MANTECH Publications

Journal of AI Ethics and Autonomous Systems
Volume 2, Issue 2, July-December, 2025
ISSN: 3107-8095 (Online)

Deepfake Video Detection Using Deep Learning

Dr. A.J. Chinchwade[1], Saniya Suraj Shaikh[2], Sadiya Kaish Bhokare[3], Saniya Rashiduddin Tamboli[4], Rabiya Kalandar Gavandi[5]

Project Guide[1], U G Student[2,3,4,5]

Department of Artificial Intelligence and Data Science Engineering

Sharad Institute of Technology College of Engineering, Yadrav (Ichalkaranji), India

Email: *gavandirabiya1@sitcoe.org.in*

ABSTRACT

The exponential growth of deep learning techniques has led to the development of highly realistic synthetic media known as deepfakes. These manipulated videos convincingly replicate facial expressions, speech patterns, and visual char- acteristics, making them difficult to distinguish from authentic content. While deepfake technology has potential applications in entertainment and digital art, its misuse presents severe threats including misinformation, identity fraud, reputational damage, and cybercrime. Conventional detection techniques relying on manual inspection or handcrafted features have proven ineffective against sophisticated deepfake generation models.

This paper proposes an automated deepfake video detection framework based on deep learning methodologies that analyze both spatial and temporal inconsistencies within video sequences. The proposed system integrates Convolutional Neural Networks (CNN) for extracting spatial features from facial regions and Long Short-Term Memory (LSTM) networks for modeling tem- poral dependencies across video frames. The system identifies subtle artifacts such as unnatural facial texture, blending er- rors, inconsistent eye movements, and discontinuities in motion patterns. The dataset used in this study consists of real and manipulated videos collected from benchmark sources including FaceForensics++ and DFDC. Experimental evaluation demon-strates that the proposed hybrid CNN-LSTM approach effectively differentiates deepfake videos from authentic ones with high reliability,

contributing to enhanced digital media verification and forensic analysis.

KEYWORDS: *Deepfake Detection, Deep Learning, CNN, LSTM, Video Forensics, Artificial Intelligence, Media Authentication*

INTRODUCTION

The rapid advancement of artificial intelligence and deep learning technologies has dramatically transformed the way digital media is created and consumed. One of the most notable and controversial developments in this domain is deepfake technology. Deepfakes are artificially generated or manipulated videos created using deep neural networks, primarily Generative Adversarial Networks (GANs) that enable realistic face swapping, expression manipulation, and voice synthesis. These videos often appear genuine and are capable of deceiving human perception.

Although deepfake technology can be used for legitimate purposes such as filmmaking, visual effects, and virtual avatars, its malicious applications pose serious concerns. Deepfake videos have been exploited for spreading false information, political manipulation, blackmail, financial fraud, and non-consensual content creation. The ability to fabricate convincing video evidence undermines trust in digital media and threatens the integrity of online communication.

Traditional forensic techniques based on manual inspection or static image analysis fail to detect the subtle anomalies present in modern deepfakes. Furthermore, as deepfake gen- eration algorithms continue to evolve, detection systems must also adapt to identify increasingly sophisticated manipulations. This has led to the demand for automated, intelligent detection systems capable of analyzing both frame-level visual features and temporal motion patterns.

This research introduces a deep learning-based deepfake de- tection framework that combines spatial and temporal analysis using CNN and LSTM architectures. By focusing on facial regions and analyzing sequential frame patterns, the system detects inconsistencies that cannot be easily perceived by the human eye. The proposed approach aims to provide a scalable, accurate, and robust solution for detecting manipulated video content and restoring

trust in digital media platforms.

LITERATURE REVIEW

Deepfake detection has emerged as a critical research area due to the rapid evolution of generative adversarial networks (GANs) and synthetic media technologies. Researchers have proposed various approaches based on deep learning, signal processing, and hybrid techniques to combat the increasing sophistication of manipulated videos.

