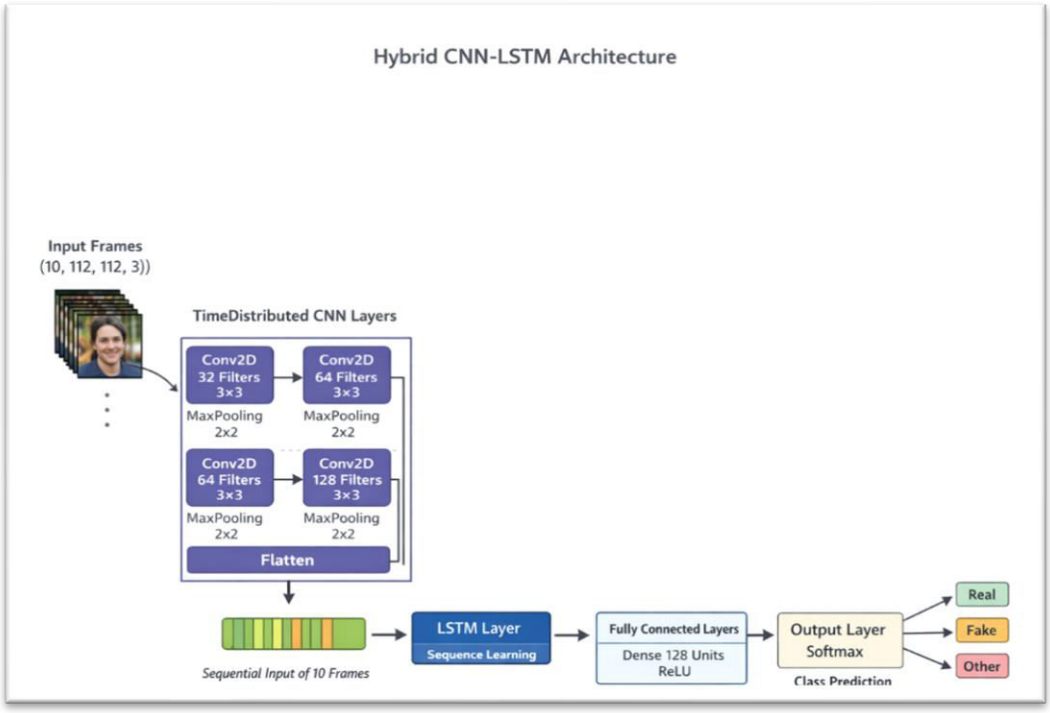
- High detection accuracy using deep learning
- Combines spatial and temporal analysis
- Fully automated system
- Effective against advanced deepfake techniques

- Scalable for large video datasets
- Reduces human effort
- Robust to variations in video quality
- Captures motion-based inconsistencies

IV. Methodology

The methodology begins with collecting a dataset of real and manipulated videos. Videos are converted into frames using frame extraction techniques. Preprocessing is applied to normalize images and remove noise. The CNN model extracts spatial features from each frame. These features are passed to the LSTM model to learn temporal relationships across frames. The combined model is trained using labeled data. The dataset is split into training and testing sets. Model performance is evaluated using accuracy, precision, recall, and F1-score. Hyperparameter tuning is applied to improve results. The trained model is deployed for detecting manipulated videos in real time. The system uses publicly available datasets such as FaceForensics++, DeepFake Detection Challenge (DFDC), or custom-collected video data. The dataset contains both real and manipulated videos. Data balancing techniques are applied to ensure equal representation of classes. Data is labeled properly for supervised learning. Videos are converted into frames using frame extraction techniques. Instead of using all frames, **frame sampling** is applied (e.g., every nth frame) to reduce computational cost while preserving important temporal information. Keyframes may also be selected based on motion or scene changes.

System Architecture



The system architecture for the CNN-LSTM based manipulated video detection model is designed as a sequential pipeline that processes video data through multiple

stages. Initially, the system takes a video as input, which is then passed to a frame extraction module where the video is divided into individual frames. These frames undergo preprocessing steps such as resizing, normalization, and noise reduction to ensure consistency and improve model performance. The processed frames are then fed into a Convolutional Neural Network (CNN), which extracts important spatial features like facial details, textures, and visual artifacts. The extracted features are subsequently passed to a Long Short-Term Memory (LSTM) network, which analyzes temporal dependencies and sequential patterns across frames. This combination enables the system to capture both spatial and temporal inconsistencies present in manipulated videos. The output from the LSTM is then sent to a fully connected classification layer, where the system predicts whether the video is real or fake using an activation function such as sigmoid or softmax. Finally, the result is displayed to the user, and the system may optionally store the input and output data for further analysis.

V. Result and Output

Layer (type)	Output Shape	Param #
time_distributed (TimeDistributed)	(None, 10, 110, 110, 32)	896
time_distributed_1 (TimeDistributed)	(None, 10, 55, 55, 32)	0
time_distributed_2 (TimeDistributed)	(None, 10, 53, 53, 64)	18,496
time_distributed_3 (TimeDistributed)	(None, 10, 26, 26, 64)	0
time_distributed_4 (TimeDistributed)	(None, 10, 43264)	0
lstm (LSTM)	(None, 64)	11,092,224
dense_8 (Dense)	(None, 64)	4,160
dense_9 (Dense)	(None, 5)	325

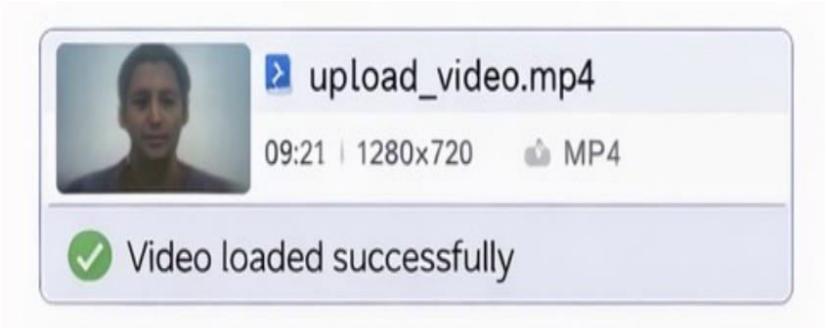
```
uploaded_file = upload.value[0]

video_path = "/kaggle/working/uploaded_video.mp4"

with open(video_path, "wb") as f:
    f.write(uploaded_file['content'])

print("Video saved at:", video_path)
```

```
fake_video_path = os.path.join(fake_folder, videos[0])
print(fake_video_path)
```



VI. Conclusion

The proposed hybrid CNN-LSTM deepfake detection system provides an effective and reliable solution for identifying manipulated videos by combining spatial and temporal feature analysis. The CNN component successfully extracts frame-level artifacts such as facial texture inconsistencies, blending errors, and lighting mismatches, while the LSTM network captures temporal abnormalities like lip-sync mismatch, unnatural facial movements,

and frame flickering. This integrated approach improves detection accuracy and makes the system more robust than conventional frame-based methods. The use of key frame extraction and preprocessing reduces computational complexity while maintaining performance. The system also offers a user-friendly interface for clear result visualization, making it suitable for practical applications such as social media verification, digital forensics, and cybersecurity. Although certain limitations exist, such as sensitivity to video quality and evolving deepfake techniques, the proposed model provides a strong foundation for secure multimedia authentication. Future improvements with advanced architectures and larger datasets can further enhance the system's accuracy, scalability, and real-world applicability.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.