International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comprehensive Review on Deepfake Detection Techniques Based on Deep Learning

Kalpana Kumari[1], Nitesh Gupta[2]

[1]MTech Scholar, CSE, Department, NIIST, Bhopal Kalpanaraj11433@gmail.com
[2]Associate Professor, CSE, Department, NIIST, Bhopal 9.nitesh@gmail.com

DOI : https://doi.org/10.55248/gengpi.6.0825.3068

ABSTRACT:

A Deepfake fake media that successfully manipulate visual or auditory content have proliferated due to the quick development of deep learning and generative models, posing serious risks to digital trust, privacy, and security. Detecting such fabricated content has become a critical research challenge, as deepfakes continue to advance in realism and complexity. This paper presents a comprehensive review on deepfake detection techniques leveraging deep learning approaches. Also analyze state-of-the-art methodologies including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformer-based architectures, and hybrid models, highlighting their strengths, limitations, and application domains. In this review discuss the benchmark datasets, performance evaluation metrics, and real-world implementation and challenges. This work aims to provide valuable insights for developing explainable and adaptive solutions to combat the growing threat of deepfakes.

Keywords— DeepFake Detection, Image classification, Computer Vision, Deep learning, Model generalization, Machine Learning

INTRODUCTION

The beginning of artificial intelligence (AI) and, more specifically, deep learning has revolutionized various domains, including computer vision, natural language processing (NLP), and multimedia generation. Among the most prominent developments is the rise of generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders, which have enabled the creation of highly realistic synthetic media known as deepfakes. A deepfake is a manipulated piece of content either text, image and video, or audio that is generated or altered using deep learning techniques to imitate real-world data with near-perfect accuracy. Even as these technologies have demonstrated remarkable potential for creative and educational purposes, their misuse for malicious activities such as identity theft, political misinformation, cyber bullying, and financial fraud has raised significant concerns regarding digital security and societal trust. The main purpose of deepfake research is to achieve high accuracy in detecting various forms of fake videos on social media or the web while minimizing false positives [1].

The main reason deepfakes pose a severe threat is their ability to deceive even the most vigilant human observers and traditional forensic detection systems. As various machine learning and deep learning models continue to develop, so do the sophistication and realism of fake content, making detection an increasingly challenging task. Unlike traditional photo or video editing, which often leaves visible artifacts, deepfakes utilize complex neural architectures that minimize detectable inconsistencies. Consequently, there is an urgent need for robust and adaptive detection methods capable of identifying subtle traces of manipulation while maintaining high accuracy across diverse datasets and real-world conditions. To counter this growing threat, researchers have explored deep learning-based detection approaches, which leverage the power of neural networks to automatically extract discriminative features and identify anomalies introduced during content generation. These methods encompass various architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformer-based models, and hybrid solutions that combine spatial and temporal cues for improved performance [2]. CNN-based techniques are widely used for image and frame-level analysis, detecting visual artifacts and inconsistencies, whereas RNN and LSTM networks focus on temporal dependencies in videos. More recently, attention-based and Transformer models have demonstrated promising results by capturing long-range dependencies in multimodal data.

Many researchers uses hybrid deep learning algorithms and natural evolution optimization approach is chosen due to its potential to improve the detection rates of deepfake compared to traditional models [3]. Other techniques, such as traditional image and video analysis, may not be effective in detecting deepfake due to the advanced AI and ML techniques used to create them [4]. Deepfake are highly realistic, and traditional detection methods may fail to identify them, as they are not designed to account for the complexity and sophistication of AI-generated media. Therefore, the proposed approach of using advanced deep learning algorithms and optimization techniques is essential for improving the accuracy and reliability of deepfake detection. Considering how common deepfakes are becoming and the possible harm they can inflict, such as identity theft, extortion, sexual exploitation, reputational damage, and harassment, it is crucial to have efficient detecting techniques to mitigate their effects. This research addresses this need by leveraging advanced technologies to improve the accuracy of deepfake detection, ultimately contributing to a safer and more secure digital environment [5].

