PAST KNOWLEDGE PEOPLE

Engineering : Open Access ISSN: 2993-8643

Review Article

Engineering: Open Access

Advances in Deepfake Detection: A Simple Review

Anjali Kshatriya*, Tulsi Patel and Komal Prajapati

Department of Artificial Intelligence and Data Science School of Engineering and Technology College Gujarat Technological University, India

***Corresponding Author**
Anjali Kshatriya, Department of Artificial Intelligence and Data Science School of Engineering and Technology College Gujarat Technological University, India.

Submitted: 2025, Jul 23; **Accepted:** 2025, Aug 25; **Published:** 2025, Sep 03

Citation: Kshatriya, A., Patel, T., Prajapati, K. (2025). Advances in Deepfake Detection: A Simple Review. *Eng OA, 3*(9), 01-04.

Abstract

Deepfake technology, powered by AI and deep learning, provides hyper-realistic synthetic media that may serve entertainment uses as well as endanger reality with misinformation, privacy, and ethical threats. This surveyed study provides a review of recent progress in deepfake detection for facial manipulation, voice synthesis, and audio-visual forgery, detailing techniques that employ CNNs, graph neural networks, BiLSTM with attention, multimodal fusion, and meta-learning. We also discuss broader challenges of real-world deployment, dataset bias, fairness, and cross-domain generalization. We aim to provide a resource for researchers wanting to research better deepfake detection techniques and mitigate the misuse of deepfake technology in digital ecosystems.

Index Terms: Deepfake Detection, Video Forensics, Fake Speech Detection, Fake Video Detection, Deep Learning, Forgery Detection

I. Introduction

1.1 Overview

Deepfake technology includes generative adversarial networks (GA Ns) and advanced artificial intelligence (AI) that allow for a hyper-realistic manipulation of audio, video, and images. While deepfake technology was first popularized for facial swaps, deepfake technology now encompasses synthetic speech and audio and even videos that incorrectly depict someone acting or saying something; therefore, there are an increased risk of activities like misinformation campaigns, identity fraud, and ethical breaches. Several well-documented examples, including falsifications of political speeches and celebrities impersonating, have highlighted concerning threats to fundamental tenets of society such as trust, legal systems, and cybersecurity. Given that deepfakes are now surpassing traditional forensic techniques in creation, there is a leading urgency to establish new and enhanced methods to detect misuse and therefore bolster detection efforts.

Detection techniques at the outset of deepfake detection used normal handcrafted features and traditional methods, existing state-of-the-art detection techniques used armory of cutting-edge deep learning methods including mix graph neural nets, BiLSTM with attention, meta-learning, multi-modal fusion with bias analysis, were used to combat sophisticated forgeries. This scoping review will study seven significant studies conducted in 2022-2024 that include audio, facial and audiovisual deepfakes. The goal is to scan the literature and identify state-of-the-art techniques and performance, de-termine what is useful in the real world and consider fairness and scaling, investigate specific issues with deployment, and provide future research implications. The review conducted an exhaustive search of IEEE and EURASIP journals, synthesiz- ing knowledge and advances in the field as well as trying to provide both an outline of literature gaps and informed detection strategies in an evolving field of deepfake detection.

1.2 Types of Deepfakes

1.2.1 Facial Swapping Deepfakes: This includes identity swapping, expression swapping, and age or gender swapping.

1.2.2 Audio Deepfakes: This includes voice cloning, emotion manipulation, and language translation.

1.2.3 Full-Body Deepfakes: This includes body swap-ping, body gesture manipulation, and tailoring clothing.

1.2.4 Text-to-Speech Deepfakes: This includes normal- sounding voices, custom voice generation, and new language adaptation.

1.2.5 Deepfake-Generated Images: This includes face manipulation, object manipulation, and scene ma- nipulation.

1.3 Deepfake Algorithms

Deepfake algorithms, therefore, can be organized into two broad categories:

1.4 Face Swapping

This involves the replace- ment of a person's face in videos or images. They did this using CNN early on (2017), which was able to produce high-quality still image swaps, but dropped the ball with the temporal dynamics associated with a video. Recent research was able to improve blending and occlusion while enabling techniques such as FaceSwap-GAN and FaceShifter [1]. Furthermore, and in lieu of an improvement in access, users now have access to an open-source complex to the project, such as Deep- FaceLab, and with numerous concerns for the misuse of this technology in spreading misinformation or defam- atory reputations. The generalization of detection even with diverse datasets has proved difficult.

