Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250402034
Volume 4, Issue2 (May-August 2025), PP: 246-252.
www.indjcst.com
ISSN No: 2583-5300

Deepfake Detection

Aakanksha Toutam[1], Sanika Gongale[2], Amey Zade[3], Kaushal Kamde[4], Vidish Worah[5], Manoj Chittawar[6]

[1,2,3,4,5]*Students, Department of Computer Science and Engineering, RCERT, Chandrapur, Maharashtra, India.*
[6]*Guide and Assistant Professor, Department of Computer Science and Engineering, RCERT, Chandrapur, Maharashtra, India.*

To Cite this Article: *Aakanksha Toutam[1], Sanika Gongale[2], Amey Zade[3], Kaushal Kamde[4], Vidish Worah[5], Manoj Chittawar[6], "Deepfake Detection", Indian Journal of Computer Science and Technology, Volume 04, Issue 02 (May-August 2025), PP: 246-252.*

Abstract: *The proliferation of deepfake technology, which utilizes advanced deep learning techniques to manipulate video and audio, poses significant threats to media integrity and information authenticity. Deepfakes are synthetic media created using neural networks, particularly Generative Adversarial Networks (GANs), to produce realistic but fake video content. While these technologies have legitimate applications in entertainment and digital media, they are increasingly exploited for malicious purposes such as spreading misinformation, impersonating individuals, and conducting fraud*

To address these challenges, this paper proposes a deepfake detection framework that leverages the combined power of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The CNN component is built upon the ResNeXt architecture, which enhances traditional CNNs by increasing the cardinality—or number of parallel pathways—within each layer. This design allows for richer and more diverse feature extraction from individual video frames. These spatial features, which capture the fine-grained details of facial structures, expressions, and inconsistencies, are then passed into an LSTM (Long Short-Term Memory) network.

The LSTM is a type of RNN designed to process sequences and maintain long-term dependencies, making it ideal for video analysis where temporal coherence is essential. It models the transitions and motion patterns across frames, learning cues such as irregular eye movements, mismatched lip-syncing, or abrupt changes in facial features—hallmarks of deepfake manipulation. This dual-stage architecture enables our system to simultaneously analyze spatial patterns within frames and dynamic behaviors across time

Overall, the combination of ResNeXt and LSTM networks allows for comprehensive detection of deepfakes by capturing both the static and dynamic features of video content. The architecture is not only accurate and interpretable but also adaptable for future improvements, such as incorporating attention mechanisms or multimodal data. This research contributes to the growing need for dependable tools in the fight against synthetic media and misinformation in the digital age.

Key Words: *Deepfake Detection, Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), ResNeXt.*

I.INTRODUCTION

Deepfake technology represents one of the most significant developments in artificial intelligence and multimedia content manipulation. At its core, a deepfake involves the use of deep learning techniques—particularly Generative Adversarial Networks (GANs)—to create or alter video, audio, and image content so convincingly that it is often indistinguishable from authentic media. The term "deepfake" is a combination of "deep learning" and "fake," highlighting the technology's foundation in neural network architectures.

The emergence of deepfakes has opened up new possibilities for innovation in fields like entertainment, film production, gaming, and education. For instance, deepfakes are used to generate realistic avatars, voiceovers, and historical reenactments. However, these same capabilities have also introduced serious threats. When misused, deepfakes can be deployed to spread disinformation, carry out identity fraud, manipulate public opinion, or even blackmail individuals. As such, the detection of deepfakes has become a critical area of research in computer vision and cybersecurity.

Detecting deepfakes is a complex task due to the increasing realism of generated content. Early detection methods often relied on visual artifacts or inconsistencies that are no longer as apparent with modern-generation deepfakes. These include blinking irregularities, head pose mismatches, or image quality inconsistencies. As generative models improve, these visual cues become harder to detect, necessitating more advanced and data-driven solutions.

To combat this issue, researchers have turned to deep learning-based detection models that can automatically learn intricate features from data. One popular approach is the use of Convolutional Neural Networks (CNNs), which are well-suited for capturing spatial features from images and video frames. CNNs can identify subtle inconsistencies in facial textures, lighting, and edges that may indicate manipulation. However, CNNs alone are limited in their ability to process temporal information—how features change and evolve over time.

To address this limitation, this research integrates CNNs with Long Short-Term Memory (LSTM) networks—a specialized form of Recurrent Neural Networks (RNNs). LSTMs are designed to analyze sequences, making them ideal for video data where temporal consistency is key. When combined, CNNs extract spatial features from each frame while LSTMs capture the temporal

dependencies between frames. This dual architecture significantly enhances the ability to detect deepfakes, especially in video formats where manipulations are often hidden across time rather than in individual frames.