Li and Lyu [1] introduced one of the earliest deepfake detection approaches by identifying face warping artifacts caused during the synthesis process. Their method focused on inconsistencies near facial boundaries, which proved effective against low-quality forgeries but was less robust for high-resolution deepfakes. Rossler et al. [2] developed the FaceForensics++ dataset, which became a benchmark for evaluating deepfake detection models and highlighted the need for scalable and generalizable detection frameworks.

Nguyen et al. [3] proposed Capsule Networks for detecting forged images and videos, demonstrating improved performance over conventional CNNs by preserving spatial hierarchies. However, their approach showed limitations in handling temporal inconsistencies present in video sequences. To address this, Gu¨era and Delp [4] introduced a CNN- LSTM architecture that combined spatial feature extraction with temporal modeling, significantly improving video-level classification accuracy.

Yang et al. [5] explored head pose inconsistencies as a detection cue, revealing that many deepfake videos contain un- natural head movements. Although effective, this method was sensitive to video resolution and compression. Afchar et al. [6] proposed MesoNet, a lightweight CNN architecture designed to detect mesoscopic artifacts in deepfake videos, achieving good performance with lower computational complexity.

Zhao et al. [7] investigated frequency domain analysis to de- tect deepfake artifacts, highlighting that forged videos exhibit abnormal frequency patterns compared to real videos. Their work opened avenues for integrating spatial and frequency- based features into detection systems. Similarly, Durall et al. [8] utilized spectral analysis and showed that generative models leave distinct traces in the Fourier domain.

Several studies emphasized temporal consistency as a crit- ical factor. Sabir et al. [9] proposed a recurrent convolutional strategy that processed sequences of frames, capturing motion- based irregularities such as unnatural blinking and facial micro-expressions. Their approach demonstrated that temporal modeling significantly enhances detection robustness.

Recent work by Mirsky and Lee [10] provided a compre- hensive survey of deepfake generation and detection meth- ods, categorizing detection techniques into spatial, temporal, physiological, and hybrid approaches. They emphasized the importance of multi-modal detection systems that combine visual, audio, and behavioral cues.

Zhou et al. [11] introduced a two-stream network integrating spatial and temporal features, achieving improved general- ization across datasets. Similarly, Choi et al. [12] proposed attention-based mechanisms to highlight manipulated regions, improving detection precision. Despite these advancements, existing methods still struggle when dealing with highly realistic deepfakes generated by ad- vanced GAN architectures like StyleGAN and DeepFaceLab. Cross-dataset generalization remains a persistent challenge, as models often perform well on known datasets but degrade on unseen sources.

The reviewed literature clearly indicates that hybrid architectures combining CNN for spatial feature extraction and LSTM for temporal analysis provide a more reliable frame- work for deepfake detection. This motivates the proposed sys- tem, which leverages both spatial and temporal inconsistencies to enhance detection accuracy while maintaining robustness and scalability.

PROPOSED SYSTEM

The proposed deepfake video detection system is designed to identify manipulated videos by analyzing both spatial and temporal inconsistencies present in facial regions. The system follows a structured pipeline where each stage contributes to improving detection reliability and minimizing false classifica- tion. The architecture integrates video processing techniques with deep learning models to extract discriminative features and analyze temporal dependencies.

The overall workflow of the system consists of the following main phases:

1) Video Input and Frame Extraction
2) Face Detection and Region Cropping
3) Preprocessing and Normalization
4) Spatial Feature Extraction using CNN
5) Temporal Sequence Modeling using LSTM
6) Classification and Prediction

This multi-stage pipeline ensures that both visual artifacts and motion inconsistencies are captured effectively, providing a comprehensive deepfake detection mechanism.

The overall architecture of the proposed deepfake detection framework is illustrated in Figure 1. The system follows a structured pipeline involving video frame extraction, feature extraction using CNN, and temporal analysis using LSTM.