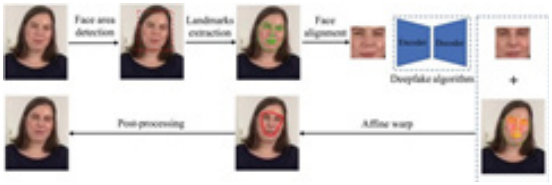

Figure 1: Generation of Face Swapping Video Frames (Driven by CNN) [1].

1.5 Face Reenactment

The second technique re- purposes facial expressions to alter one's facial expressions to mimic another person and poses significant ethical issues. Earlier (2006) template-based techniques lacked temporal coherence Specifically, this method uses sequencing, for example, algorithms such as Face2Face (2016) have real-time expression transfer to (mediocrely) mimic person A or B's expression. Furthermore, mouth synthesis, gaze direction, head motion has also been improved upon, however complete seamless integration has still been an elusive process, if any- thing the knock on effects are significant. [2].

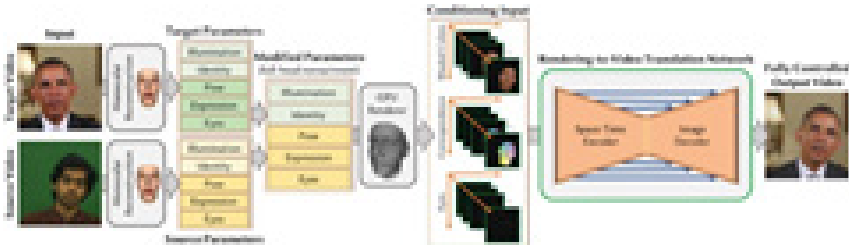

Figure 2: Real-time Face Reenactment Video Synthesis [2].

2. Related Work

This section presents an analysis of seven deep- fake detection studies between 2022 and 2024 in chronological order, discussing methodologies, datasets, results, strengths, weaknesses, and implications in relation to audio and video modalities to provide a broad view of the field.

2.1 Analyzing Fairness in Deepfake Detection with Massively Annotated Databases (2024) [3].

This study investigated biases in deepfake detection by annotating five datasets (e.g., FaceForen- sics++) with 47 different attributes (e.g., age, gender, ethnicity) and testing with three CNN-based models. The results showed that detection accuracy can change by 10% between demographic groups, raising an issue of fairness. This approach is a step forward to recognizing biases in datasets, but it has limitations because it only takes into consideration the known attributes. It suggests important future work in developing useful bias mitigation efforts. This work will add to fairer designs of models for eventual deployment in the real world at places like social media verification.

2.2 BMNet: Enhancing Deepfake Detection Through BiLSTM and Multi-Head Self-Attention Mechanism (2024) [4].

This paper's research proposes BMNet which utilizes BiLSTM to model deepfake via a tem- poral sequence and for a multi-head self-attention mechanism to extract important video frames from the identified use of the deepfake. The testing was done on FaceForensics++ and Celeb-DF datasets before obtaining the the testing evaluations. Over- all BMNet achieved 95.8% accuracy, while this is according to the authors of the paper; BM- Net outperformed the relevant cross-dataset base- line performance of CNNs by 4%. While it offers the best opportunity for identifying temporal in- consistencies, BMNet only consumes high-quality videos, indicating tests assessing deepfake instances in compressed videos would recommended and identify any other testing viability. The efficiency of BMNet enhances forensics in video forensics, which is a pressing topic for both social media, apps, and/or user-generated content platforms.

