

A Review: Unveiling the Unreal: Deepfake Face Detection Using CNN, BiLSTM and Attention Mechanisms.

Ms. Rutuja N. Gandhile, Ms. Akanksha S. Nagdeve

Department of Computer Science and Engineering, Deogiri Institute of Engineering and Management
Studies, Chhatrapati Sambhajanagar

ABSTRACT: Deepfake technology has rapidly evolved with the advancement of Artificial Intelligence (AI) and Deep Learning techniques, enabling the creation of highly realistic manipulated images and videos. Although deepfake generation methods provide benefits in entertainment, education, and media production, they also create serious threats related to misinformation, identity theft, privacy violations, and cybercrime. The increasing accessibility of deepfake generation tools has made the detection of manipulated media an important research area in computer vision and cybersecurity. Various deep learning approaches such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Attention Mechanisms have been developed to improve deepfake detection accuracy. CNN models are effective in extracting spatial facial features, while BiLSTM networks help capture temporal inconsistencies in video sequences. Attention mechanisms further enhance model performance by focusing on important manipulated regions within frames. This review paper presents a comprehensive study of AI-based deepfake face detection techniques with special focus on CNN, BiLSTM, and Attention-based approaches. The paper reviews recent research works, datasets, methodologies, advantages, and limitations of existing deepfake detection systems. Furthermore, the study identifies major research gaps such as poor cross-dataset generalization, high computational complexity, and limited real-time detection capability. The review concludes that hybrid deep learning models integrating CNN, BiLSTM, and Attention mechanisms can significantly improve the robustness and accuracy of deepfake detection systems and support future development of secure digital media authentication technologies.

KEYWORDS: Deepfake Detection, Artificial Intelligence, Deep Learning, CNN, BiLSTM, Attention Mechanism, Face Manipulation, Computer Vision, Video Forensics, Cyber security

I. INTRODUCTION

Deepfake technology has emerged as one of the most advanced applications of Artificial Intelligence (AI) and Deep Learning in recent years. The term “deepfake” refers to digitally manipulated images, audio, or videos generated using deep learning algorithms, particularly Generative Adversarial Networks (GANs). These techniques can create highly realistic fake media in which a person’s face, voice, or expressions are replaced or modified to appear authentic. Due to the rapid advancement of AI technologies and the easy availability of deepfake generation tools, manipulated digital content has become increasingly difficult to identify with the human eye.

Deepfake technology has several positive applications in entertainment, virtual reality, education, gaming, and film production. However, its misuse has created serious concerns related to misinformation, fake news, identity theft, political manipulation, cybercrime, financial fraud, and privacy violations. Fake videos and images can spread rapidly through social media platforms and significantly affect public trust and digital security. As a result, deepfake detection has become an important research area in computer vision, multimedia forensics, and cybersecurity.

Traditional image and video verification methods are often unable to detect modern AI-generated fake media because deepfake generation techniques continuously improve visual quality and realism. To overcome these challenges,

researchers have developed intelligent deep learning-based detection systems capable of identifying subtle inconsistencies present in manipulated media. Various Artificial Intelligence techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Attention Mechanisms are widely used for deepfake face detection.

CNN models are highly effective in extracting spatial facial features and identifying manipulation artifacts within images and video frames. BiLSTM networks improve detection performance by learning temporal dependencies and frame-level inconsistencies in deepfake videos. Attention mechanisms further enhance the system by focusing on important facial regions and manipulated patterns that contribute most to classification accuracy. The integration of CNN, BiLSTM, and Attention techniques provides a powerful hybrid framework for improving deepfake detection robustness and reliability.

This review paper presents a comprehensive study of AI-based deepfake face detection techniques with special emphasis on CNN, BiLSTM, and Attention-based approaches. The paper reviews existing research works, datasets, methodologies, advantages, and limitations of current deepfake detection systems. Additionally, the study identifies major research gaps and discusses future research directions for developing more accurate, explainable, scalable, and real-time deepfake detection frameworks.

II. REASEARCH REVIEW

[1] Alhudhaif et al. (2025) proposed a hybrid deepfake face detection framework using Dense-Switch CNN, BiLSTM, and Attention Mechanisms for detecting manipulated facial videos. The Dense-Switch CNN model extracted deep spatial facial features and improved feature selection efficiency. BiLSTM networks captured temporal inconsistencies between video frames, while the Attention Mechanism focused on important manipulated facial regions to improve classification accuracy. The proposed framework achieved higher detection accuracy and better robustness compared to traditional CNN-based models. The researchers evaluated the system on benchmark deepfake datasets and demonstrated improved performance for complex manipulated videos. However, the framework required high computational power and large-scale datasets for effective real-time implementation of a hybrid approach of deepfake detection techniques used.

