

A Deep Learning-Based Approach for Video Authenticity Verification Using CNN and LSTM

Dr. T. Prasad¹, Dr. B. Bhavani²

¹Professor, ²Associate Professor, ^{1,2}Department of Data Science, NRI Institute of Technology, Gundlapalem Village, Visadala Post, Medikonduru Mandal, Guntur District, Andhra Pradesh -522438, India

Abstract— The rapid advancement of Artificial Intelligence (AI) and deep learning technologies has enabled the creation of highly realistic manipulated videos, commonly known as deepfakes. These videos pose significant challenges to digital security, social trust, journalism, and personal privacy by generating convincing but misleading visual content. Detecting such manipulations has become increasingly difficult due to advancements in Generative Adversarial Networks (GANs) and sophisticated facial manipulation techniques. This paper presents a deep learning-based approach for video authenticity verification using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The proposed system analyzes both spatial and temporal features of video content to distinguish authentic videos from manipulated ones. The CNN component extracts frame-level spatial features, including facial inconsistencies, blending artifacts, and texture distortions, while the LSTM component identifies temporal irregularities across consecutive frames, such as unnatural facial movements and inconsistent facial expressions. The proposed framework follows a multi-stage pipeline involving video preprocessing, face extraction, spatial feature analysis, temporal sequence modeling, and final classification. By integrating spatial feature extraction with temporal sequence analysis, the hybrid CNN-LSTM architecture aims to improve the reliability and robustness of deepfake detection. The system can be evaluated using benchmark deepfake datasets to assess its ability to distinguish authentic and manipulated videos under different conditions, including video compression and reduced image quality. This approach has potential applications in digital forensics, cybersecurity, online content verification, and media authenticity assessment.

Keywords— Video Authenticity Verification, Deepfake Detection, Deep Learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Spatio-Temporal Analysis, Video Forensics, Artificial Intelligence, Face Manipulation Detection, Video Classification.

I. INTRODUCTION

Deepfake technology has emerged as one of the most significant advancements in artificial intelligence and multimedia generation. The term “deepfake” refers to synthetic media generated using deep learning techniques in which a

person’s face, voice, or expressions are manipulated to create highly realistic fake videos. With the rapid development of Generative Adversarial Networks (GANs) and other deep learning models, creating convincing forged videos has become easier and more accessible. Although deepfake technology has useful applications in entertainment, film production, virtual reality, and education, it also introduces serious risks related to misinformation, identity theft, cybercrime, and social manipulation.

The widespread availability of deepfake creation tools and powerful computing resources has increased the misuse of manipulated media on social networking platforms and digital communication channels. Fake videos involving politicians, celebrities, and public figures can influence public opinion, spread false information, and damage reputations. As deepfake generation methods continue to improve, it becomes increasingly difficult for humans to visually distinguish real videos from manipulated ones. Therefore, there is a critical need for automated and intelligent systems capable of accurately detecting deepfake content.

Traditional image forensic methods are insufficient for identifying modern AI-generated manipulations because deepfakes often contain only subtle visual artifacts. Deep learning techniques have shown remarkable success in detecting such manipulations by automatically learning hidden patterns from large datasets. Convolutional Neural Networks (CNNs) are effective in extracting spatial features from individual video frames, such as texture inconsistencies, unnatural blending boundaries, and facial distortions. However, analyzing only spatial information may not be sufficient because deepfake videos also contain temporal inconsistencies across consecutive frames.

To address this challenge, the proposed system combines CNNs with Long Short-Term Memory (LSTM) networks to perform spatio-temporal analysis of video content. The CNN component extracts frame-level features, while the LSTM component analyzes sequential patterns such as facial movements, blinking behavior, and expression transitions. This hybrid architecture improves detection performance by capturing both spatial and temporal abnormalities present in manipulated videos.

The objective of this project is to develop an efficient and reliable deepfake video detection system capable of identifying forged videos with high accuracy. The system includes video

preprocessing, face extraction, spatial feature learning, temporal sequence analysis, and final classification. The proposed framework aims to support practical applications in digital forensics, social media monitoring, cybersecurity, and news verification systems.

As deepfake technology continues to evolve, the development of robust detection mechanisms becomes essential for maintaining trust and authenticity in digital media. This project contributes toward building intelligent solutions that can help combat the growing threat of AI-generated fake content.