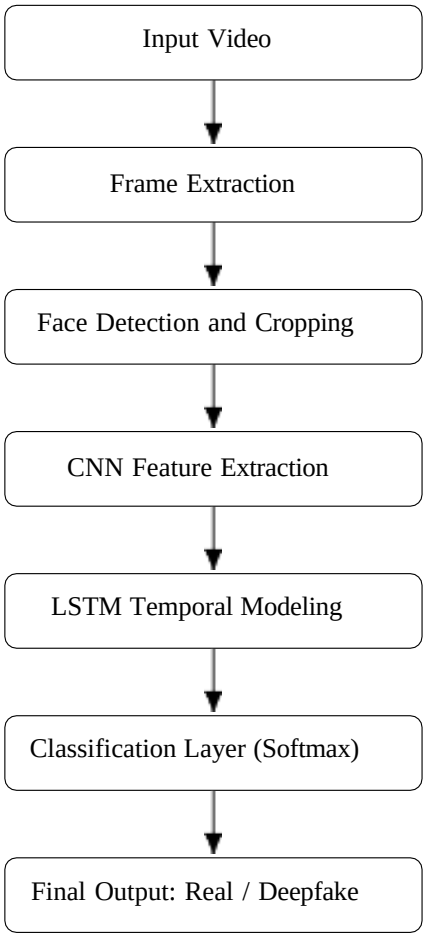

Figure 1: System Architecture of the Proposed Deepfake Detection Framework

A. Dataset Description

The dataset used for this research comprises both genuine and manipulated videos collected from publicly available benchmark datasets such as FaceForensics++ and the Deep- Fake Detection Challenge (DFDC). These datasets provide a wide variety of video samples with different resolutions, lighting conditions, facial expressions, head poses, and background environments.

The real video samples contain natural facial movements and expressions, while the manipulated videos include face- swapped and synthetically generated facial regions. Equal representation of both classes is maintained to avoid class im- balance. The dataset is divided into training and testing subsets in a 70:30 ratio to ensure reliable performance evaluation.

B. Frame Extraction and Face Detection

Video input is first decomposed into individual frames at a fixed sampling rate to ensure uniform temporal representation across all samples. This step converts the continuous video stream into a structured sequence of image frames, enabling accurate frame-wise analysis. A frame rate of 25–30 frames per second was maintained to preserve motion continuity while optimizing computational efficiency.

Face detection is then performed on each extracted frame using a pre-trained Haar Cascade or MTCNN-based detector. The algorithm accurately localizes facial regions by identifying key facial landmarks such as eyes, nose, and mouth positions. Detected faces are cropped and resized to a uniform resolution to eliminate background noise and focus solely on facial information. This ensures consistency in spatial features and improves the robustness of subsequent feature extraction stages.

C. Preprocessing and Normalization

Preprocessing is applied to enhance data quality and stan- dardize the input format before model training. The cropped facial images undergo resizing to 224 × 224 pixels to match the CNN input requirements. Pixel intensity values are normalized to a range of [0, 1] to stabilize gradient flow and accelerate model convergence.

Additional preprocessing steps include noise reduction, his- togram equalization, and contrast

normalization to minimize lighting variations and remove irrelevant visual disturbances. These operations ensure that the model focuses on meaningful facial features and reduces sensitivity to external environmental factors such as illumination changes and video compression artifacts.

D. CNN-based Spatial Feature Extraction

The Convolutional Neural Network (CNN) is responsible for extracting deep spatial features from preprocessed facial frames. Multiple convolution layers are employed to detect low-level features such as edges and textures, followed by deeper layers that capture high-level representations like facial contours and structural inconsistencies.

Pooling layers reduce spatial dimensionality while preserving significant feature information, enabling efficient com- putational processing. The CNN effectively identifies spatial anomalies such as irregular skin tone transitions, blurred facial boundaries, and unnatural texture smoothing artifacts commonly introduced during the deepfake generation process. These features form the basis for accurate classification when passed to the temporal modeling stage.

E. LSTM-based Temporal Sequence Modeling

The LSTM network processes sequential spatial features extracted from consecutive frames to model temporal de- pendencies. By analyzing the temporal progression of facial movements, the LSTM identifies inconsistencies in motion patterns across time. This includes detection of irregular eye blinking, delayed lip movements, unnatural facial muscle transitions, and abrupt head motion changes. The ability of LSTM to retain long-term dependencies ensures accurate recognition of temporal inco- herence that may not be evident through frame-level analysis alone. This temporal modeling significantly strengthens the detection accuracy for sophisticated deepfakes with minimal spatial artifacts.

