World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

A survey on deep-fake detection algorithms

Kavitha Soppari, Minnu Sri Thumnoori, Sumanth Gangalam * and Dheeraj Kumar Raju Bhatraju

Department of CSE (Artificial Intelligence and Machine Learning) of ACE Engineering College Hyderabad, India.

World Journal of Advanced Research and Reviews, 2025, 26(03), 1123-1127

Publication history: Received on 27 April 2025; revised on 06 June 2025; accepted on 09 June 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.3.2251

Abstract

Since AI technology has been on the rise, applications based in this field are also increasing rapidly. However, some of them are utilizing AI to generate images and videos that display explicit activities with manipulated faces of celebrities or other innocent people, incorporated into them. These images and videos are called Deep Fakes. It causes harm by spreading false information or fake news using social media and other similar applications. Deep fakes are generated using Generative Adversarial Networks also known as GANs and other algorithms which utilize machine learning. However, GANs also perform video deep-fake detection along with Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs). We also used feature extraction to derive basic facial expressions. The best results are obtained using methods based on EfficientNet B7. The accuracy for the state-of-the-art approach in detection is around 88%. Using such mentioned deep learning models, we aim to improve them and increase the accuracy to 93%, with minimal fluctuations to enhance the reliability and robustness of deep fake detection systems.

Keywords: Deep Fake; Deep Fake Detection; Deep Learning; Convolutional Neural Networks (CNNs); Generative Adversarial Networks (GANs); LSTMs

1 Introduction

Deepfakes are a subset of AI-generated media that have grown in notoriety for their ability to deceive viewers. Using deep learning methods, particularly GANs, these systems can learn to generate realistic videos by training on large datasets. Deepfakes have gained widespread attention due to their ability to manipulate public features' appearances and voices, often with harmful intent. The ease with which deepfake videos can be produced has led to serious concerns about their potential for misuse in areas such as cyberbullying, political manipulation, and the violation of personal privacy. GANs, the foundational technology behind many deepfake applications, work by pitting two neural networks against each other: a generator that creates new data, and a discriminator that evaluates the generated data. The generator and discriminator work together in a feedback loop to improve the quality of the generated content. As a result, deepfakes can appear exceedingly convincing, making the task of detection a complex challenge.

To address these challenges, researchers have turned to deep learning-based detection systems. These systems use various machine learning algorithms, such as CNNs and LSTMs, to recognize patterns, artifacts, or inconsistencies in videos and images that may suggest manipulation. CNNs are particularly effective for analyzing visual content, while LSTMs help detect temporal anomalies by analyzing the sequence of frames in a video. While deepfake detection systems have made significant progress, there is still a need for improvement. Existing methods often struggle with maintaining high accuracy rates across diverse datasets and varying types of deepfakes. The proposed model aims to address these shortcomings by leveraging advanced deep learning architectures, particularly EfficientNet B7, a cutting-edge model known for its efficiency and accuracy.

* Corresponding author: Sumanth Gangalam

Deepfake technology, powered by artificial intelligence (AI), has rapidly evolved over recent years. These advancements include its capacity to manipulate and generate highly realistic multimedia content. Leveraging advanced machine learning algorithms such as GANs, deepfakes are allowed to construct and manipulate images, videos, and audios that are difficult to differentiate from real content. This growing ability to fabricate reality presents significant challenges, especially as it pertains to misinformation, fake news, privacy violations, and malicious defamation. With the rise of deepfakes, the need for accurate detection mechanisms has become paramount.

The increasing popularity of deepfake technology is due to its ability to superimpose faces onto videos of people, often innocent civilians, celebrities, or politicians, making them appear to engage in activities in which they did not have a part in. These videos are then circulated through various social media platforms, fueling the spread of harmful and misleading content. As a result, it has become a priority for technologists and researchers to understand and tackle this issue.

The model is focused on using deep learning techniques to detect deepfake videos. These techniques particularly include Long Short-Term Memory Networks, Generative Adversarial Networks, and Convolutional Neural Networks. The goal is to improve upon previous existing detection systems to provide a more reliable, accurate, cutting-edge model to distinguish between the original and manipulated content.