This paper presents a deepfake detection framework that employs the ResNeXt CNN architecture and LSTM networks. ResNeXt introduces cardinality as an additional dimension in neural network design, improving performance and feature diversity. The integration of ResNeXt with LSTM offers a powerful mechanism for both spatial and temporal analysis, resulting in a highly effective detection system.

II. LITERATURE SURVEY

Title	Author(s)	Algorithm Used	Key Findings
Deepfake Detection: A Systematic Literature	Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, Andrew H. Sung	CNNs (ResNet, XceptionNet, VGG), RNNs (LSTM), SVM, Random Forest, Blockchain-based methods	Deep learning models dominate; CNN + RNN hybrids perform well; generalization across datasets is a major challenge; lack of standardized benchmarks.
Deep Fake Detection and Classification Using ELA and Deep Learning	Rimsha Rafique, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, Asma Hassan Alshehri	Error Level Analysis (ELA), CNNs (ResNet18, GoogLeNet, SqueezeNet), SVM, KNN	Hybrid ELA + CNN improves pixel-level manipulation detection; ResNet18 + KNN achieves 89.5% accuracy; suitable for image-based detection but computationally intensive.
Deepfake Detection Using Deep Learning methods: A systematic and comprehensive review	Arash Heidari, Nima Jafari Navimipour, Hasab Dag, Mehmat Unal	CNNs (VGGNet, ResNet), adversarial training with GAN-generated samples	Adversarial training enhances robustness; CNN-based models perform better with adversarially augmented datasets; emphasizes need for continuous model adaptation.
The Deep Fake Detection Challenge (DFDC) Dataset	Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, Cristian Canton Ferrer	Deep CNNs (EfficientNet, XceptionNet, ResNet, 3D CNNs); extensive data augmentation; ensemble models	Introduces largest public deepfake dataset (>100k fake videos); winning models used strong CNNs + augmentation + ensembles; generalization remains difficult at internet scale.
Multi-Attentional Deepfake Detection	Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, Nenghai Yu	Multi-attentional network; EfficientNet-b4; Bilinear Attention Pooling; Regional Independence Loss; Attention-Guided Data Augmentation (AGDA)	Treats detection as fine-grained classification; multi-attentional heads capture local artifacts; outperforms vanilla binary classifiers; strong cross-dataset generalization; somewhat sensitive to video compression.

III.METHODOLOGY

Dataset:

We use the Celeb-DF dataset hosted on Kaggle, which contains 4,536 videos comprising a mix of real and deepfake content generated using advanced generative models. Prior to training and evaluation, all videos undergo a standardized preprocessing and normalization pipeline to ensure consistency. The dataset is partitioned into training, validation, and testing sets in a 70:15:15 ratio, resulting in 3,176 training videos, 680 validation videos, and 680 test videos.