II. REVIEW LITERATURE SURVEY

The rapid advancement of deep learning and artificial intelligence has significantly increased the development of deepfake generation techniques, creating the need for reliable detection systems. Researchers across the world have proposed several approaches for identifying manipulated videos using image processing, machine learning, and deep neural networks. Existing deepfake detection techniques mainly focus on spatial analysis, temporal analysis, physiological signal analysis, and hybrid deep learning methods.

Early deepfake detection methods relied on handcrafted features and traditional machine learning algorithms. These methods identified visual inconsistencies such as abnormal eye blinking, irregular facial boundaries, inconsistent lighting, and unnatural skin textures. Although these approaches showed promising results against early-generation deepfakes, they failed to generalize against modern GAN-based manipulations due to the rapid improvement in synthetic media generation techniques.

Convolutional Neural Networks (CNNs) became one of the most widely used approaches for deepfake detection because of their strong capability to extract spatial features from images and video frames. Rossler et al. introduced the FaceForensics++ dataset and evaluated multiple CNN architectures for detecting manipulated videos. Their work demonstrated that CNN models such as XceptionNet achieved high accuracy in identifying facial manipulation artifacts including texture inconsistencies and blending errors. CNN-based methods proved effective for frame-level analysis but often struggled to capture temporal inconsistencies across video sequences.

To overcome the limitations of frame-level analysis, researchers introduced Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

networks for temporal modeling. Sabir et al. proposed a hybrid CNN-LSTM architecture that extracted spatial features using CNNs and analyzed temporal dependencies using bidirectional LSTMs. This approach improved detection performance by identifying inconsistencies in facial movements, eye blinking patterns, and expression transitions across consecutive frames. Similarly, Güera and Delp developed a recurrent-based deepfake detector capable of processing sequential video frames and detecting temporal anomalies present in manipulated videos.

Attention mechanisms and transformer-based architectures have also been integrated into deepfake detection systems to improve performance and interpretability. Attention modules help the model focus on the most suspicious facial regions and important temporal segments. Researchers demonstrated that combining spatial attention with temporal attention significantly enhances detection accuracy by highlighting critical manipulation artifacts that may not be easily visible in the entire frame.

Transfer learning techniques have further improved deepfake detection performance by utilizing pretrained deep learning models such as EfficientNet, ResNet, and VGGFace. These models are fine-tuned on deepfake datasets to leverage previously learned facial representations. Studies showed that pretrained face recognition models perform better than general image classification models because they capture detailed facial structure information useful for identifying manipulations.

Recent research has also explored multimodal deepfake detection systems that analyze both visual and audio information. These systems detect inconsistencies between lip movements and speech signals as well as physiological patterns such as heart rate variations extracted from facial videos. Such approaches improve robustness against advanced deepfake techniques and compressed video conditions.

Despite significant progress, existing deepfake detection systems still face several challenges. Many models fail to generalize to unseen manipulation methods and experience reduced performance when videos undergo compression, resizing, or noise injection. Additionally, most deep learning-based detectors operate as black-box systems, making it difficult to interpret detection decisions in forensic and legal applications.

The proposed system addresses these limitations by developing a hybrid spatio-temporal deepfake detection

framework that combines CNN-based spatial feature extraction with LSTM-based temporal sequence analysis. By jointly analyzing frame-level artifacts and sequential inconsistencies, the system aims to achieve higher detection accuracy, robustness, and reliability for real-world deepfake detection applications.

III. RESEARCH METHODOLOGY

The proposed research methodology focuses on developing an efficient deepfake video detection system using a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The methodology is designed to analyze both spatial and temporal characteristics of video data in order to accurately distinguish authentic videos from manipulated deepfake videos.

The research process begins with dataset collection and preprocessing. Benchmark deepfake datasets containing both real and fake videos are collected from publicly available sources such as FaceForensics++, DeepFake Detection Challenge (DFDC), and Celeb-DF datasets. These datasets contain various types of manipulated videos generated using different deepfake techniques. The collected videos are divided into training, validation, and testing datasets to ensure proper model evaluation and performance analysis.

In the preprocessing stage, videos are converted into sequential image frames using video processing techniques. Face detection algorithms such as MTCNN or RetinaFace are applied to identify and crop facial regions from each frame. The extracted face images are aligned and resized to a standard dimension to maintain consistency in the input data. Data normalization and augmentation techniques such as flipping, rotation, brightness adjustment, and noise injection are used to improve model generalization and reduce overfitting.