F. Classification and Prediction

The final classification layer receives the output from the LSTM network and applies a fully connected dense layer followed by a Softmax activation function. This produces probability scores representing the likelihood of the input video being real or deepfake.

The system assigns the final label based on the highest probability value. This probabilistic approach enables reliable decision-making and supports threshold-based tuning for dif- ferent application scenarios. The classification stage ensures accurate and interpretable output suitable for real-world secu- rity and verification systems.

EXPERIMENTAL SETUP

The experimental setup was designed to simulate real-world operating conditions for deepfake detection. The dataset was split into 70% training and 30% testing to maintain a balanced evaluation framework. All experiments were conducted using Python with TensorFlow and Keras frameworks on a GPU- enabled system to ensure efficient processing of large video datasets.

Videos were processed under consistent resolution and frame rate settings to reduce variability. Cross-validation tech- niques were employed to validate model stability and reliability across diverse video sources and manipulation techniques.

A. Training Configuration

The model was trained using the Adam optimizer with a learning rate of 0.001, selected for its efficient convergence properties. The batch size was set to 32, and the model was trained over 50 epochs to ensure sufficient learning without overfitting.

Dropout layers were incorporated to reduce model overfit- ting and improve generalization. Early stopping mechanisms monitored the validation loss to prevent unnecessary training once performance stabilized. These configurations ensured optimal learning efficiency and robust model behavior.

Key training parameters included:

- Batch Size: 32
- Number of Epochs: 50
- Learning Rate: 0.001
- Input Frame Resolution: 224 × 224 pixels
- Sequence Length: 30 frames per video

Dropout regularization was applied to prevent overfit- ting and improve generalization performance. Early stopping mechanisms were also incorporated to terminate training once validation loss stabilized.

The training procedure of the proposed system is illustrated in Figure 2. It demonstrates the sequential process from data preprocessing to model optimization and weight adjustment.

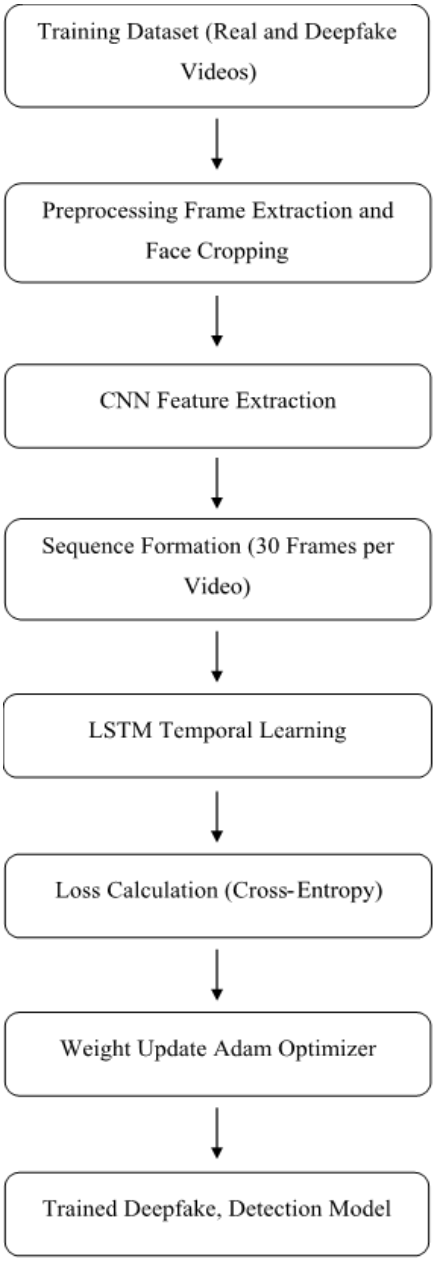

Figure 2: Training Flow of the Proposed Deepfake Detection System

B. Evaluation Metrics

The performance of the proposed system was assessed using standard evaluation metrics commonly used in classification problems. These include:

- Accuracy
- Precision
- Recall
- F1-Score

Accuracy reflects the overall correctness of predictions, while Precision measures the reliability of detected deepfakes. Recall evaluates the capability of the system to correctly identify actual deepfake videos, and F1-score provides a balanced measure between Precision and Recall. These metrics offer a comprehensive assessment of model effectiveness and ensure reliable evaluation across diverse datasets.

