Review Article

The Emerging AI Technology Deepfake

[1]Sushmita Bose, [2]Oshin Hathi, [3]Keya Pandey, [4]Udai Pratap Singh

[1]M.Sc. Forensic Science, Department of Anthropology, University of Lucknow, [2]Assistant Professor, Forensic Science, Department of Anthropology, University of Lucknow, [3]Professor, Department of Anthropology, University of Lucknow, [4]Professor and Head, Department of Anthropology, University of Lucknow

How to cite this article: Sushmita Bose, Oshin Hathi, Keya Pandey, Udai Pratap Singh. The Emerging AI Technology Deepfake. Medico-Legal Update / Vol 25 No. 3, July - September 2025

Abstract

Artificial Intelligence (AI) has given rise to a new technology called Deepfake. Deepfake Technology (DT) involves replacing one person's face with that of another in digital content, often for malicious purposes. The most prominent model that operates behind Deepfake is the Generative Adversarial Network (GAN), which operates autonomously. It creates convincingly forged content from a variety of inputs. DT raises significant concerns among legal experts and scholars and has become a worldwide phenomenon due to its easy accessibility and wide availability. The impact of DT can be seen in areas such as politics, journalism, and the legal system. Although DT has potential positive uses as well as negative ones, the detrimental consequences tend to overshadow the benefits. The rapid global spread of DT underscores the urgent need for effective detection methods. Several techniques, such as Convolutional Neural Networks (CNNs), MesoNet, face detection, and multimedia forensics, are discussed here. This paper provides a brief overview of Deepfake technology, its developing field, the associated threats, and the methods used for detection. Deepfake encompasses a broad range of applications, which are also covered here. It can offer creative advantages while simultaneously posing significant risks. The technology for detection is advancing quickly in tandem with improvements in Information and Communication Technology (ICT). The primary emphasis of this paper is on examining the different techniques available for detecting this emerging field.

Keywords: Artificial Intelligence, Deepfake, Media manipulation, Generative Neural Network, Convolutional Neural Networks, deep learning.

Introduction

With the rise and progress of Artificial Intelligence (AI), a range of new technologies have emerged. One such notable technology being, Deepfake. Deepfake, or AI faceswap, is the combination of "deep learning" with "fake" [1, 17]. It constitutes a type of synthetic media that can mislead, defame, or deceive individuals or organizations by altering images, videos, or audio suggesting events that never actually happened. This phenomenon gained significant attention in the early and mid 2017 through a social media platform known as "Reddit" [3]. On Reddit, an application called "FaceSwap" was launched to the international

Corresponding Author: Oshin Hathi, Assistant Professor, Forensic Science, Department of Anthropology, University of Lucknow

E-mail:oshinh75.oh@gmail.com

Submission: Feb 28, 2025 **Revision:** April 17, 2025 **Published date:**August 14 , 2025

audience. This app helped users create forged digital content by simply following a few steps, leading to its high level of popularity.

Deepfake is a software tool driven by machine learning that leverages AI to alter data. Being machine learning-based, it operates without the necessity of human oversight. It functions through the Generative Adversarial Network (GAN)[24], which learns autonomously [8]. Deepfake's first successful execution was seen through the app called "FakeApp", that swapped faces of individuals in videos.

Individuals are becoming victims of this method because it is readily available, accessible, and user-friendly. Even someone without any prior experience or knowledge in machine learning or AI can produce convincingly forged digital content simply by adhering to the provided instructions. Tools like FaceSwap and Reface[2] are surfacing the internet and are utilized by numerous users across the globe to create fabricated content. This method poses serious threats to significant areas of society, including politics, businesses and organizations, the legal system and courtrooms, revenge porn, cyberbullying, are a few to be mentioned[4].

This paper seeks to explore several questions: What exactly is deepfake? What are its uses? What potential dangers does it pose, and what detection methods exist? Lastly, what steps can be taken to alleviate the risks associated with deepfake?