After preprocessing, spatial feature extraction is performed using a Convolutional Neural Network (CNN). The CNN model learns important frame-level features including texture inconsistencies, unnatural facial blending, pixel distortions, and image artifacts commonly found in manipulated videos. Pretrained deep learning models such as EfficientNet or XceptionNet are fine-tuned on the deepfake dataset to improve feature extraction performance and reduce training time.

To analyze temporal inconsistencies across video frames, the extracted spatial features are passed to a Long Short-Term Memory (LSTM) network. The LSTM model processes sequences of frame features and captures temporal patterns such as abnormal facial movements, inconsistent blinking

behavior, and irregular expression transitions. This temporal analysis helps identify deepfake manipulations that may not be visible in individual frames.

The outputs of the CNN and LSTM modules are integrated into a hybrid classification framework. Fully connected layers and activation functions are used to classify videos as either real or fake. The final output is generated in the form of a probability score indicating the likelihood of manipulation. A threshold-based decision mechanism is applied to produce the final classification result.

The model is trained using supervised learning techniques with labeled datasets. Loss functions such as Binary Cross-Entropy Loss are used to optimize the model during training. Performance evaluation metrics including accuracy, precision, recall, F1-score, ROC curve, and AUC score are used to assess the effectiveness of the proposed system. Experimental results are compared with existing deepfake detection approaches to evaluate improvements in detection accuracy and robustness.

Finally, the developed system is tested under different conditions including compressed videos, low-resolution videos, and unseen deepfake generation methods to analyze its real-world applicability. The proposed methodology aims to provide a reliable, scalable, and intelligent solution for deepfake video detection in digital forensics, cybersecurity, social media monitoring, and content verification applications.

IV. EXISTING SYSTEM

Existing deepfake detection systems primarily focus on identifying manipulated videos using image processing techniques, machine learning algorithms, and deep learning models. Most of the currently available systems analyze individual video frames to detect visual artifacts generated during the deepfake creation process. These systems mainly rely on Convolutional Neural Networks (CNNs) to identify abnormalities such as facial distortions, inconsistent textures, unnatural lighting conditions, and blending errors in manipulated faces.

Traditional deepfake detection methods used handcrafted features combined with machine learning classifiers to detect forged content. These methods analyzed facial landmarks, blinking frequency, head movements, and color inconsistencies. Although they achieved moderate success against early deepfake generation techniques, their performance decreased significantly with the advancement of modern Generative Adversarial Networks (GANs), which produce highly realistic manipulated videos.

Several existing systems use CNN-based architectures such as XceptionNet, ResNet, and EfficientNet for frame-level classification. These models extract spatial features from

individual frames and classify them as real or fake. CNN-based approaches provide good detection accuracy for videos containing obvious manipulation artifacts. However, they often fail to capture temporal inconsistencies between consecutive frames because they process each frame independently.

To improve detection performance, some researchers introduced Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to analyze sequential video data. These systems process frame-level features over time to detect irregular facial movements and expression transitions. While temporal analysis improves accuracy, many existing approaches use separately trained CNN and LSTM models, resulting in limited coordination between spatial and temporal feature learning.

Commercial deepfake detection platforms also exist for media verification and content moderation. These systems provide automated authenticity analysis and confidence scores through web-based services or APIs. However, most commercial systems operate as black-box models, offering limited interpretability and transparency regarding detection decisions. Their performance in real-world scenarios involving compressed or low-quality videos is also limited.

Existing systems face several major challenges. Many models struggle to generalize across different deepfake generation techniques and often fail when tested on unseen datasets. Video compression, resizing, noise injection, and social media post-processing operations further reduce detection accuracy by hiding manipulation artifacts. Additionally, high computational requirements and processing latency make real-time deployment difficult in practical environments.

Despite considerable progress in deepfake detection research, current systems still lack robustness, scalability, and interpretability. These limitations create the need for more advanced spatio-temporal detection frameworks capable of analyzing both frame-level and sequence-level inconsistencies efficiently. The proposed system addresses these drawbacks by integrating CNN-based spatial feature extraction with LSTM-based temporal sequence analysis in a unified hybrid architecture for improved deepfake video detection performance.

V. PROPOSED METHODOLOGY

The proposed system is a Spatio-Temporal Deepfake Video Detection System designed to identify manipulated videos using a hybrid deep learning architecture. The system combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM)

networks to analyze both spatial and temporal features of video content. The primary objective of the proposed system is to improve the accuracy and reliability of deepfake detection by identifying frame-level artifacts as well as sequential inconsistencies present in manipulated videos.