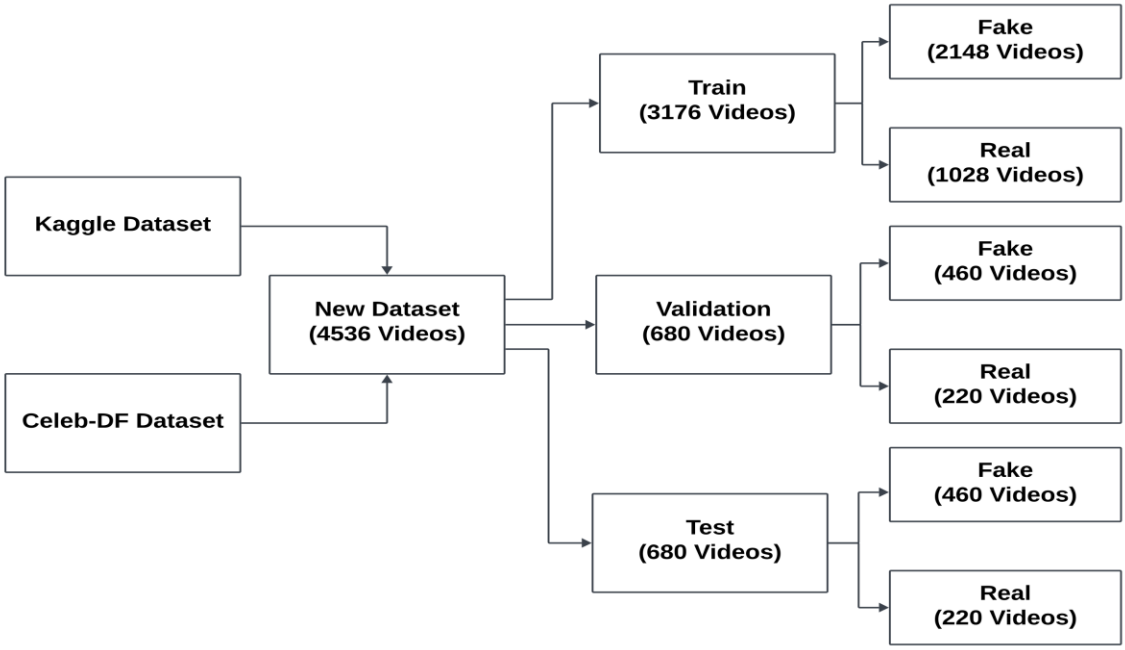

Fig.1: Dataset

IV.PREPROCESSING

The preprocessing workflow begins with loading the dataset and accessing associated labels. From each video, ten frames are uniformly extracted to capture key visual information while avoiding temporal redundancy. Each extracted frame is then converted to RGB format and resized to a standardized resolution. Next, we apply normalization transformations aligned with ImageNet standards, followed by stacking the processed frames into a fixed-size tensor sequence suitable for sequential modeling.

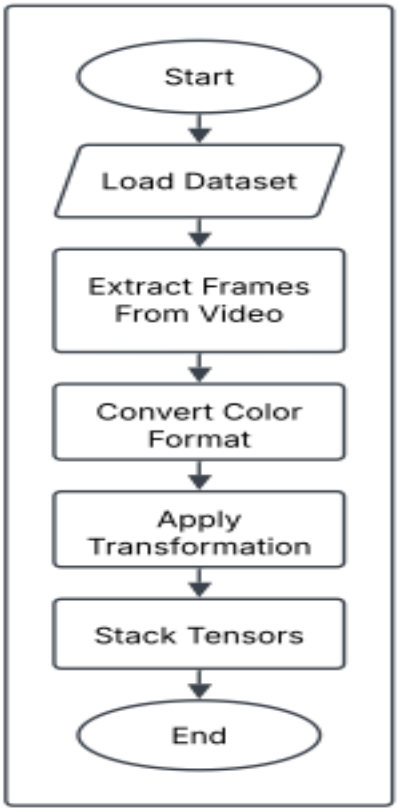

Fig. 2: Preprocessing

V. ARCHITECTURE

The core architecture of our model is divided into two parts: spatial feature extraction and temporal sequence classification. The feature extraction is handled by a ResNeXt CNN, which processes each frame individually and outputs a spatial embedding. These frame-level feature vectors are then passed into an LSTM network, which models the temporal progression of features across the sequence of frames. The LSTM's final output is fed into a dense layer with a SoftMax activation to classify the video as real or fake.

The architecture utilizes a ResNeXt50_32x4d backbone followed by an LSTM layer. The DataLoader handles the loading of preprocessed, face-cropped video clips and divides them into training and testing subsets. The extracted frames from these videos are then fed into the model in mini-batches for both training and evaluation.