2.3 Comprehensive Multiparametric Analysis of Human Deepfake Speech Recognition (2024) [5].

This research assesses human understanding of synthetic speech

using YourTTS and adds a realism score similar to Mean Opinion Score. Informed par- ticipants displayed better fake detection but overall high-quality audio fooled most and exposed their vulnerabilities. Conducted in Czech and Slovak, the authors recommend multilingual data to ex- pand upon monolingual restrictions. This research serves as a connection between cognitive science and forensics by providing responses to audio chal- lenges, like spoofing threats such as phishing.

2.4 A Novel Approach for Detecting Deep Fake Videos Using Graph Neural Network (2024) [6].

This paper presents a framework for using the GNN, CNN, and the fusion networks, achieving 99.3% accuracy on controlled data. Though it is great at spatial-temporally decoding data, it does not test against real-world data (e.g., compres- sion). While suggesting to validate against multiple datasets, this paper adds to video forensics - de- tection and appears to relate to content verification, but, has not considered if the GNN is suitable for deployment.

2.5 Fake Speech Detection Using VGGish with Attention Block (2024) [7].

This research employs a VGGish model with CBAM for audio spoofing detection to achieve 95.8% accuracy on ASVspoof 2019. The attention mechanism improves generalization of the model across spoofing types, but we need to test with more diverse datasets. ASVspoof incorporates various attack scenarios, improving the robustness of the model. The results show promising evidence that the model will produce robust voice authentication systems to combat misinformation and possibly be validated more broadly.

2.6 Generalization of Forgery Detection with Meta Deepfake Detection Model (2023) [8].

This research presented a new Meta Deepfake Detection (MDD) meta-learning framework that achieves 92% accuracy on Celeb-DF without requiring re-training on the new domains. The authors optimize multi-domain weights for deeper learning but admit the complexity of the rankings will require exploration of a more simple meta-learning approach. MDD uses realistic deepfakes from Celeb- DF to leverage training's depth in effectiveness, and it will be able to apply in increasingly dynamic contexts by taking less costs for re-training in real- world settings.

2.7 A Deepfake Video Detection Method Based on Multimodal Deep Learning (2022) [9].

This study utilizes a combination of CNN-RNN and audiovisual fusion that achieved 93.5% ac- curacy on DFDC by multi-modal approach distinguishes itself from other models by conducting a primarily multimodal model. This study did struggle with low-quality audio from DFDC, and the authors offered recommendations for noise robust features instead. Given the social media context of the DFDC dataset, this is assuredly grounded in practical relevance. The use of a multimodal approach and model is practical to social media to accept that there will be complex deepfake occurrences that may provide an ineffective deepfake detector.

3. Comparative Analysis of Existing Works

Table I summarizes key aspects of seven deepfake detection works from 2022 to 2024, ordered by year.

Source	Journal	Year	Techniques	Dataset	Key Findings	Highlights	Limitations
Analyzing Fairness in Deepfake Detection with Massively Annotated Databases [1]	IEEE Trans. Multi-media	2024	CNN, bias analysis	FaceFore nsics++	10% acc. variance	Exposes demo-graphic biases	Limited to known attributes
BMNet: Enhancing Deepfake Detection Through BiLSTM and Multi-Head Self-Attention Mecha- nism [2]	IEEE Trans. Pattern Anal.	2024	BiLSTM, self-attention	Celeb-DF	95.8% acc. Celeb-DF	Robust temporal detection	High-quality inputs needed
Comprehensive Multipara- metric Analysis of Human Deepfake Speech Recogni- tion [3]	EURASIP J. Image Video	2024	YourTTS, human exp.	Czech/ Slovak data	Humans miss fakes	Reveals human vulnera- bilities	Monolingual study
A Novel Approach for Detecting Deep Fake Videos Using Graph Neural Network [4]	J. Big Data	2024	GNN, CNN, fusion	Controlled datasets	99.3% acc.	Strong feature extraction	Lacks real-world testing
Fake Speech Detection Using VGGish with Attention Block [5]	EURASIP J. Audio Speech	2024	VGGish, CBAM	ASVspoof 2019	95.8% acc. ASVspoof	Robust audio detection	Limited dataset scope
Generalization of Forgery Detection with Meta Deepfake Detection Model [6]	IEEE Access	2023	Meta-learning	Celeb-DF	92% acc. Celeb- DF	Adapts across domains	Complex architec- ture
A Deepfake Video Detection Method Based on Multimodal Deep Learning [7]	IEEE Trans. Cir- cuits Syst.	2022	CNN- RNN, fusion	DFDC	93.5% acc. DFDC	Effective multi-modal fusion	Low- quality audio issues