Applications

Positive Impacts

- *In educational field*

  The Deepfake technology has the ability to generate altered videos by replacing a person's face. This feature of Deepfake can be highly beneficial in education. We can create Deepfake representations of historical figures such as freedom fighters, scientists, doctors, and others who have made significant contributions and are no longer alive[7]. One specific system, known as "LumièreNet," aims to simplify the process of producing educational videos and presentations for online learning platforms like Udacity[1].

- *In Entertainment and art generation*

  In Fast and Furious 7, following Paul Walker's untimely death in a car accident, his brother Cody Walker, who closely resembles him, stepped in to finish the last scenes of the movie. If he hadn't had a brother, the filmmakers would have probably invested heavily in CGI[7]. The use of Deepfake technology could have minimized both costs and time significantly[1].

- *In the health sector*

  The World Health Organization (WHO) has launched an AI-driven tool called "Florence," designed to assist individuals in breaking free from tobacco dependence. Users can have virtual conversations with "Florence" to bolster their resolve to quit smoking by creating a plan to track their progress[1]. Researchers are also exploring the application of Generative Adversarial Networks (GANs) to detect anomalies in X-rays and their potential to reveal early signs of diseases.

- *In fashion industry*

  This technique can be used to generate patterns and designs by combining classic designs and motifs, helping fashion designers to create innovative clothing, shoes, bags, and wallets. By applying Generative Adversarial Networks, designers can upload images of apparel and bags to produce new footwear designs or input pictures of shoes and wallets to design clothing[7]. Imagine a runway show where you can virtually "try on" outfits that models are wearing or even alter their appearances.

Negative Impacts

- *Threat to Journalism*

  The level of distrust in journalism has already reached significant heights globally, and segments of society have made little progress in discerning which news and photos to

believe. Yet, we frequently observe that both intentional and unintentional fake content, whether in writing or photographs, is circulated as if it were legitimate. There are plenty of genuine reports concerning false information[20]. While it's common to assume that content intended for humor is factual, the primary threat lies in purposefully fabricated news.

- *Threat to politics*

  The extensive accessibility of deepfakes can undermine public confidence in both politicians and the media, complicating the distinction between truth and deception. External organizations may exploit deepfakes to meddle in elections, tarnish political rivals, or stir discord within a nation[10]. A prominent example featured a manipulated video of American politician Nancy Pelosi that became widely circulated on social media. In this clip, she appeared intoxicated and had difficulty expressing her words. Despite requests from parties to remove the video, a Facebook spokesperson stated that the platform does not have policies that mandate the elimination of false information[1].

- *Threat to individuals and businesses*

  Deepfakes are being used to create non-consensual sexual content. "Twitter, Reddit, and Pornhub have all recently decided to prohibit AI-generated pornography, also referred to as 'deepfakes,' labelling it as non-consensual porn" [5,8,16]. Deepfakes have the potential to enable financial fraud. This may result in data breaches, theft of trade secrets, or intentional damage. Deepfakes can also spread false information or create chaos among employees, customers, or investors, hindering the company's operations[8]. Organizations might face legal repercussions for the dissemination of deepfakes generated by their employees or linked to their brand, even if they did not directly create the deepfake

- *Threat to the Judicial system*

  Deepfakes can be employed to produce false evidence, such as altered videos or audio recordings, which could be used to wrongfully implicate innocent people or clear the guilty. This might result in unjust convictions or dismissals. They could be utilized to fabricate false testimonies or to undermine actual witnesses by altering their image or voice. This could jeopardize the reliability of witnesses and complicate the process of uncovering the truth in legal proceedings[10]. This could result in a loss of faith in the legal process and threaten the integrity of the rule of law.

Deepfake Detection Methods

Over the last few decades, the rapid development of AI and its related aspects like deepfake are posing a threat to the mankind. Hence, deepfake detection methods are the need of the hour.