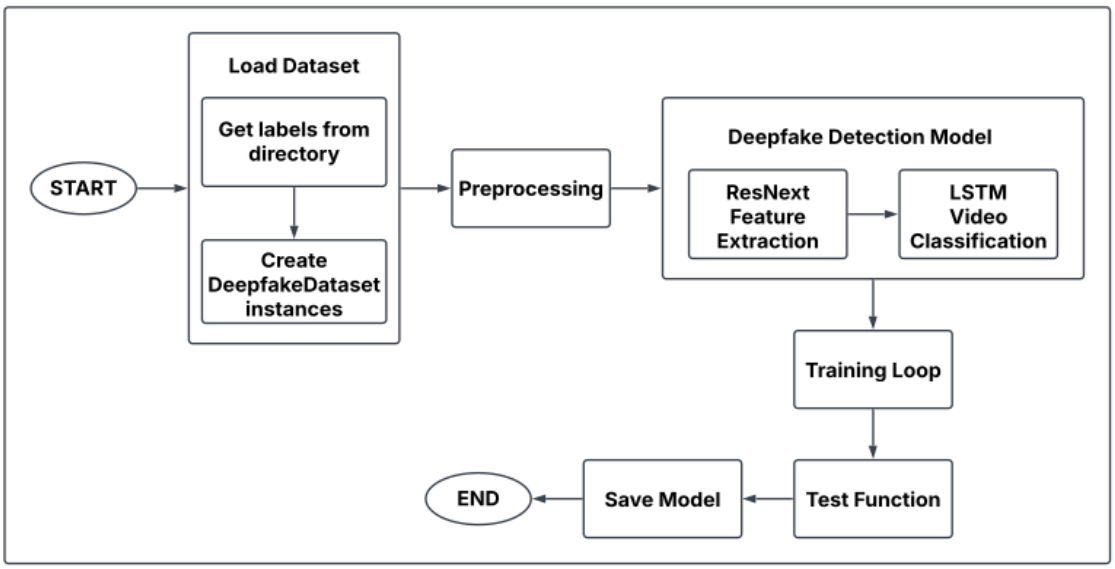

Fig. 3: Architecture

For training, we utilize a batch size of 4, a learning rate of 0.0001, and the Adam optimizer. The binary cross-entropy loss function guides the model's learning, and training is conducted over 5 epochs on an NVIDIA GeForce RTX 2060 GPU. Model performance is monitored using validation accuracy, and the best-performing model weights are saved for final evaluation.

ResNeXt CNN:

ResNeXt is a type of convolutional neural network (CNN) architecture introduced by Facebook AI Research (FAIR) in the 2017 paper **"Aggregated Residual Transformations for Deep Neural Networks"** by Xie et al. It builds upon ideas from **ResNet** and **Inception**, combining their strengths in a more efficient and scalable way. ResNeXt introduces the idea of **cardinality**, which is the number of parallel paths or "branches" within a block. Instead of increasing depth (more layers) or width (more channels), ResNeXt increases **cardinality** to improve performance. This is done through grouped convolutions.

For feature extraction, instead of building a classifier from scratch, we utilize the ResNeXt CNN model to extract rich, frame-level features with high accuracy. The network is then fine-tuned by appending additional layers as needed and carefully selecting a suitable learning rate to ensure effective convergence during gradient descent. The 2048-dimensional feature vectors obtained from the final pooling layer are subsequently used as input sequences for the LSTM module.

LSTM:

Long Short-Term Memory (LSTM) is an advanced form of Recurrent Neural Network (RNN), introduced by Hochreiter and Schmidhuber, specifically designed to handle long-range dependencies in sequential data. This makes LSTMs particularly well-suited for applications such as language translation, speech processing, and time series prediction.

In contrast to standard RNNs that rely solely on a single hidden state passed through each time step, LSTMs incorporate a dedicated memory cell. This cell is capable of retaining important information over longer durations, effectively addressing the limitations of traditional RNNs in learning from long sequences.

For sequence processing, consider a series of feature vectors extracted from video frames using the ResNeXt CNN as input. The goal is to classify the entire sequence using a two-node output layer that predicts whether the video is deepfake or authentic. A primary challenge lies in designing a model that can effectively interpret the sequence over time. To address this, we propose using an LSTM layer with 2048 units and a dropout rate of 0.4. This configuration allows the model to capture temporal dependencies by analyzing how the features evolve across frames. The LSTM processes the sequence in order, enabling comparison between the frame at time **t** and previous frames (e.g., **t–n**), where **n** can vary depending on the length of temporal context required.

Prediction:

When a new video is introduced for prediction, it undergoes preprocessing to match the format expected by the trained model. The video is first broken down into individual frames, and facial regions are cropped from each frame. Instead of saving these cropped frames to local storage, they are directly fed into the trained model for deepfake detection.

VI.RESULT

The deepfake detection model was trained using a dataset consisting of 3,176 training videos and 680 validation videos. The model architecture employed a CNN-based feature extractor (ResNeXt) followed by temporal sequence modeling using LSTM. The training was conducted on a GPU-enabled environment (NVIDIA GeForce RTX 2060) with a batch size of 4, learning rate of 0.0001, and 5 epochs.

During training, the model demonstrated rapid convergence with significant improvements in validation accuracy from the first epoch itself. Notably, at **epoch 1**, the model achieved a **validation accuracy of 99.85%**, and continued to maintain high validation accuracy above 96% throughout the training process. The final model was saved based on the best validation performance as shown in fig. 4.

Fig. 4: Training Result

To evaluate real-world applicability, two videos were tested using the trained model via a web interface:

• **Authentic Video (Figure 5)**: The model classified the uploaded video as **authentic**, with an accuracy of **99.43%**, showcasing its strong capability to correctly identify real content.