Table 1: Comparative Analysis of Deepfake Detection Research Papers (2022-2024)

3. Problem Statement Identification

Current deepfake detection methods excel in con- trolled settings, with accuracies like 99.3% and 95.8% but struggle in real-world scenarios. Generalization to unseen domains remains limited, as seen with only 92% accuracy on new datasets, and biases in datasets (e.g., demographic disparities causing 10% accuracy variance [1]) hinder fairness. Real-world challenges like compression and noise reduce performance significantly while multi- modal approaches, though promising (93.5% accu- racy [7]), falter with low-quality inputs. Scalability and computational efficiency are under addressed, and human detection remains unreliable [3]. A unified, fair, and scalable detection system that generalizes across modalities and conditions is urgently needed to counter sophisticated deepfake threats.

4. Conclusion

This survey underscores rapid progress in deep- fake detection, with techniques like GNNs BiLSTM with self-attention meta-learning and multimodal fusion pushing boundaries. Yet, challenges persist in generalization fairness and real-world robustness [1-7]. Future research should focus on bias-free datasets, lightweight mod- els for scalability, and integrated audio-visual detection to address complex forgeries. Combining human insights with automated systems will enhance reliability, mitigating the growing threat of synthetic media in our digital ecosystem [10,11].

Reference

1. Z. Zhang and Q. Wang, ”Generalization of forgery detection with meta deepfake detection model,” *IEEE Access,* vol. 11, Art. no. 10376174, 2023.
2. Rashid, M. M., Lee, S. H., & Kwon, K. R. (2021). Blockchain technology for combating deepfake and protect video/image integrity. 멀티미디어학회논문지, *24*(8), 1044-1058.
3. Fernandez and B. Smith,”Analyzing fairness in deepfake detection with massively annotated databases,” *IEEE Trans. Multimedia,* vol. 26, Art. no. 10438899, 2024.
4. Xiong, D., Wen, Z., Zhang, C., Ren, D., & Li, W. (2025). BMNet: Enhancing Deepfake Detection Through BiLSTM and Multi-Head Self-Attention Mechanism. *IEEE Access.*
5. Malinka, K., Firc, A., Šalko, M., Prudký, D., Radačovská, K., & Hanáček, P. (2024). Comprehensive multiparametric analysis of human deepfake speech recognition. *EURASIP Journal on Image and Video Processing, 2024*(1), 24.
6. El-Gayar, M. M., Abouhawwash, M., Askar, S. S., & Sweidan, S. (2024). A novel approach for detecting deep fake videos using graph neural network. *Journal of Big Data, 11*(1), 22.
7. Kanwal, T., Mahum, R., AlSalman, A. M., Sharaf, M., & Hassan, H. (2024). Fake speech detection using VGGish with attention block. *EURASIP Journal on Audio, Speech, and Music Processing, 2024*(1), 35.
8. S. Anil and S. S. Nair, ”Deepfake detection using deep learning,” *Int. Res. J. Eng. Technol*., vol. 10, no. 5, pp. 27–32, May 2023.
9. X. Li and H. Yang, ”A deepfake video detection method based on multimodal deep learning,” *IEEE Trans. Circuits Syst. Video Technol*., vol. 32, no. 12, Art. no. 9999180, Dec. 2022.
10. CyberPeace Foundation, ”The efforts of social media platforms to counter deepfake,” CyberPeace. Available at https://www.cyberpeace.org/resources/blogs/the-efforts- of-social-media-platforms-to-counter-deepfake.
11. Spiceworks, ”What is deepfake? Definition, examples, and how to identify,” Spiceworks. Available at https://www.spiceworks.com/it-security/cyber-risk- management/articles/what-is-deepfake/, [Accessed: Apr. 15, 2025].