Deepfakes generally used the GAN algorithm to manipulate the digital content and make the forged look authentic. We have categorized different detection methodologies in the following ways:

1. **MACHINE LEARNING BASED METHODS**

   Conventional machine learning (ML) algorithms play a crucial role in understanding the reasoning behind decisions in a way that can be articulated in human language. These techniques are particularly well-suited for the Deepfake field due to the enhanced understanding of data and procedures[26]. Several ML-focused techniques aim to identify specific anomalies present in videos or images generated by GANs. A core method of creating Deepfakes involves altering human faces to mislead viewers. In this context, the consistency of biological indicators is assessed along both spatial and temporal dimensions, utilizing different facial landmark points (like the eyes, nose, mouth, etc.) as distinctive features to verify the authenticity of videos or images produced by GANs. Regarding the performance of machine learning-based

Deepfake detection techniques, it has been noted that these methods can reach up to 98% accuracy in identifying Deepfakes. However, the effectiveness largely depends on the dataset employed, the features chosen, and the alignment of the training and testing sets.

a. Face Detection-

Deepfake technology is designed to swap one person's face for another in a digital image or video. While the algorithm utilizes artificial intelligence to generate a convincing altered image or video, it struggles to replicate subtle nuances such as eye blinks. Research conducted by indicates that a person typically blinks every two to ten seconds, with each blink lasting approximately one-tenth to one-fourth of a second. Deepfake techniques fail to accurately reproduce faces to capture this subtle feature of human blinking, which can help identify whether a video is authentic or manipulated. The effectiveness of deepfake relies heavily on the availability of photographs and images from the internet. Consequently, an individual with limited online images will have even fewer resources depicting their eyes closed [3]. Typically, a person blinks approximately every 1 to 10 seconds, which is unlikely to occur in Deepfake videos unless images of the individual with both closed and open eyes are supplied[7]. Variations in eye color are also utilized to identify deepfakes. For this process, the hue of each eye is extracted using computer vision techniques. Following detection, all images are cropped from the facial region and resized to 768 pixels[3]. This resizing is performed to guarantee that samples undergo processing at a uniform resolution.

b. Watermarking-

It enables the straightforward identification of modified digital sources by revealing concealed markers. It helps determine if any editing has occurred. Watermarks are integrated when content undergoes changes. These marks are partially visible, so even if the material is distributed on social media or online platforms, the altered components will probably carry such markers, alerting recipients to the fraudulent content [3].

2. DEEP LEARNING-BASED METHODS

When it comes to detecting Deepfakes in images, numerous studies have utilized deep learning techniques to identify specific artifacts created by the generation process[26]. A GAN simulator was introduced to mimic the collective artifacts associated with GAN-generated images and provide them as input to a classifier for Deepfake identification. Another network was proposed to extract standard features from RGB images, while a similar but more general approach was also suggested.

a. MesoNet-

MesoNet can automatically identify facial manipulation in Deepfake videos using deep learning techniques. This approach distinguishes between computer-generated images and genuine images within Deepfake videos by employing two architectural networks, meso-4 and mesoinception-4. The primary aim of these two structures is to accurately detect facial video forgery, allowing for differentiation in image properties such as noise, accuracy, classification, and aggregation. After evaluating these image features, both meso-4 and mesoinception-4 can successfully identify Deepfake videos with an accuracy ranging from 95% to 98%[7,12].

b. Convolutional Neural Networks (CNNs)-

In contrast to human detection methods, Convolutional Neural Networks (CNNs) and similar techniques operate on machine learning principles and can effectively identify deepfake content through advanced image analysis capabilities[3]. These AI algorithms can be integrated into information-sharing platforms and social media. Researchers at SUNY Albany, Yuezun Li and Siwei Lyu, employed Convolutional Neural Networks (CNNs) to help identify face-warping artifacts.

By comparing the deepfake and source videos, CNNs assist researchers in identifying similarities and determining whether a video is a deepfake or not. The model developed by the SUNY Albany researchers was more time-efficient than others, achieving a detection rate between 84.5% and 99.1% in trials, marking a significant advancement in the fight against deepfakes.