Deepfakes usually come under as a branch of AI-generated media that have grown disreputable for their ability to misguide viewers. Deepfakes have gained enormous attention due to the manipulative ability of public figures' appearances, and voices, often with harmful intent. Since deepfakes are easily constructed, it has led to serious concerns about their potential for misuse in areas such as cyberbullying, political nuisance, and the violation of personal privacy. They use GANs for the producing these deepfakes, they work by taking two neural networks and pit them against each other: a generator and a discriminator. The generator creates new data while the discriminator evaluates the generated data. Both these networks work together in a feedback loop and try to improve the caliber of the generated content. As a result, deepfakes can appear exceedingly believing, making the process of detection a very complex challenge.

To manage these challenges, researchers have begun to use the same methods and systems which are used to construct them to help detect them. In these methods, CNNs are particularly effective for analyzing visual content, while temporary anomalies are detected through LSTMs. These anomalies in the input data are analyzed based on the sequence of frames in a video. Existing methods often have a hard time with accuracy rates across multiple datasets and varying types of deepfakes. The model is aimed to address these limitations by utilizing advanced deep learning architectures, particularly EfficientNet B7, a state-of-the-art tool known for its efficiency and accuracy.

2 Literature Survey

2.1 Li, Y., and Siwei Lyu., et al. (2019). Exposing Deep Fake Videos by Detecting Face Warping Artifacts

This study employed the method of detecting artifacts by comparing the synthesized face regions and their neighboring areas with a specialized Convolutional Neural Network model. There were two-fold of Face Artifacts in this work. Their approach is premised on the fact that the existing deepfake algorithm only allows for generation of images of lower resolutions, which then need to be further processed to match the faces to be replaced within the source video. Their approach has not accounted for the temporal analysis of frames. It detects inconsistencies in face warping and spatial artifacts in the facial regions. It uses, CNNs, VGGNet-based architecture and Binary Classification.

2.1.1 Algorithms and Methods

- **Face Warping Artifact Detection:** The system focuses on spatial inconsistencies in facial regions by comparing synthesized face areas with their neighboring regions.
- **CNN-Based Binary Classifier**: A VGGNet-like architecture is used to classify video frames into real or fake.
- **Limitations**: Ineffective against high-resolution deepfakes with minimal artifacts; lacks temporal analysis capabilities.

2.2 Li, Y., et al. (2018). Exposing AI Generated Fake Face Videos by Detecting Eye Blinking

This study explains a novel approach to identifying the deepfakes by eye blinking as an important parameter resulting in classification of the videos as deepfake or genuine. Temporal analysis of cropped frames of eye blinking was done using the Long-term Recurrent Convolution Network (LRCN). As today the deepfake generation algorithms have become so

advanced that absence of eye blinking cannot be the sole indicator for detection of the deepfakes. There have to be certain other parameters must be taken into consideration for the detection of deepfakes such as teeth enchantment, wrinkles on faces, improper placement of eyebrows etc. This method works on detecting disturbances in the eye blinking patterns and lack of realism in the blinking. The algorithms used are – LSTM, CNNs and Eye state classification.

2.2.1 Algorithms and Methods

- **Temporal Analysis of Eye Blinks**: Uses Long-term Recurrent Convolutional Network (LRCN) for tracking eye blinking patterns across frames.
- **CNN + LSTM Architecture**: CNN for spatial feature extraction and LSTM for sequence modeling.
- **Limitations**: Advanced deepfake generators now simulate blinking, reducing the effectiveness of this cue alone.

2.3 Nguyen, H. H., et al. (2018). Using Capsule Networks to Detect Forged Images and Videos

This study uses a technique that employs a capsule network for detecting fake, tampered images and videos in various applications, such as replay attack detection and computer-generated video detection. In their work, they utilized random noise during the training process which is not an optimal solution. Nevertheless, the model worked well in their data but could fail on real time data as a result of noise during training. Our approach is suggested to be trained using noiseless and real time data. It employs Capsule Networks to model the spatial relationships between the facial features. Algorithms – Capsule Networks, Dynamic Routing, Binary Classification.

2.3.1 Algorithms and Methods

- **Capsule Networks (CapsNet)**: Capture orientation and pose of facial features using dynamic routing between capsules.
- **Noise-Augmented Training**: Introduces random noise to improve generalization (though this reduces real-world applicability).
- **Limitations**: High computational cost and sensitivity to occlusions or low-quality inputs.