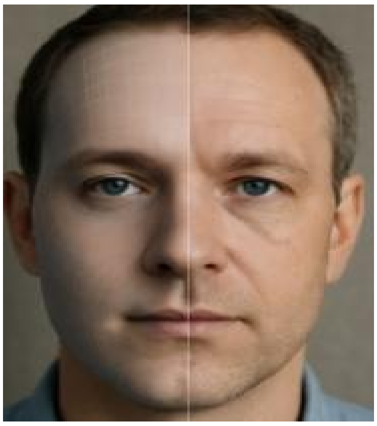

Figure 1.1: Deep Fake Image

This paper provides a comprehensive review of deepfake detection techniques utilizing deep learning, with an emphasis on categorizing existing methods, comparing their effectiveness, and identifying limitations that hinder real-world deployment.

LITRETURE REVIEW

Various Researchers studies have explored deep learning-based techniques for deepfake detection, leveraging architectures such as CNNs, RNNs, LSTM and Transformers to identify subtle inconsistencies in manipulated content. Existing research focuses on diverse approaches, including image-based analysis, temporal modeling for videos, and multimodal frameworks to enhance detection accuracy.

Authors [1] investigate the fundamental technologies, such as deep learning models, and evaluate their efficacy in differentiating real and manipulated media. In addition, we explore novel detection methods that utilize sophisticated machine learning, computer vision, and audio analysis techniques. This work represents the newest developments in the deepfake research area. In an era where distinguishing fact from fiction is paramount, Author’s objectives to enhanced the security and awareness of the digital ecosystem by advancing our understanding of autonomous detection and evaluation methods.

In this work [2], a hybrid convolutional neural network (CNN) and recurrent neural network (RNN) with a particle swarm optimization (PSO) algorithm is utilized to demonstrate a deep learning strategy for detecting deepfake videos. High accuracy, sensitivity, specificity, and F1 score were attained by the proposed approach when tested on two publicly available datasets that is Celeb-DF and the Deepfake Detection Challenge Dataset (DFDC). The proposed method achieved an average accuracy of 97.26% on Celeb-DF and an average accuracy of 94.2% on DFDC. The results were compared to other state-of-the-art methods and showed that the proposed method outperformed many. This method can effectively detect deepfake videos, which is essential for identifying and preventing the spread of manipulated content online.

This research work [3] proposed a deep learning (DL)-based techniques for detecting deepfakes. The system comprises three components: preprocessing, detection, and prediction. Preprocessing includes frame extraction, face detection, alignment, and feature cropping. Convolutional neural networks (CNNs) are employed in the eye and nose feature detection phase. A CNN combined with a vision transformer is also used for face detection. The model is trained on various face images using Face Forensics?? and DFDC datasets. Multiple performance metrics, including accuracy, precision, F1, and recall, are used to assess the proposed model’s performance. The experimental results indicate the potential and strengths of the proposed CNN that achieved enhanced performance with an accuracy of 97%, while the CViT-based model achieved 85% using the Face Forences?? Dataset and demonstrated significant improvements in deepfake detection compared to recent studies, affirming the potential of the suggested framework for detecting deepfakes on social media. This work contributes to a broader understanding of CNN-based DL methods for deepfake detection.

In this work [4], Author implemented a customized CNN algorithm to identify deepfake pictures from a video dataset and conducted a comparative analysis with two other methods to determine which way was superior. The Kaggle dataset was used to train & test our model. Convolutional neural networks (CNNs) have been used in this research to distinguish authentic & deepfake images by training three distinct CNN models. A customized CNN model, which includes several additional layers such as a dense layer, MaxPooling, as well as a dropout layer, has also been developed and implemented. These methods follows the frames extraction, face feature extraction, data preprocessing, and classification phases in determining whether real or fake images in the video. This work achieves 91.4 % accuracy and loss value of 0.342.

Authors [5] reviewed a deep learning approach that has shown a remarkable performance in deepfake detection, the quality of deepfake has been increasing. the current deep learning methods need to improve as well to successfully identify fake videos and images. Additionally for the current deep learning methods, there is not a clear method to know the number of layers needed and which architecture is appropriate for deepfake detection. Another area of investigation is the incorporation of identification of deepfake detection methods into social media platform in order to improve their effectiveness in coping with the pervasive effects of deepfakes and reduce its impacts.