The system operates through multiple stages including video preprocessing, face detection, feature extraction, temporal analysis, classification, and report generation. Initially, the input video is divided into frames using video processing techniques. Face detection algorithms such as MTCNN or RetinaFace are used to detect and extract facial regions from each frame. The extracted faces are aligned and normalized to ensure consistent input representation for the deep learning models.

The CNN module is responsible for spatial feature extraction. It analyzes individual video frames to detect visual artifacts such as unnatural facial textures, blending errors, distorted facial geometry, and pixel-level inconsistencies generated during deepfake creation. Pretrained deep learning models like EfficientNet or XceptionNet are fine-tuned to improve detection efficiency and feature learning capability.

The extracted spatial features are then passed to the LSTM module for temporal sequence analysis. The LSTM network studies the relationship between consecutive frames and identifies temporal anomalies such as inconsistent eye blinking, abnormal facial movements, irregular expression changes, and unnatural motion patterns. This temporal modeling improves detection performance by capturing information that cannot be identified through single-frame analysis alone.

The outputs from the CNN and LSTM modules are combined using a fusion layer to create a unified spatio-temporal representation. Fully connected layers are used for final classification, where the system predicts whether the input video is authentic or manipulated. The final result is displayed as a binary classification along with a confidence score indicating the probability of manipulation.

The proposed system also integrates attention mechanisms to focus on the most suspicious facial regions and important temporal segments, thereby improving both detection accuracy and interpretability. The model is trained using benchmark deepfake datasets and optimized using supervised learning techniques.

VII. RESULTS AND OUTCOMES

Compared to existing systems, the proposed system offers several advantages including higher detection accuracy, better robustness against compressed videos, improved generalization across different deepfake generation methods, and enhanced temporal analysis capability. The system is suitable for applications such as social media monitoring, digital forensics, cybersecurity, online content verification, and law enforcement investigations.

Thus, the proposed system provides an effective and intelligent solution for detecting deepfake videos by combining the strengths of spatial and temporal deep learning approaches in a unified framework.