[2] Khan et al. (2024) presented a comprehensive study on deepfake detection model generalization across different architectures, datasets, and pre-training strategies. The researchers evaluated several deep learning models for identifying manipulated facial media using benchmark datasets such as FaceForensics++, DFDC, and Celeb-DF. Their study highlighted that Convolutional Neural Networks (CNNs) achieved strong feature extraction performance, while Transformer-based models improved cross-dataset generalization capability. The research emphasized that most deepfake detection systems perform well only on trained datasets and fail under unseen real-world conditions. The study concluded that improving model generalization is essential for building reliable deepfake detection systems. However, high computational requirements and dataset dependency remained major limitations.

[3] Omar et al. (2024) proposed an ensemble CNN model integrated with self-attention mechanisms for deepfake video detection. The system combined multiple CNN architectures to improve feature extraction and classification accuracy. The self-attention mechanism helped the model focus on manipulated facial regions and subtle forgery artifacts present in video frames. The researchers demonstrated that attention-based deep learning models significantly improve detection performance compared to traditional CNN systems. Their approach achieved higher robustness against complex facial manipulations and compressed video datasets. However, the model required large-scale training datasets and increased computational complexity due to the ensemble architecture.

[4] Masud et al. (2023) developed LW-DeepFakeNet, a lightweight CNN-LSTM-based deepfake video detection framework designed for real-time applications. The proposed model used Convolutional Neural Networks (CNN) for

extracting spatial facial features and Long Short-Term Memory (LSTM) networks for capturing temporal inconsistencies across video frames. The system improved detection efficiency while reducing training time and computational overhead. The researchers showed that combining CNN with temporal sequence learning significantly enhances deepfake video classification performance. The framework also addressed scene changes and class imbalance problems in deepfake datasets. However, the model showed reduced performance on highly compressed and low-quality videos.

[5] Sarah Tipper et al. (2024) investigated the effectiveness of hybrid CNN-LSTM architectures for deepfake video detection. The study analyzed how spatial and temporal feature learning improves the detection of manipulated media content. CNN layers extracted important facial texture features, while LSTM networks learned sequential frame dependencies present in deepfake videos. The researchers demonstrated that hybrid deep learning frameworks achieved higher detection accuracy compared to standalone CNN models. The study also discussed the importance of temporal learning in identifying facial inconsistencies generated by GAN-based deepfake systems. However, the architecture required high processing power and large video datasets for effective performance.

[6] Al-Adwan et al. (2024) proposed a hybrid CNN-RNN model optimized using the Particle Swarm Optimization (PSO) algorithm for deepfake media detection. The proposed system combined CNN feature extraction with recurrent neural networks to improve classification accuracy. The PSO optimization technique helped enhance parameter tuning and model efficiency. The researchers demonstrated that hybrid architectures can effectively identify manipulated images and videos with improved prediction performance. The system achieved high accuracy on multiple benchmark datasets and showed better adaptability to various deepfake generation techniques. However, optimization processes increased model complexity and training time.

[7] Al-Dulaimi et al. (2024) introduced a transfer learning-based hybrid CNN-LSTM framework for precision deepfake image detection. The proposed model utilized pre-trained CNN networks for extracting deep visual features and LSTM layers for sequential pattern analysis. Transfer learning reduced training requirements and improved detection accuracy on limited datasets. The study showed that combining CNN and LSTM models improves the ability to identify manipulated facial artifacts and inconsistencies. The researchers highlighted that transfer learning approaches reduce computational costs while improving system efficiency. However, the framework still struggled with highly realistic GAN-generated images.

[8] Thing (2023) conducted a comparative analysis between CNN-based and Transformer-based deepfake detection systems. The study evaluated multiple deep learning architectures using popular deepfake datasets including FaceForensics++, Google DFD, Celeb-DF, and DFDC. CNN models demonstrated strong spatial feature extraction capability, while Transformer models improved global feature learning and dataset generalization. The research concluded that hybrid architectures combining CNN and attention mechanisms can improve detection robustness and accuracy. However, Transformer-based systems required significantly higher computational resources and training data.

[9] Hashmi et al. (2023) proposed AVTENet, an Audio-Visual Transformer-based Ensemble Network for multimodal deepfake video detection. The framework combined visual and audio analysis using transformer-based architectures to detect manipulated media content. The proposed system improved detection performance by analyzing inconsistencies between facial movements and speech patterns. The study demonstrated that multimodal detection systems outperform single-modality approaches in identifying sophisticated deepfakes. However, the framework required complex architecture design and high-performance hardware for real-time deployment.