Authors [6] work on the data challenges such as unbalanced datasets and inadequate labelled training data. Training challenges include the need for many computational resources. It also addresses reliability challenges, including overconfdence in detection methods and emerging manipulation approaches. The research emphasises the dominance of deep learning-based methods in detecting deepfakes despite their computational efficiency and generalisation limitations. on the other hand it also acknowledges the drawbacks of these approaches, such as their limited computing efficiency and generalisation. The research also critically evaluates deepfake datasets, emphasising the necessity for good-quality datasets to improve detection methods. The study also indicates major research gaps, guiding future deepfake detection research.

FINDINGS OF THE SURVEY

This Review on deepfake detection techniques using deep learning provides many key observations about the current research landscape. Firstly, CNN-based architectures dominate image-based detection, as they stand out in identifying spatial inconsistencies such as blending artifacts, unnatural textures, or pixel-level anomalies introduced during manipulation. These methods are effective for static images but often fail to capture temporal cues necessary for video-based detection. To address this, researchers have incorporated RNNs and LSTM networks, which analyze frame sequences to identify motion-related artifacts, leading to improved performance in detecting dynamic manipulations. Recently, Transformer-based models and attention mechanisms have emerged as powerful alternatives due to their ability to capture long-range dependencies and multi-modal relationships across data types. Another major finding is the increasing shift toward multimodal detection frameworks that combine visual, audio, and textual features for improved robustness. These systems demonstrate superior performance when compared to single-modality approaches. However, the survey also highlights limitations, including poor generalization to unseen datasets, vulnerability to adversarial attacks, and the need for large, diverse training datasets. Current models often suffer a drop in accuracy when tested on real-world or compressed videos, making practical deployment challenging. Table 2.1 provides a comparative summary of recent research work focused on deepfake detection. It outlines the proposed methodologies, key results, relevant observations, limitations in trending deep learning based detection techniques. The key observations from the findings highlights that convolution neural networks (CNNs) continue to dominate static image based deepfake detection due to their strong feature extraction capabilities. Recurrent models such as RNNs and LSTMs enhance temporal analysis in videos, capturing frame-wise dependencies effectively. Furthermore, challenges related to scalability and interpretability hinder the practical deployment of these models in real-world application.

Table 2.1: Comparative Summary of Deep Learning Based Deepfake Detection

Reference	Proposed Work	Result	Remark
[1] Reshma Sunil et. Al	Investigation of deep learning and multimedia analysis techniques for deepfake detection.	Focus on evaluating foundational models and advancing autonomous detection.	Highlights importance of digital security and awareness.
[2] Aryaf Al-Adwan et. Al	Hybrid CNN-RNN model with PSO algorithm for deepfake detection.	Achieved 97.26% (Celeb-DF), 94.2% (DFDC).	Outperformed many existing methods.
[3] Ahmed Hatem Soudy1 et. Al.	CNN and Vision Transformer (CViT)-based model for face feature-based deepfake detection.	CNN accuracy: 97%, CViT accuracy: 85% on Face Forensics and DFDC.	Showed strong results in social media context.
[4] Usha Kosarkar et. Al.	Customized CNN model with multiple layers to classify deepfake images.	Accuracy: 91.4%, Loss: 0.342 using Kaggle dataset.	Compared with two other models, performed well.
[5] Abdulqader M. Almars	Review of current DL-based detection systems and their limitations.	Noted increase in deepfake quality vs. detection capability.	Suggested integration into social media and architectural optimization.
[6] Achhardeep Kaur et. Al.	Analysis of data imbalance, model reliability, and detection challenges.	Identified need for better datasets and handling overconfidence.	Outlined key future research gaps and challenges.

Furthermore, interpretability remains critical gaps in existing systems. Many detection models function as black boxes, creating trust issues for legal or forensic use cases. Therefore, future research should integrate Explainable AI (XAI), federated learning, and lightweight architectures for real-time and privacy-preserving applications.