3. STATISCAL MEASUREMENTS BASED METHOD

Calculating the various statistical metrics like-average normalized cross-correlation scores between authentic and questioned data aids in assessing the authenticity of data. Investigate the photo response non-uniformity (PRNU) for identifying Deepfakes within video frames.

a. The DFDC Dataset-

It is the largest existing Deepfake dataset and one of the few that includes footage specifically captured for machine learning purposes. Brian Dolhansky and colleagues [25] contributed the DeepFake Detection Challenge (DFDC) Dataset. The DFDC Dataset is the largest Deepfake dataset available today and ranks among the few specifically recorded datasets designed for machine learning tasks (the others include the considerably smaller Google Deepfake Detection Dataset and an earlier version of this dataset)[14]. In addition to developing and releasing a dataset, the second key contribution is the resulting analysis. Firstly, the financial incentives offered encouraged specialists in computer vision or Deepfake detection to invest time and computational power into developing models for performance evaluation. Secondly, organizing a public competition eliminates the necessity for the authors of a study to train and assess a model on their own dataset. Launching a dataset and a benchmark at the same time can introduce bias since the dataset creators have detailed knowledge of the methods applied in its creation. Lastly, collecting thousands of submissions and testing them against actual Deepfake videos that participants do not see provides an exceptionally accurate reflection of the current state of Deepfake detection technology.

b. The Celeb-DF Dataset-

In order to offer the more context-related corpus to assess and assist the future growth Deepfake detection methods, Yuezun Li et al. Suggested the Celeb-DF Dataset.Celeb-DF dataset contains 590 real and 5,639 DeepFake videos (with over 2 million video frames). With a standard frame rate of 30 frames per second, the average length of all videos is roughly 13 seconds. The videos are sourced from publicly accessible YouTube content featuring interviews with 59 celebrities who exhibit a diverse range of ages, genders, and ethnicities. Additionally, the videos showcase a wide array of variations in aspects such as the sizes (in pixels) of the subjects' faces, their orientations, backgrounds, and lighting conditions.Celeb-DF dataset contains 590 real and 5,639 DeepFake videos (with over 2 million video frames).

4. BLOCKCHAIN BASED METHOD

As a result, no established approach exists for validating onscreen post office the original of a digitized video, audio, or image. It is not feasible to establish a COA for such digital content. Thus the is a huge requirement, for a Proof of Authenticity (POA) system for online digital content, to mark the authenticity of published sources and thus be able to fight against deep fake videos, audios and images. A decentralized Proof of Authenticity (POA) using the cutting-edge technology the blockchain. As the technology can provide some of the key features, this technology can be exploited to prove the authenticity and originality of digital assets in a decentralized, highly trusted and secured way. The permission less or public blockchain is most suitable for such deepfakes. In this paper, we base our solution on the public Ethereum blockchain utilizing smart contracts to govern and track the history of transactions for digital content [9].

NOTE - The percentage of studies focusing on machine learning techniques is 18%, while those using statistical methods account for 3%. In this analysis, the proportion of research centered on the Blockchain-based approach is 2% [26].

The global community cannot afford to remain passive while measures such as detection or legislation are implemented against deepfakes; the consequences could be too severe. Instead, we can adapt to a world inundated with deepfakes by focusing on raising awareness and educating the public on this issue. More crucially, we call upon all decision-makers—including technologists, social media platform leaders, and policymakers—to assist organizations and society in preventing and mitigating the harmful side of content manipulation. With this objective in mind, Kietzmann et al. [6] suggests a R.E.A.L. framework to manage deepfake risks:

- Record original content to ensure deniability
- Expose deepfakes promptly
- Advocate for legal safeguards
- Utilize trust.

Legislation must be created and implemented to address the issue of deepfakes in legal settings, such as in evidence presentation, while digital forensics will need to adapt by applying emerging deepfake detection technologies. Techniques in multimedia or image forensics are the most effective for identifying manipulated images or video content. Image forensics analyze subtle parameters such as pixel correlation, continuity of the image, and lighting conditions. Multimedia forensics consider each stage of an image's lifecycle, including how it was saved in a compressed or alternate format, the acquisition method, or any post-processing that may leave behind a unique data trace, akin to a fingerprint[3].