[10] Researchers on multi-attentional deepfake detection proposed attention-guided networks for identifying subtle manipulation artifacts in facial images and videos. The system used multiple spatial attention heads to focus on important manipulated facial regions and improve classification accuracy. The study highlighted that attention-based mechanisms significantly enhance the ability of deep learning systems to detect fine-grained forgery patterns. The approach achieved state-of-the-art performance on several benchmark datasets. However, attention-based models increased overall computational complexity and training requirements.

III. REASEARCH GAP

From the reviewed literature, it is observed that many existing deepfake detection systems mainly focus on improving detection accuracy using individual deep learning models such as CNN, LSTM, or Transformer-based architectures. However, several limitations still exist in current deepfake detection approaches. Many models perform effectively only on specific datasets and fail to generalize under real-world conditions, compressed videos, or unseen deepfake generation techniques.

Existing systems also face difficulties in detecting highly realistic GAN-generated deepfakes due to subtle facial manipulations and improved visual quality. Some deep learning frameworks require high computational resources, large training datasets, and expensive hardware, making real-time implementation challenging. Additionally, many detection systems lack explainability and transparency, making it difficult to understand how classification decisions are generated.

Although hybrid models combining CNN, BiLSTM, and Attention Mechanisms have shown improved performance, limited research has focused on developing lightweight, scalable, and computationally efficient architectures for practical deployment. Furthermore, several existing studies do not effectively utilize temporal inconsistencies and important facial attention regions together within a unified framework.

Therefore, there is a need for an efficient and robust deepfake face detection system that integrates Dense-Switch CNN, BiLSTM, and Attention Mechanisms to improve spatial feature extraction, temporal learning, classification accuracy, and real-time detection performance across diverse datasets and real-world environments.

IV. CONCLUSION

Deepfake technology has rapidly advanced with the development of Artificial Intelligence and Deep Learning techniques, creating serious challenges related to misinformation, digital security, privacy, and cybercrime. This review paper analyzed various AI-based deepfake face detection approaches using CNN, BiLSTM, Attention Mechanisms, Transformer models, and hybrid deep learning frameworks. The study observed that CNN models are highly effective for spatial feature extraction, while BiLSTM networks improve temporal sequence analysis in manipulated videos. Attention mechanisms further enhance detection performance by focusing on important manipulated facial regions.

The literature review highlighted that hybrid architectures combining CNN, BiLSTM, and Attention techniques provide better accuracy and robustness compared to traditional deepfake detection systems. However, several challenges still exist, including poor cross-dataset generalization, high computational complexity, lack of explainability, and limited real-time deployment capability.

The study concludes that integrating Dense-Switch CNN, BiLSTM, and Attention Mechanisms can significantly improve deepfake detection performance and support the development of more reliable and scalable detection systems. Future research should focus on developing lightweight, explainable, and real-time deepfake detection frameworks capable of handling diverse real-world manipulated media environments.

REFERENCES

- [1] A. Alhudhaif et al., “Deepfake Face Detection Using Dense-Switch Convolutional Neural Network, BiLSTM, and Attention Mechanism,” *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3570761.
- [2] S. Khan et al., “Deepfake Detection: Analysing Model Generalisation Across Architectures, Datasets and Pre-Training Paradigms,” 2024.
- [3] H. Omar et al., “Deepfake Video Detection Using Ensemble CNN and Self-Attention Mechanisms,” *Neural Computing and Applications*, vol. 36, 2024.
- [4] M. Masud et al., “LW-DeepFakeNet: Lightweight CNN-LSTM Framework for Deepfake Video Detection,” *Signal, Image and Video Processing*, vol. 17, 2023.
- [5] S. Tipper et al., “Hybrid CNN-LSTM Architectures for Deepfake Video Detection,” *Applied Sciences*, vol. 14, no. 2, 2024.
- [6] M. Al-Adwan et al., “Hybrid CNN-RNN Deepfake Detection Using Particle Swarm Optimization,” *Computers*, vol. 13, no. 4, 2024.
- [7] A. Al-Dulaimi et al., “Transfer Learning-Based CNN-LSTM Framework for Deepfake Image Detection,” *Electronics*, vol. 13, no. 9, 2024.
- [8] V. Thing, “A Comparative Study of CNN and Transformer-Based Deepfake Detection Systems,” 2023.
- [9] M. Hashmi et al., “AVTENet: Audio-Visual Transformer Ensemble Network for Deepfake Video Detection,” 2023.
- [10] Y. Li et al., “Multi-Attentional Deepfake Detection Network for Face Manipulation Detection,” *IEEE Transactions on Information Forensics and Security*, vol. 18, 2023.
- [11] S. Dasgupta et al., “Lightweight CNN with SE Block Attention for Efficient Deepfake Detection,” 2025.