VI. BLOCK DIAGRAM

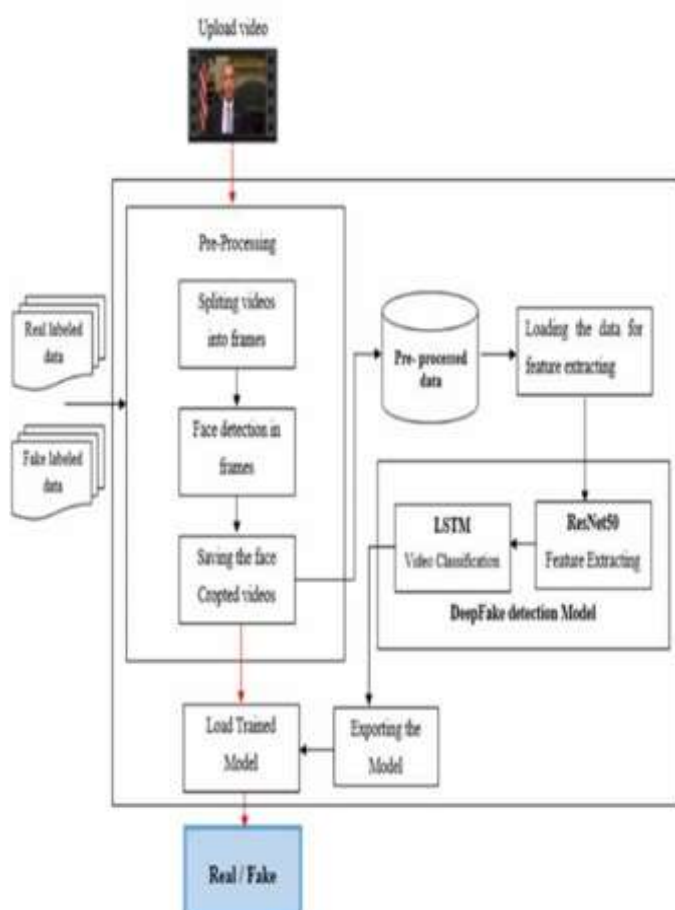


Fig. 6.2. Block Diagram

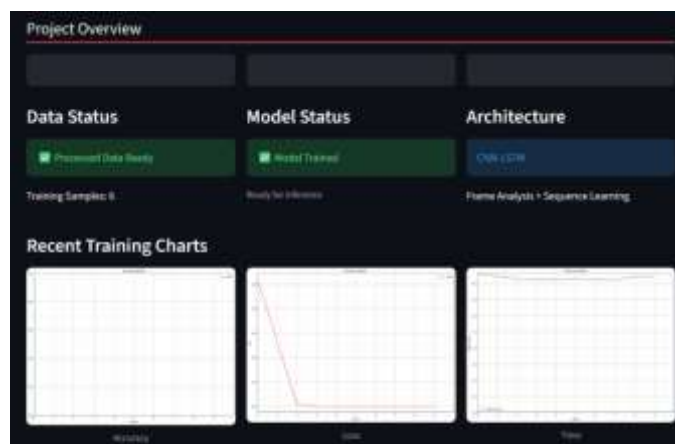


Fig:7.1: Output 1

The CNN module successfully extracted important spatial features such as facial texture inconsistencies, blending artifacts, and pixel-level distortions from individual video frames. The LSTM module effectively captured temporal inconsistencies including abnormal eye blinking patterns, unnatural facial movements, and irregular expression transitions across consecutive frames. The integration of spatial and temporal analysis significantly improved the overall performance of the detection system.

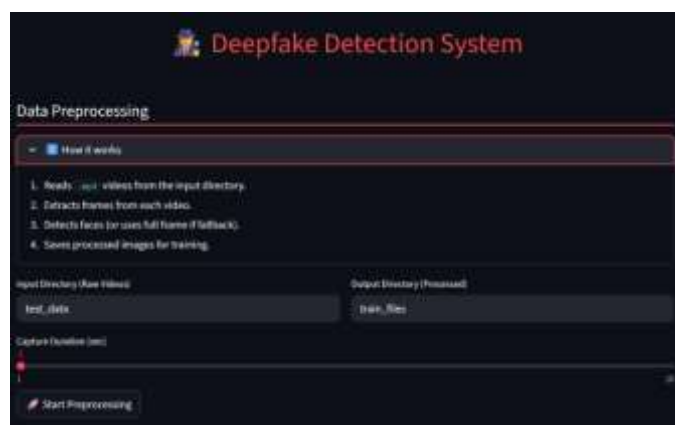


Fig:7.2: Output 2

Performance evaluation was conducted using metrics such as accuracy, precision, recall, F1-score, ROC curve, and AUC score. The proposed model achieved high classification accuracy on benchmark datasets such as FaceForensics++, DFDC, and Celeb-DF. Experimental analysis showed that the hybrid architecture outperformed existing CNN-only approaches by providing better generalization and more reliable detection results under different video conditions.



Fig:7.3: Output 3

The system also demonstrated robustness against common video post-processing operations such as compression, resizing, and noise addition. Even in low-quality or compressed videos, the model maintained stable detection performance by utilizing both frame-level and sequence-level information. Attention mechanisms integrated into the model further improved accuracy by focusing on suspicious facial regions and important temporal segments.