Since businesses are particularly susceptible to deepfake schemes, such as impersonating owners and extortion through blackmail or smear tactics, they can implement preventative measures against deepfakes right away. Educating employees is critical, as their awareness of deceptive practices will enable them to report any suspicious activities. Companies can establish two-step verification processes, for example, by confirming phone call information through emails or requiring a second employee to authorize actions such as fund transfers[8]. Furthermore, organizations can enhance security protocols and restrict access to their images or videos, which hinders cybercriminals from utilizing this data to fabricate more convincing deepfakes. Overall, while these strategies might slightly hinder business operations due to confirmation protocols or privacy measures, they could ultimately save businesses a significant amount of money.

Limitations & Challenges

Deepfakes represent an intriguing technological innovation, yet they come with numerous considerable limitations and challenges:

A. Ethical Concerns: Deepfakes can be utilized to fabricate and disseminate false information, skew public perception, and harm reputations, which can lead to significant repercussions in fields like politics, social justice, and national security. It has the potential to generate non-consensual pornography or impersonate individuals for harmful reasons, resulting in serious breaches of privacy and psychological distress.

B. Technical Limitations: This complicates efforts to mitigate the spread of misinformation and hold creators responsible. Producing high-quality deepfakes commonly necessitates substantial amounts of high-quality source material, which may not always be readily available or easy to access.

C. Legal and Regulatory Challenges: The legal framework surrounding deepfakes is still in development, posing difficulties in prosecuting creators and holding them accountable for the misuse of this technology.It is essential for regulations to find a balance between fighting against harmful deepfakes and protecting freedom of speech and expression.

D. Social Impact: Being exposed to deepfakes can have mental health effects on individuals, especially those who are victims of malicious deepfakes. The increasing prevalence of deepfakes can foster a general atmosphere of skepticism and anxiety within society.

These limitations and challenges underscore the necessity for responsible progress and application of deepfake technology. Tackling these issues requires a comprehensive strategy involving technological improvements, ethical standards, legal measures, and public awareness.

Conclusion

This paper concludes that the ongoing evolution of cybercrime has led to the emergence of deepfakes, which significantly enhance the risks associated with traditional fraud. Deepfakes continue to present various dangers, including misinformation in politics, fraud, and tampering with evidence in legal proceedings. While current technical measures can be utilized to combat deepfake attacks, relying solely on these methods is insufficient. Thus, it is crucial to also focus on raising awareness and providing training to help recognize the early indicators of deepfake attacks. Technology companies and governments should contemplate enacting laws that criminalize the malicious use of deepfakes aimed at damaging individuals' reputations. This would ensure that malicious perpetrators face suitable penalties and repercussions. One of the most threatening applications of deepfakes involves seeking revenge by inserting an individual's face into adult films, with female celebrities frequently becoming targets. There are various methods available for detecting deepfakes through the application of different algorithms. Despite the effectiveness of deepfake technology in generating deceptive videos, its outputs appear highly realistic and believable to the human eye, making detection more challenging and optimal. Therefore, the realm of deepfakes still holds much to be examined.

Source of Funding - Self

Conflict of Interest – There are no conflict of interests.