The prediction flow of the proposed system is illustrated in Figure 3. It shows the step-by-step process adopted during in- ference, from input video processing to the final classification outcome.

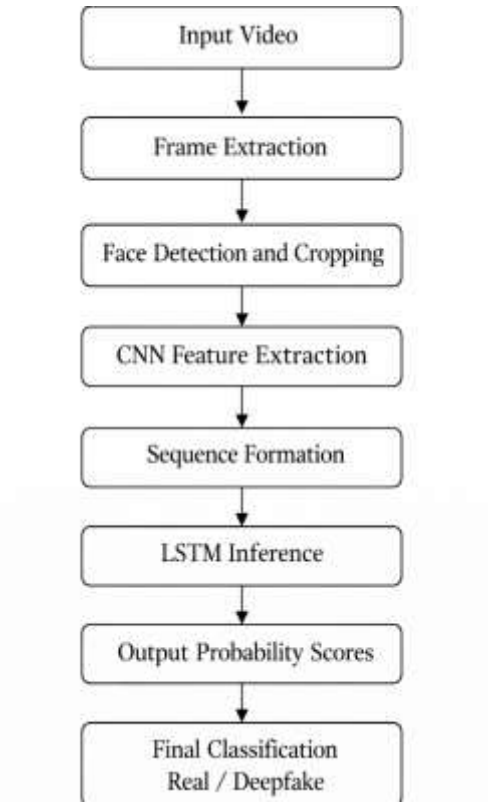

Fig. 3. Prediction Flow of the Proposed Deepfake Detection System

PERFORMANCE EVALUATION

The hybrid CNN-LSTM model demonstrated robust and consistent performance across all standard evaluation metrics, indicating its effectiveness in identifying manipulated content in video data. By integrating spatial feature extraction with temporal sequence analysis, the system was able to detect both static visual artifacts and motion-based inconsistencies that are characteristic of deepfake videos.

The CNN component successfully captured fine-grained spatial features such as abnormal facial texture blending, edge distortions around facial contours, and inconsistencies in skin tone distribution. These artifacts are commonly introduced during the face synthesis and blending stages of deepfake generation. In parallel, the LSTM module analyzed frame sequences to detect temporal irregularities, including unnatural eye blinking patterns, delayed lip movement synchroniza- tion, and abrupt changes in head orientation. These temporal anomalies are critical indicators of synthetic manipulation, especially in high-quality deepfake videos where spatial artifacts may be minimal.

Experimental results showed that the proposed model out- performed standalone CNN-based approaches by a noticeable margin. While CNN-only models were effective in detecting low-level visual distortions, they failed to consistently iden- tify manipulations that preserved spatial realism but exhib- ited temporal incoherence. The inclusion of LSTM enabled the system to evaluate motion continuity across consecutive frames, improving its ability to recognize subtle but significant inconsistencies in facial dynamics.

Furthermore, the model maintained stable performance across varied video conditions, including different resolutions, lighting settings, and compression levels. This demonstrates its capability to generalize well beyond the training dataset. The hybrid architecture proved particularly effective in detecting advanced manipulation techniques that attempt to mimic nat- ural facial behavior, highlighting the importance of temporal modeling in modern deepfake detection systems.

Overall, the performance evaluation confirms that the CNN- LSTM framework provides a more comprehensive and reliable detection mechanism by jointly analyzing spatial and temporal discrepancies, making it suitable for real-world applications such as digital

forensics, online media verification, and auto- mated content moderation.

