A Review on Deepfake Detection Techniques using Machine Learning and Deep Learning

Anmol Verma[1], Aadit Sharma[2], Ayush Tawar[3], Ansh Dubey[4], Abbas Electric[5]*

Computer Science and Engineering, Indore Institute of Science and Technology, Rau, Pithampur Road, Indore, 453331, Madhya Pradesh, India

How to Cite this Article:
Verma, A., Sharma, A., Tawar, A., Dubey, A. & Electric, A. (2026). A Review on Deepfake Detection Techniques using Machine Learning and Deep Learning. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.540

https://doi.org/10.55041/ijcope.v2i5.540

Abstract. Deepfake technology has rapidly evolved with the advancement of artificial intelligence, enabling the creation of highly realistic synthetic audio and video content. While such technology has beneficial applications in entertainment and virtual reality, it also poses serious threats in terms of misinformation, identity fraud, and privacy violations. Consequently, deepfake detection has emerged as an important research area within cyber security and artificial intelligence. This paper presents a comprehensive review of deepfake detection techniques based on machine learning and deep learning approaches. Existing methods are analyzed with respect to detection strategies, datasets, and performance characteristics. A comparative analysis of prominent techniques is provided to highlight their strengths and limitations. Furthermore, major challenges such as generalization, dataset bias, and computational complexity are discussed along with potential future research directions. This review aims to provide a concise understanding of current advancements in deepfake detections and to serve as a reference for researchers and practitioners in this domain.

Keywords: Deepfake Detection, Machine Learning, Deep Learning, Cyber Security, Artificial Intelligence

I. Introduction

The rapid growth of artificial intelligence has enabled the development of sophisticated media generation techniques capable of producing highly realistic synthetic images, videos, and audio. One such phenomenon is deepfake technology, which primarily relies on deep learning models to manipulate or generate human facial expressions, speech, and gestures. Although deepfakes have legitimate uses in areas such as film production and virtual avatars, their malicious misuse has become a serious concern.

Deepfakes can be exploited for spreading misinformation, political manipulation, financial fraud, and identity theft. The increasing realism of such forged media makes it difficult for humans to distinguish between authentic and manipulated content. As a result, automated deepfake detection has become essential to ensure media integrity and digital trust.

Recent research has focused on detecting deepfakes using machine learning (ML) and deep learning (DL) techniques. These approaches analyze visual artifacts, temporal inconsistencies, and physiological cues present in forged media. This paper reviews major deepfake detection techniques, commonly used datasets, and highlights the challenges faced by existing systems in real-world deployment scenarios.

II. Deepfake Generation Techniques

Deepfake generation techniques are primarily driven by advances in deep learning architectures. One of the most widely used approaches is Generative Adversarial Networks (GANs), where a generator and a discriminator compete to produce increasingly realistic synthetic content. GAN-based methods are commonly used for face swapping and facial reenactment.

Another category includes autoencoder-based models that learn compact facial representations and reconstruct manipulated faces. These methods are computationally efficient and were among the earliest deepfake generation techniques. Additionally, recent approaches employ diffusion models and transformer-based architectures to improve visual fidelity and temporal consistency.

Audio deepfakes involve voice cleaning using neural text-to-speech systems and speaker embedding models. The continuous evolution of generation techniques makes detection increasingly challenging, requiring robust and adaptive detection mechanisms.

III. Literature Review

Several studies have explored deepfake detection using different strategies and datasets. Rana et al. conducted a systematic literature review covering more than one hundred deepfake detection studies and categorized existing approaches into deep learning, classical machine learning, statistical, and blockchain-based methods. Their analysis indicates that deep learning approaches generally outperform traditional machine learning techniques.

Rossler et al. introduced the FaceForensics++ dataset, which has become a standard benchmark for evaluating deepfake detection methods. Their work demonstrated that deep convolutional neural networks, such as Xception, achieve strong detection performance even under compressed video conditions.

Dolhansky et al. presented the Deepfake Detection Challenge (DFDC) dataset, designed to improve the robustness of detection models by incorporating diverse actors, environments, and manipulation techniques. This dataset highlighted the importance of generalization across different real-world conditions.

Afchar et al. proposed MesoNet, a compact convolutional neural network that focuses on mesoscopic image properties. Their model achieved high detection accuracy while maintaining low computational complexity. Li and Lyu introduced a method based on detecting face warping artifacts caused by resolution inconsistencies during manipulation. Additionally, Ciftci et al. proposed FakeCatcher, which detects deepfakes using physiological signals such as subtle blood flow patterns extracted from facial re- gions.

IV. Comparative Analysis

A comparison of prominent deepfake detection techniques is summarized in Table I.

Table 1. Comparison of Deepfake Detection Techniques

Author	Year	Technique	Dataset	Limitation
Rana et al.	2022	Systematic Survey	FF++, DFDC	No new detector
Rossler et al.	2019	CNN (Xception)	FaceForensics++	Limited real-world diversity
Dolhansky et al.	2019	Dataset-based	DFDC	Dataset-focused
Afchar et al.	2018	Compact CNN	Public datasets	Generalization issues

Li & Lyu	2019	Artifact-based CNN	FaceForensics++	Reduced artifacts in new fakes
Ciftci et al.	2019	Physiological signals	Curated datasets	Sensitive to compression

This comparison highlights that while deep learning-based methods achieve high accuracy, they often struggle with generalization and computational requirements.

V. Challenges in Deepfake Detection

Despite significant progress, deepfake detection faces several challenges. One major issue is the rapid improvement of deepfake generation techniques, which reduces detectable artifacts. Dataset bias is another challenge, as models trained on specific datasets often fail to generalize to unseen manipulations.

High computational cost limits the deployment of deepfake detectors in real-time systems. Additionally, video compression, low resolution, and varying lighting conditions significantly degrade detection performance. These challenges necessitate the development of more robust and adaptive detection methods.

VI. Future Scope

Future research in deepfake detection may focus on developing generalized models that perform consistently across datasets and manipulation techniques. Hybrid approaches combining spatial, temporal, and physiological cues show promise in improving robustness.

Explainable AI techniques can enhance transparency and trust in detection systems. Furthermore, real-time detection systems optimized for low-resource environments are crucial for practical deployment. The creation of large- scale, diverse, and continuously updated datasets will also play a vital role in advancing this field.

VII. Conclusion

Deepfake technology poses a growing threat to digital media authenticity and cyber security. This paper reviewed existing deepfake detection techniques based on machine learning and deep learning approaches. A comparative analysis highlighted the strengths and limitations of popular methods. Although deep learning-based detectors achieve high accuracy, challenges such as generalization, computational complexity, and evolving manipulation techniques remain unresolved. Continued research is required to develop robust, efficient, and trustworthy deepfake detection systems capable of addressing real-world threats at scale across evolving digital ecosystems.

Acknowledgement

The author would like to express sincere gratitude to the Department of Computer Science and Engineering, Indore Institute of Science and Technology, for providing the necessary support and academic environment to carry out this work. The author is also thankful to the faculty members for their guidance and encouragement throughout the preparation of this paper.

References

[1] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," IEEE Access, vol. 10, pp. 25494–25513, 2022.

[2] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2019.

[3] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) Preview Dataset," arXiv:1910.08854, 2019.

[4] D. Afchar et al., "MesoNet: a Compact Facial Video Forgery Detection Network," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018.

[5] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[6] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," arXiv:1901.02212, 2019