2.4 Guera, D., and Delp, E. J., et al. (2018). Deepfake Video Detection Using Recurrent Neural Networks

This study which is used for deepfake detection employed the strategy of employing RNN for sequential processing of the frames combined with the use of ImageNet pre-trained model. Their method utilized the HOHO dataset comprising only 600 videos. Their dataset includes small set of videos and same category of videos, which might not work that well on the actual time data. We will be training our model with large amount of Realtime data. GHRCEM Wagholi, Pune, Department of Computer Engineering 2019-20205. Leverages temporal inconsistencies in between the frames of a video and capture fake patterns. Uses, CNNs, RNNs, and GRU specifically.

2.4.1 Algorithms and Methods

- **RNN-Based Temporal Modeling**: Employs a Recurrent Neural Network to learn temporal frame patterns.
- **CNN + RNN Hybrid**: CNN extracts feature from each frame, RNN captures interframe dynamics.
- **Limitations**: Limited dataset (600 videos); struggles with subtle frame transitions.

2.5 Ciftci, U. A., and Demir, I., et al. (2020). Detection of Synthetic Portrait Videos Using Biological Signals

In this study, *Ciftci et al.* extract biological signals among the facial regions on clean and deepfake video-pairs. Utilized transformations to calculate the temporal consistency and spatial coherence, record the signal properties in feature vector and PhotoPlethysmography (rPPG) maps. They also train a probabilistic Support Vector Machine and a Convolutional Neural Network. Then, the average of authenticity probabilities is used to classify whether the video is a deepfake or a pristine. Fake Catcher detects fake content with high accuracy, independent of the generator, content, resolution, and quality of the video. Due to lack of discriminator leading to the loss in their findings to preserve biological signals, formulating a differentiable loss function that follows the proposed signal processing steps is not straight forward process. Analyses remote PhotoPlethysmography (rPPG) signals. These are biological signals which only appear in real people. The algorithms utilized are – rPPG, CNNs and RNNs

2.5.1 Algorithms and Methods

- **Remote Photoplethysmography (rPPG)**: Extracts biological signals from face videos by tracking color variations.
- **CNN + rPPG + SVM**: Combines spatial analysis, temporal consistency, and statistical classification.
- **Limitations**: Sensitive to lighting/motion conditions; relies on high-resolution facial tracking.