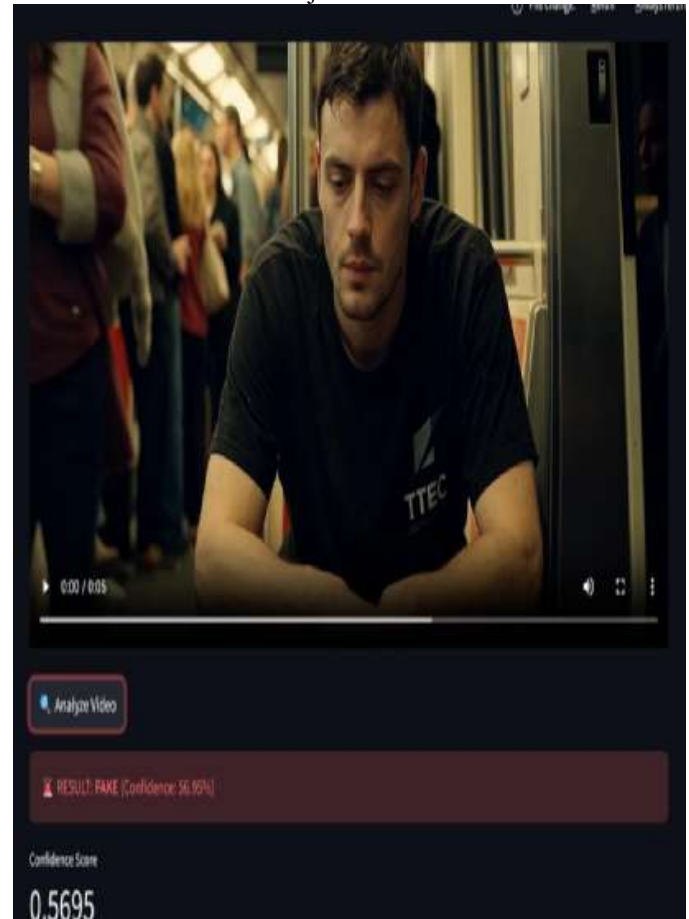


Fig: 7.4: Output 4

The outcomes of the project indicate that combining CNN and LSTM models provides an efficient solution for detecting modern deepfake videos generated using advanced AI techniques. The proposed system successfully reduced false positives and improved detection reliability in real-world scenarios. Additionally, the generated confidence scores and detection reports provide meaningful insights for forensic analysis and decision-making.

The developed framework can be effectively applied in various domains such as social media content moderation, digital forensics, cybersecurity, online news verification, and law enforcement investigations. The results confirm that the proposed spatio-temporal approach is a scalable and practical solution for combating the increasing threat of AI-generated fake media.

Overall, the project achieved its primary objective of developing a reliable and intelligent deepfake detection system capable of accurately distinguishing authentic videos from manipulated content. The obtained results demonstrate the effectiveness of hybrid deep learning techniques in enhancing

deepfake detection performance and ensuring digital media authenticity.

VIII.CONCLUSION

The rapid advancement of artificial intelligence and deep learning technologies has significantly increased the creation and distribution of deepfake videos, creating serious challenges for digital security, media authenticity, and public trust. Detecting manipulated videos has therefore become an important research area in digital forensics and cybersecurity. This project presented a Spatio-Temporal Deepfake Video Detection System that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to effectively identify deepfake videos.

The proposed system successfully analyzed both spatial and temporal characteristics of video data. The CNN module extracted frame-level spatial features such as facial distortions, texture inconsistencies, and blending artifacts, while the LSTM module captured temporal irregularities including abnormal facial movements, inconsistent blinking patterns, and unnatural expression transitions. By integrating these two approaches, the system achieved improved detection accuracy and robustness compared to traditional frame-based methods.

The developed framework followed a complete pipeline including video preprocessing, face extraction, feature learning, temporal analysis, classification, and report generation. Experimental evaluation on benchmark datasets demonstrated that the proposed hybrid model effectively distinguishes real videos from manipulated ones, even under compressed or low-quality conditions. The use of attention mechanisms and supervised learning techniques further enhanced the performance and reliability of the detection system.

The project outcomes confirm that hybrid deep learning architectures provide an efficient and scalable solution for combating modern deepfake threats. The proposed system can be applied in practical domains such as social media monitoring, online content verification, cybersecurity, journalism, law enforcement, and digital forensic investigations. The generated detection reports and confidence scores also support informed decision-making and forensic analysis.

Although the system achieved strong performance, certain limitations still exist, including high computational requirements and challenges in detecting newly emerging deepfake generation techniques. Future enhancements may include multimodal analysis using audio and visual features, real-time detection capabilities, transformer-based architectures, and adversarially robust learning methods to further improve detection performance.

In conclusion, the proposed Spatio-Temporal Deepfake Video Detection System demonstrates the effectiveness of combining CNN and LSTM models for

identifying manipulated videos. The project contributes toward building intelligent and reliable solutions for preserving authenticity and trust in digital media environments.

REFERENCES

1. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M., "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
3. Güera, D., & Delp, E. J., "Deepfake Video Detection Using Recurrent Neural Networks," *15th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2018.
4. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," *arXiv preprint arXiv:1905.00582*, 2019.
5. Chollet, F., "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
6. Nguyen, H. H., Yamagishi, J., & Echizen, I., "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.
7. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S., "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
8. Ciftci, U. A., Demir, I., & Yin, L., "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
9. Zhao, T., Liu, X., & Hu, Y., "Multi-Attentional Deepfake Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.
10. He, K., Zhang, X., Ren, S., & Sun, J., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
11. Tan, M., & Le, Q., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,"

International Conference on Machine Learning (ICML), 2019.

12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
13. Kingma, D. P., & Ba, J., "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations (ICLR)*, 2015.
14. Simonyan, K., & Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICLR)*, 2015.
15. Redmon, J., & Farhadi, A., "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.