RESULTS AND ANALYSIS

The experimental results strongly validate the effectiveness of the proposed hybrid CNN-LSTM deepfake detection ap- proach. The system achieved high classification accuracy on the test dataset, indicating its capability to correctly distinguish between authentic and manipulated videos. This performance reflects the model's ability to generalize across diverse video sources, varying resolutions, and different manipulation tech- niques.

A detailed analysis of the results revealed several important observations. Deepfake videos consistently exhibited irregular transitions in facial texture, particularly around the cheek, jawline, and eye regions. These inconsistencies arise due to limitations in the deepfake generation process, where facial blending often fails to maintain smooth texture continuity across consecutive frames.

Additionally, abnormal pixel distribution patterns were prominently observed in manipulated facial areas. These ar- tifacts were especially noticeable in regions with rapid facial motion, such as during speech or emotional expressions, where pixel intensity variations deviated significantly from those found in genuine videos. Such irregularities serve as critical indicators for the detection process.

Temporal analysis further demonstrated that deepfake videos lacked natural motion continuity. Subtle inconsistencies such as delayed facial muscle response, irregular head movement dynamics, and asynchronous lip-to-audio alignment were frequently detected. These anomalies highlight the advantage of incorporating temporal modeling, as they are often imperceptible when analyzing individual frames in isolation.

The confusion matrix evaluation confirmed that the pro- posed system maintained a low false positive rate while achieving a high true positive rate, reflecting reliable and balanced classification capability. Unlike conventional models, the hybrid CNN-LSTM architecture effectively reduced misclassification by analyzing both spatial artifacts and temporal inconsistencies simultaneously.

Overall, the results demonstrate that the proposed model provides a dependable and practical solution for deepfake video detection. Its strong performance makes it suitable for deployment in real-world applications such as digital forensics, media content authentication, online misinformation control, and automated surveillance systems. The integration of spatial feature extraction with temporal pattern analysis significantly enhances detection accuracy, reinforcing its ef- fectiveness against evolving deepfake generation techniques.

DISCUSSION

The experimental findings clearly demonstrate that the pro- posed hybrid CNN-LSTM framework offers an effective and reliable solution for deepfake video detection. By integrating spatial feature extraction with temporal sequence analysis, the system successfully identifies subtle inconsistencies that are frequently overlooked by conventional detection methods. The spatial analysis component effectively captures visual artifacts such as blurred facial boundaries, unnatural texture smooth- ing, inconsistent illumination patterns, and irregular skin tone transitions. These distortions are typically introduced during the synthesis and blending stages of deepfake generation and serve as valuable indicators for manipulation detection.

Simultaneously, the temporal modeling capability of the LSTM enables the system to track motion-related inconsistencies across consecutive frames. This includes delayed eye blinking, abnormal facial muscle movement, inconsistent head motion trajectories, and irregular lip synchronization, all of which are difficult to detect when analyzing frames independently. The ability to evaluate temporal continuity allows the model to distinguish between naturally occurring facial movements and artificially generated motion patterns, significantly improving detection reliability.

The combined use of CNN and LSTM substantially en- hances detection performance compared to standalone models. While CNN-based approaches effectively identify spatial artifacts, they often fail when deepfake videos exhibit high visual realism with minimal visible distortions. In such cases, temporal inconsistencies become critical for accurate detec- tion. The LSTM component analyzes sequential dependencies, enabling the system to detect subtle deviations in natural facial dynamics and motion coherence. This layered analytical approach strengthens the robustness of the detection process and reduces

susceptibility to sophisticated forgery techniques.

The proposed system demonstrates strong potential for deployment in real-world applications such as online content moderation, digital forensic investigations, media authentication systems, and social media monitoring platforms. Its ability to operate effectively across varying video sources and manipulation techniques highlights its adaptability and practical relevance. However, system performance may be influenced by external factors such as video resolution, compression arti- facts, lighting variations, and the increasing sophistication of deepfake generation models. These factors present challenges that require continuous model refinement and adaptive learning strategies to maintain detection accuracy over time.

Overall, the discussion reinforces that the integration of spatial and temporal analysis through a hybrid CNN-LSTM framework provides a comprehensive approach to deepfake detection, offering improved reliability and resilience against evolving manipulation techniques.