2.6 Comparison of Algorithms in Previous/Existing Models

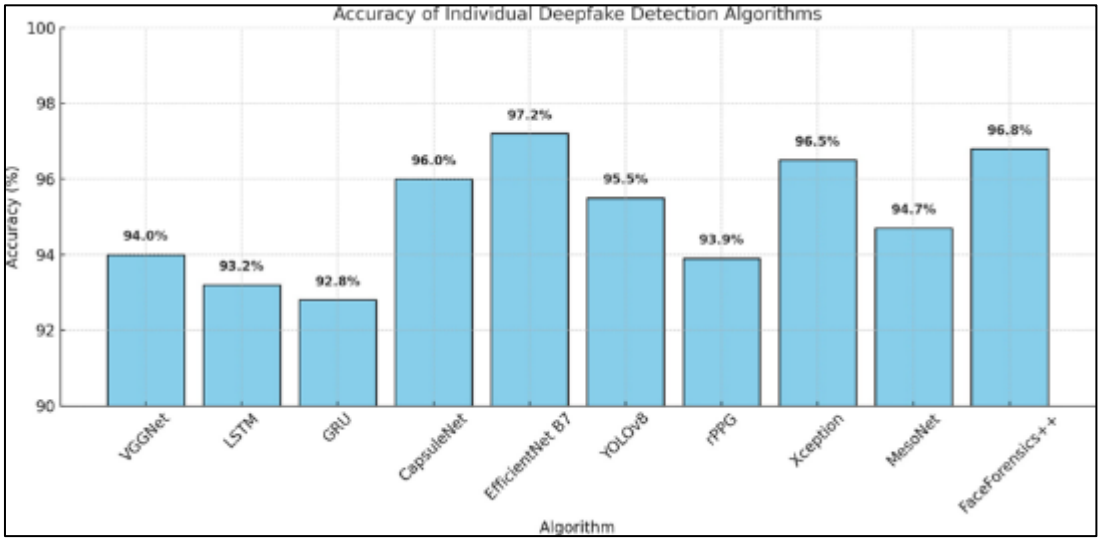

Figure 1 Accuracy of Individual Deep Fake Detection Algorithms

The above chart displays a comparison of the accuracy parameter of individual deep fake detection algorithms, used across the field of computer research. The EfficientNet b model achieves the highest accuracy at 97.2%, followed closely by FaceForensics++ (96.8%), Xception (96.5%), and CapsuleNet (96.0%). Traditional models such as VGGNet (94.0%), LSTM (93.2%), and GRU (92.8%) show lower accuracy rates, indicating their limitations in handling complex and diverse deepfake patterns. YOLOv8 and MesoNet perform moderately well with accuracies of 95.5% and 94.7%, respectively. rPPG-based detection, which leverages biological signals, achieves an accuracy of 93.9%, demonstrating its potential when used under ideal conditions. Overall, the chart highlights that advanced architectures, especially those optimized for feature efficiency and multimodal learning, yield significantly higher detection accuracies.

Table 1 Comparative Analysis of Existing Models

Paper Name	Year	Algorithm	Accuracy	Drawbacks
Exposing Deep Fake Videos by Detecting Face Warping Artifacts	2018	CNN (VGGNet)	~94%	Fails on high-quality Deep Fakes; limited to spatial artifacts
Exposing AI Generated Fake Face Videos by Detecting Eye Blinking	2018	CNN+LSTM	~95%	Ineffective if the Deep Fake model learns to replicate blinking
Using Capsule Networks to Detect Forged Images and Videos	2018	Capsule Networks (CapsNet)	~96%	High computational cost; sensitive to noise and occlusions
Deepfake Video Detection Using Recurrent Neural Networks	2019	CNN+RNN (LSTM/GRU)	~93.5%	Requires high temporal consistency; may miss subtle fake transitions
Detection of Synthetic Portrait Videos Using Biological Signals	2020	CNN+ rPPG+ Attention	~97%	Sensitive to lighting/motion; requires high-quality facial region tracking

2.7 Research Gaps

Despite recent advancements, deepfake detection systems face several persistent research gaps. One major challenge is **temporal inconsistency**, where models fail to capture smooth transitions between video frames, leading to flickering artifacts and reduced reliability. Additionally, high computational complexity—particularly in advanced models like EfficientNet B7 and diffusion architectures—limits their deployment in real-time or resource-constrained environments. Another significant issue is the lack of generalization, as models trained on specific datasets often struggle to detect new or evolving deepfake techniques due to domain shift. Moreover, most current systems suffer from poor controllability and explainability, operating as black boxes without interpretable outputs or fine-grained user control. A further limitation is the over-reliance on single-modal cues such as visual artifacts or eye-blinking, making

them vulnerable to well-crafted manipulations that exploit unseen modalities. Lastly, the absence of standardized evaluation frameworks across datasets and model types hampers consistent benchmarking, highlighting the need for unified, composite metrics that assess both perceptual realism and temporal coherence. Addressing these gaps is essential for building robust, generalizable, and interpretable deepfake detection systems.

3 Conclusion

The Deep Fake Detection System marks a significant step toward developing a reliable and efficient QR Code based attendance system. By integrating secure algorithms and a user-friendly interface, it enhances accuracy, reduces manual errors, and streamlines attendance tracking. The proposed system ensures seamless authentication and minimizes the chances of proxy attendance, making it a robust solution for academic and professional environments. Future improvements could involve incorporating biometric verification or real-time data analytics to further enhance security and usability. Overall, this system lays the foundation for modern, technology-driven attendance management.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Li, Yuezun, and Siwei Lyu. “Exposing DeepFake Videos by Detecting Face Warping Artifacts.” ArXiv.org, 22 May 2019, arxiv.org/abs/1811.00656v3.

[2] Li, Yuezun, et al. “In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.” ArXiv.org, 11 June 2018, arxiv.org/abs/1806.02877v2.

[3] Nguyen, Huy H, et al. “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos.” ArXiv (Cornell University), 26 Oct. 2018, https://doi.org/10.48550/arxiv.1810.11215.

[4] Guera, David, and Edward J. Delp. “Deepfake Video Detection Using Recurrent Neural Networks.” 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Nov. 2018, https://doi.org/10.1109/avss.2018.8639163.

[5] Ciftci, Umur Aybars, and Ilke Demir. “FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, pp. 1–1, arxiv.org/abs/1901.02212, https://doi.org/10.1109/TPAMI.2020.3009287.