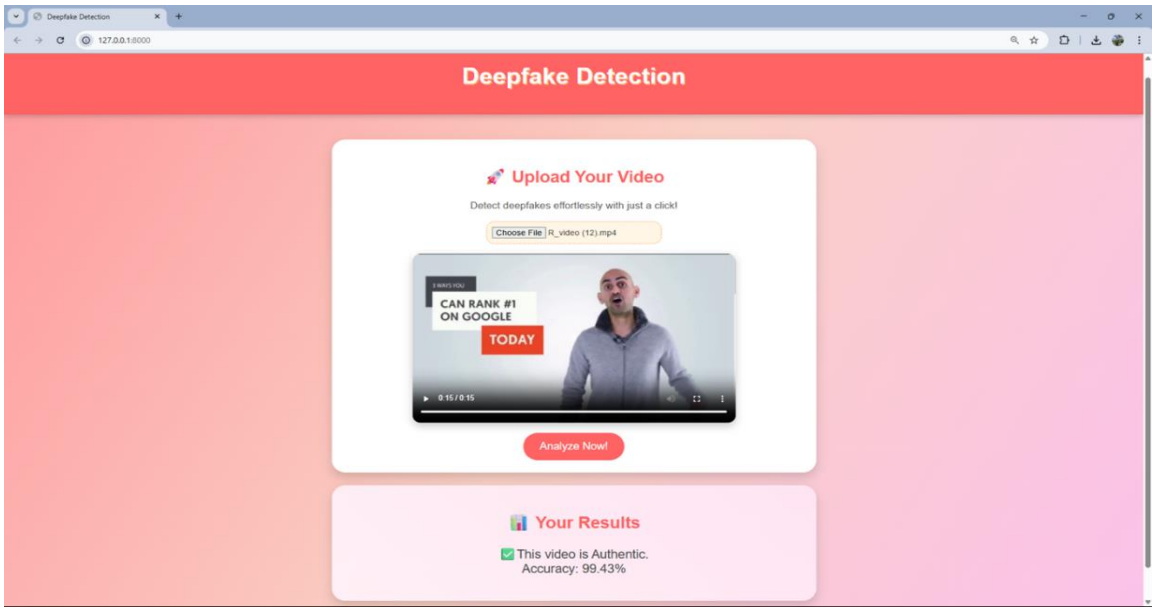

Fig 5: Output Real

- **Deepfake Video (Figure 6)**: The model accurately detected the second uploaded video as a **deepfake**, with an accuracy of **89.73%**, demonstrating effective discrimination of manipulated content.

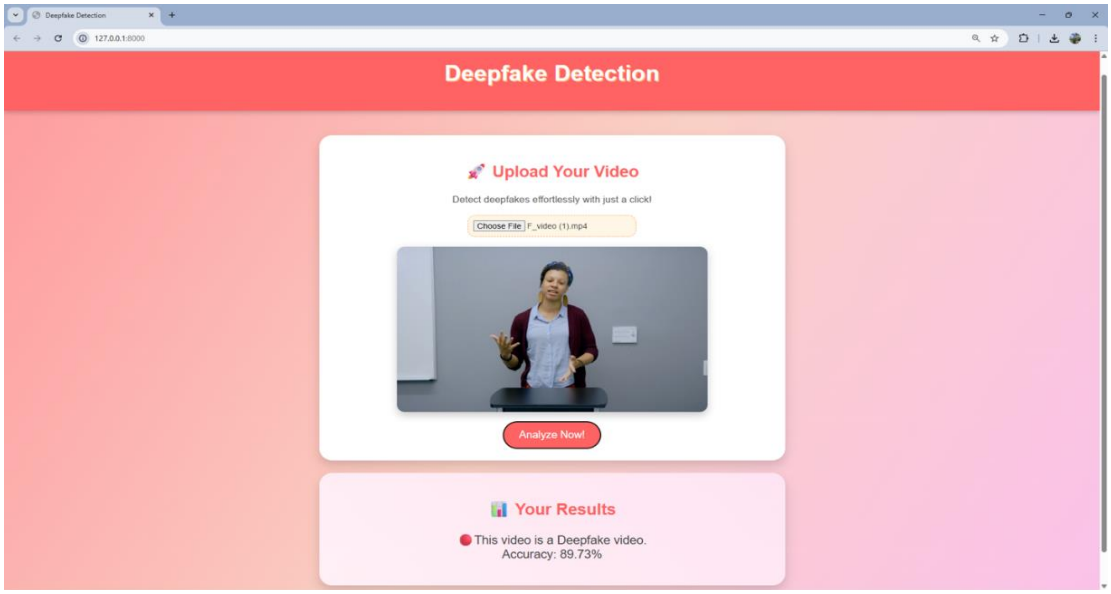

Fig 6: Output Deepfake

These results confirm the high reliability and robustness of the proposed deepfake detection framework in distinguishing between authentic and manipulated videos.