References

1. Buo SA, Department of Computing & Informatics. The emerging threats of deepfake attacks and countermeasures. [Journal article].
2. Korshunov P, Marcel S. Vulnerability assessment and detection of Deepfake videos. Idiap Research Institute, Martigny, Switzerland. 2019. https://doi.org/10.1109/icb45273.2019.8987375
3. Albahar M, Almalki J, Umm Al-Qura University. DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW. J Theor Appl Inf Technol. 2019;97(22):3242–3243. http://www.jatit.org/volumes/Vol97No22/7Vol97No22.pdf
4. Pfefferkorn R, Center for Internet and Society, Stanford Law School. "Deepfakes" in the courtroom. Public Interest Law J. 2020;29:245–275. https://ssrn.com/abstract=4321140
5. Gosse C, Burkell J. Politics and porn: how news media characterizes problems presented by deepfakes. Crit Stud Media Commun. 2020;37(5):497–511. https://doi.org/10.1080/15295036.2020.1832697
6. Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC. Deepfakes: Trick or treat? Bus Horiz. 2020;63(2):135–146. https://doi.org/10.1016/j.bushor.2019.11.006
7. Yadav D, Salmani S. Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. 2019. https://doi.org/10.1109/iccs45141.2019.9065881
8. Muna M. Technological arming: Is deepfake the next digital weapon? UC Berkeley. 2020.
9. Hasan HR, Salah K. Combating deepfake videos using blockchain and smart contracts. IEEE Access. 2019;7:41596–41606. https://doi.org/10.1109/access.2019.2905689
10. Citron DK, Chesney R. Deep Fakes: A looming challenge for privacy, democracy, and national security. Calif Law Rev. 2019;107:1753. https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=1640&context=faculty_scholarship
11. Kwok AOJ, Koh SGM. Deepfake: a social construction of technology perspective. Curr Issues Tour. 2020;24(13):1798–1802. https://doi.org/10.1080/13683500.2020.1738357
12. Shanghai Key Lab of Intelligent Information Processing, Fudan University; School of Information Technology, Deakin University. WildDeepfake: A challenging real-world dataset for deepfake detection. Proc 28th ACM Int Conf Multimedia. 2020. https://doi.org/10.1145/3394171.3413769

13. Wang X, Huang J, Ma S, Nepal S, Xu C. DeepFake Disrupter: The detector of DeepFake is my friend. Univ Sydney, School of Computer Science. 2021.
14. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for DeepFake forensics. University at Albany, SUNY, University of Chinese Academy of Sciences. http://www.cs.albany.edu/~lsw/celeb-deepfakeforensics.html
15. Lyu S. Deepfake detection: Current challenges and next steps. arXiv [Preprint]. 2020 Mar 11. Available from: https://arxiv.org/abs/2003.09234
16. Koopman M, Macarulla Rodriguez A, Geradts Z. Detection of deepfake video manipulation. Proc 20th Irish Mach Vis Image Process Conf (IMVIP). 2018.
17. Mahmud BU, Sharmin A. Deep insights of deepfake technology: A review. Chittagong Univ Eng Technol. 2021. https://arxiv.org/pdf/2105.00192
18. Mullen M. A new reality: deepfake technology and the world around us. Mitchell Hamline Law Rev. 2022. https://open.mitchellhamline.edu/mhlr/vol48/iss1/5
19. Katarya R, Lal A. A study on combating emerging threat of deepfake weaponization. 2020 Fourth Int Conf I-SMAC (IoT in Social, Mobile, Analytics and Cloud). 2020:485–490. https://doi.org/10.1109/i-smac49090.2020.9243588
20. Temir E. Deepfake: New era in the age of disinformation & end of reliable journalism. DergiPark. 2020. https://doi.org/10.18094/josc.685338
21. Liu M, Zhang X. Deepfake technology and current legal status of it. Atlantis Highlights Comput Sci. 2023:1308–1314. https://doi.org/10.2991/978-94-6463-040-4_194
22. Dash B, Sharma P. Are ChatGPT and Deepfake Algorithms Endangering the Cybersecurity Industry? A Review. Int J Eng Appl Sci (IJEAS). 2023. https://www.researchgate.net/publication/368838115
23. Zendran M, Rusiecki A. Swapping face images with generative neural networks for deepfake technology – Experimental study. Procedia Comput Sci. 2021;192:834–843. https://doi.org/10.1016/j.procs.2021.08.086
24. Arslan F. Deepfake technology: A criminological literature review. Sakarya Üniversitesi Hukuk Fakültesi Dergisi. 2023;11(1):701–720.
25. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The DeepFake Detection Challenge (DFDC) dataset. arXiv [Preprint]. 2023. https://arxiv.org/pdf/2006.07397
26. Rana MS, Nobi MN, Murali B, Sung AH. Deepfake Detection: A Systematic Literature Review. IEEE Access. 2022. https://doi.org/10.1109/ACCESS.2022.3154404