CONCLUSION

The rapid development in generative models such as GANs and autoencoders have significantly contributed to the rise of deepfakes, creating highly realistic manipulated media that pose severe threats to privacy, security, and societal trust. This survey represented a comprehensive review of deep learning-based deepfake detection techniques, highlighting state-of-the-art methods, their strengths, limitations, and application domains. From CNN-based image analysis to RNN and Transformer models for video and multimodal detection, deep learning has emerged as a powerful tool to counter this evolving challenge. Although notable progress, several critical challenges persist. Current detection models often struggle with generalization across diverse datasets and fail against adversarial attacks or unseen manipulation techniques. The lack of large, balanced, and real-world datasets further complicates the development of robust systems. Additionally, deepfake detection faces issues of interpretability, computational cost, and scalability, limiting practical deployment in real-time applications. Furthermore, collaboration between researchers, policymakers, and technology companies is essential to establish ethical guidelines and standards for combating deepfakes effectively. In conclusion, deep learning-based detection offers promising solutions to this emerging threat, but continuous innovation and interdisciplinary efforts are vital to stay ahead of increasingly sophisticated deepfake technologies and ensure digital media integrity in the AI-driven era.

References

[1] Reshma Sunil et. al. "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation" DOI:

https://doi.org/10.1016/j.heliyon.2025.e42273, www.cell.com/heliyon

[2] Aryaf Al-Adwan et. al. "Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm" https://doi.org/10.3390/computers13040099, MPDI 2024

[3] Ahmed Hatem Soudy1 et. al. "Deepfake detection using convolutional vision transformers and convolutional neural networks" Neural Computing and Applications (2024) 36:19759–19775, https://doi.org/10.1007/s00521-024-10181-7(0123456789().,-volV)(0123456789,-().volV)

[4] Usha Kosarkar et. al. "Revealing and Classification of Deepfakes Video's Images using a Customize Convolution Neural Network Model" 10.1016/j.procs.2023.01.237, Procedia Computer Science 218 (2023) 2636–2652

[5] Abdulqader M. Almars "Deepfakes Detection Techniques Using Deep Learning: A Survey" Journal of Computer and Communications, 2021, 9, 20-35 https://www.scirp.org/journal/jcc ISSN Online: 2327-5227

[6] Achhardeep Kaur et. Al. "Deepfake video detection: challenges and opportunities" Artifcial Intelligence Review (2024) 57:159

https://doi.org/10.1007/s10462-024-10810-6

[7] Al-Hussein M, Venkataraman S, Jawahar C (2020) Deepfake detection for video: an open source challenge. arXiv preprint arXiv:2006.06058

[8] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

[9] Dua M, Shakshi Singla R, Raj S, Jangra A (2021) Deep cnn models-based ensemble approach to driver drowsiness detection. Neural Comput Appl 33:3155–3168

[10] Hasan MJ et al (2020) Understanding the influence of epochs and learning rate on deep learning-based sentiment analysis. In: Proceedings of the international conference on artificial intelligence and applications, pp 553–561

[11] Ismail A, Elpeltagy M, Zaki SM, Eldahshan K (2021) A new deep learning-based methodology for video deepfake detection using xgboost. Sensors 21(16):54

[12] A. M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey," J. Comput. Commun., 2021, doi: 10.4236/jcc.2021.95003.

[13] L. Nataraj et al., "Detecting GAN generated Fake Images using Co-occurrence Matrices," 2019, doi: 10.2352/ISSN.2470- 1173.2019.5.MWSF-532.

[14] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot.. For Now," 2020,

doi: 10.1109/CVPR42600.2020.00872.

[15] C. C. Hsu, C. Y. Lee, and Y. X. Zhuang, "Learning to detect fake face images in the wild," 2019, doi: 10.1109/IS3C.2018.00104.

[16] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2019, doi: 10.1109/AVSS.2018.8639163.

[17] F. Sun, N. Zhang, P. Xu, and Z. Song, "Deepfake Detection Method Based on Cross-Domain Fusion," Secur. Commun. Networks, vol.

2021, no. 2, 2021, doi: 10.1155/2021/2482942.