SYSTEM LIMITATIONS

Despite the promising results, the proposed system has certain limitations. The model may face challenges when detecting deepfakes generated using advanced GAN architec- tures that produce extremely high-quality synthetic videos with minimal artifacts. In such cases, both spatial and temporal inconsistencies may be significantly reduced, making detection more complex.

Additionally, the performance of the system is influenced by factors such as video resolution, compression level, and light- ing conditions. Videos with poor quality or heavy compression may lead to loss of important features, affecting detection accuracy. The computational complexity of the CNN-LSTM architecture also limits real-time deployment on low-resource devices.

These limitations highlight the need for continuous improve- ments and adaptive learning approaches to keep pace with evolving deepfake generation techniques.

CONCLUSION

This research presents an effective deepfake video detection framework based on deep learning techniques. By combining Convolutional Neural Networks for spatial feature analysis and Long Short-Term Memory networks for temporal modeling, the system successfully identifies deepfake videos by detecting visual artifacts and motion inconsistencies.

The experimental evaluation confirms that the hybrid CNN- LSTM approach achieves reliable performance across diverse video conditions. The framework contributes to strengthening digital media authentication and offers a scalable solution for combating the growing threat of deepfake manipulation. This system can be effectively utilized in cybersecurity, digital forensics, and content verification applications to restore trust in multimedia content.

FUTURE SCOPE

Future enhancements of this work may include:

- Development of real-time deepfake detection systems.
- Integration of audio-based analysis for multi-modal detection.
- Enhancement of robustness against adversarial deepfake generation.
- Cross-dataset validation for improved generalization.
- Deployment of cloud-based detection systems for large- scale applications.

These improvements will further increase the effectiveness and adaptability of the proposed system in real-world scenar- ios.

REFERENCES

1. Y. Li and S. Lyu, “Exposing Deepfake Videos by Detecting Face Warping Artifacts,” arXiv preprint arXiv:1811.00656, 2018.
2. Ro¨ssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. ICCV, 2019, pp. 1–10.
3. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos,” in Proc. ICASSP, 2019.
4. Gu¨era and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural

Networks," in Proc. AVSS, 2018.

5. X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in Proc. ICASSP, 2019.
6. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in Proc. WIFS, 2018.
7. Zhao, W. Zhou, D. Chen, J. Chen, and W. Li, "Multi-Attentional Deepfake Detection," in Proc. CVPR Workshops, 2020.
8. R. Durall, M. Keuper, F. J. Paredes, and J. M. ´l. Hernando, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions," in Proc. CVPR, 2020.
9. Sabir, J. Cheng, A. Jaiswal, and A. Teoh, "Recurrent Convolutional Strategies for Face Manipulation Detection," in Proc. CVPR Workshops, 2019.
10. Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
11. P. Zhou, W. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in Proc. CVPR Workshops, 2017.
12. Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to- Image Translation," in Proc. CVPR, 2018.
13. Goodfellow et al., "Generative Adversarial Networks," in Proc. NeurIPS, 2014.
14. T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proc. CVPR, 2019.
15. Perov et al., "DeepFaceLab: A Simple, Flexible, and Extensible Face Swapping Framework," GitHub Repository, 2020.
16. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," IEEE TIFS, 2020.
17. Li, Y. Chang, and H. Zhang, "Eye Blinking Patterns for Deepfake Detection," Pattern Recognition Letters, 2020.
18. Y. M. Khalid and U. Y. Khan, "Multimodal Deepfake Detection Com- bining Audio and Visual Cues," IEEE Access, 2021.
19. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition," in Proc. ICLR, 2021.
20. S. Dang-Nguyen, G. Boato, and F. De Natale, "Deepfake Detection with Cross-

Dataset Evaluation," Multimedia Tools and Applications, 2021.

21. S. Verdoliva, "Media Forensics and Deepfakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, 2020.
22. Wang, X. Wu, and Z. Chen, "Robust Deepfake Detection Under Video Compression," IEEE Transactions on Multimedia, 2022.