VII.CONCLUSION

This study presents a deepfake detection framework that integrates ResNeXt50_32x4d for spatial feature learning and a single-layer LSTM for temporal sequence modeling. Using the Celeb-DF dataset hosted on Kaggle, consisting of 4,536 videos, the model was trained with robust preprocessing that included frame extraction, face cropping, normalization, and tensor conversion. The model was trained over 5 epochs with a batch size of 4 and learning rate of 0.0001 using the Adam optimizer on an NVIDIA RTX 2060 GPU.

The hybrid architecture allowed the model to extract rich spatial features from individual frames while learning temporal patterns across sequences, resulting in a peak validation accuracy of 99.85%. This demonstrates the effectiveness of combining CNNs and LSTMs for detecting subtle manipulations in video data.

The results affirm that deepfake detection can be effectively addressed with spatial-temporal models, even on moderately powered hardware. The framework can be extended in future research by incorporating attention mechanisms or multi-modal inputs (e.g., audio) to improve robustness. Overall, this work provides a reliable and scalable method for detecting manipulated videos and contributes to ongoing efforts in media authentication and digital forensics.

This research presents a hybrid deep learning framework for deepfake video detection, combining ResNeXt50_32x4d for spatial feature extraction with an LSTM network to model temporal dynamics. The system was trained and evaluated on a carefully curated subset of the Celeb-DF dataset hosted on Kaggle, comprising 4,536 real and manipulated videos. Our preprocessing pipeline ensured consistency by extracting and normalizing ten representative frames per video, which allowed the model to focus on key visual regions, particularly the face.

The integration of ResNeXt and LSTM enabled the model to not only detect subtle inconsistencies within individual frames but also analyze their progression over time. The training setup, utilizing a batch size of 4, a learning rate of 0.0001, and the Adam optimizer across 5 epochs on an NVIDIA RTX 2060 GPU, resulted in a peak validation accuracy of 99.85%.

The high performance and reliability of the model demonstrate the viability of using spatial-temporal neural networks for deepfake detection, even on limited computational resources. This work lays the foundation for building scalable solutions that can support digital media forensics and online content authentication. Future directions could include incorporating attention mechanisms or audio-visual modalities to further enhance detection capability.

`Overall, the results affirm the strength of combining CNN and LSTM architectures for video-based forgery detection and underscore the importance of continuous innovation to stay ahead of increasingly sophisticated deepfake generation techniques.

References

1. *Heidari, A., Jafari Navimipour, N., Dag, H. and Unal, M., 2024. Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), p.e1520.*
2. *Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327. keywords: {Videos;Visualization;Image color analysis;Decoding;YouTube;Training;Detection algorithms},*
3. *Rana, M.S., Nobi, M.N., Murali, B. and Sung, A.H., 2022. Deepfake detection: A systematic literature review. IEEE access, 10, pp.25494-25513.*

4. *Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C., 2020. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.*
5. *Ahmed, S.R., Sonuç, E., Ahmed, M.R. and Duru, A.D., 2022, June. Analysis survey on deepfake detection and recognition with convolutional neural networks. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-7). IEEE.*
6. *Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W. and Yu, N., 2021. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194).*
7. *Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A. and Alshehri, A.H., 2023. Deep fake detection and classification using error-level analysis and deep learning. Scientific reports, 13(1), p.7422.*
8. *Guarnera, L., Giudice, O. and Battiato, S., 2020. Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 666-667).*
9. *Alkishri, W.A.S.I.N. and Al-Bahri, M.A.H.M.O.O.D., 2023. Deepfake image detection methods using discrete fourier transform analysis and convolutional neural network. Journal of Jilin University (Engineering and Technology Edition), 42(2).*
10. *Raza, A., Munir, K. and Almutairi, M., 2022. A novel deep learning approach for deepfake image detection. Applied Sciences, 12(19), p.9820.*
11. *Nirkin, Y., Wolf, L., Keller, Y. and Hassner, T., 2021. Deepfake detection based on discrepancies between faces and their context. IEEE transactions on pattern analysis and machine intelligence, 44(10), pp.6111-6121.*