Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 28 (2025), No. 8, pp. 3013–3023
DOI : 10.47974/JDMSC-2445

Deepfake detection : A new frontier in cybersecurity for protecting digital identities, preventing misinformation, and ensuring data integrity

Preeti Narooka [$]
School of Computer Science and Engineering
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

Dharmendra Sharma [†]
School of Technology Management and Engineering
NMIMS University
Mumbai 400056
Maharashtra
India

Miriyala Trinath Basu [§]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Hyderabad 500075
Telangana
India

Varun Kumar [‡]
Department of Pure and Applied Mathematics
Alliance School of Sciences
Alliance University
Bangalore 562106
Karnataka
India

[$] *E-mail:* preeti.narooka@jaipur.manipal.edu
[†] *E-mail:* Dharmendra.sharma@nmims.edu
[§] *E-mail:* trinath@klh.edu.in
[‡] *E-mail:* varun.kumar@alliance.edu.in

Sushanth Chandra Addimulam [@]
Kforce Inc
2681 Sidney St
Pittsburgh
PA 15203
U.S.A.

Neha Kapila [#]
Department of Electronics and Electrical Engineering
Lovely Professional University
Phagwara 144411
Punjab
India

Neeraj Kumar Verma *
Department of Data Science and Engineering
Manipal University Jaipur
Jaipur
Rajasthan
India

Abstract

Deepfake technology, powered by artificial intelligence (AI), is advancing rapidly and enabling the creation of highly realistic but manipulated audio and video content. While impressive, it presents serious challenges to information integrity, privacy, and cybersecurity. This research examines the intricate cybersecurity risks presented by deepfakes and evaluates efficient tactics to get people and organizations ready for these dangers. Semi-structured interviews using a qualitative research methodology using cy-Experts in cybersecurity and AI were consulted to obtain understanding of the present danger landscape, deepfakes' technological development, and tactics for their identification and avoidance. The findings of this study highlight how urgent it is to create reliable, AI-driven detection tools and support a well-rounded strategy that considers both the ethical implications of these innovations as well as improvements in technology. Policymakers and cybersecurity experts are advised to invest in detection technology, encourage digital literacy, and facilitate international cooperation to set ethical AI usage norms. This thesis adds to the larger conversation about cybersecurity and AI ethics by laying the groundwork for next studies and the creation of new regulations in the age of digital manipulation. The paper examines the ethical and social ramifications of deepfakes in addition to technological ones, emphasizing the significance of educating the public about the dangers they present. Public education is essential because it equips people to recognize and handle the increasing amount of distorted content. The study also emphasizes the necessity of thorough regulatory structures that can handle the complex issues raised by deepfakes. In order to prevent the detrimental usage of AI technology

[@] *E-mail:* **sushanth93@gmail.com**
[#] *E-mail:* **neha.31020@lpu.co.in**
* *E-mail:* **neeraj.verma@jaipur.manipal.edu (Corresponding Author)**

while preserving their useful applications, these frameworks must strike a balance between innovation and responsibility.

Subject Classification: *94A60, 68T10, 68U10.*

Keywords: *Deepfake, Cybersecurity, Artificial intelligence, Experts, Convolutional neural networks (CNNs).*

1. Introduction

More realistic media, such as audio, video, and images, can now be produced thanks to recent developments in AI and machine learning. Cutting-edge deep learning techniques made it possible to create "deepfakes," which are modified or synthetic media whose authenticity is difficult for the human eye to discern [1-2]. Despite being a relatively recent phenomenon initially surfaced at the end of 2017— deepfake has grown significantly [3-4]. the figure of deepfake videos on the English-speaking internet increased at a phenomenal rate of 7,964 in December 2018 to 14,678 in July 2019 to up to 85,047 as of December 2020 - a 968 percent increase in deepfake videos between 2018 and 2020. Most deepfake generation tools had over 10,000 available by 2024 [5]. This paper will examine the contemporary research ecosystem of deepfakes in a variety of ways, including performance metrics and standards and datasets [5]. Also, we give a meta-analysis of 15 well-chosen survey articles which do not only discuss these but also other general issues like performance comparisons, key challenges, and recommendations [7, 8]. Although the world has already become familiar with the term deepfake, its definition is still a subject of debate, and it is not always easy to differentiate deepfakes and non-deepfakes. To use in this research, we take a more liberal and inclusive definition of all types of manipulated or synthetic media that can be termed deep fakes. [9, 10] As they are typically applied in presentation attacks to biometric authentication systems, and their detection is overlapping with the discipline of multimedia forensics, studies in this field are also heavily involved with closely related fields like biometrics and digital forensics. The following section provides a more detailed discussion of the varying interpretations of "deepfake." [11-12]. The very first digital hoaxes referred to as Deepfake, a portmanteau of deep learning and fake media, involve objects that appear realistic but which are fake and have been fabricated or entirely altered to say or do something that did not occur. Any person can generate deepfake contents, as the sophisticated LM-based GenAI applications (such as Live Person and

ChatGPT) are openly accessible [13-14]. When these skills are democratized, the chances of abuse become extremely high. Though it is highly related to deepfakes, misinformation can be defined as any type of intentionally spread false or misleading information in order to disorient or control netizens, or active Internet users. Although the idea of misinformation is not a recent development, the launch of large model (LM)-based generative AI (GenAI) systems has increased the magnitude, complexity, and potential effects [15–16]. These LM-based GenAI technologies carry wide-ranging and complex implications. Across the globe, democratic societies are grappling with the ways in which AI-generated content (AIGC) shapes public opinion and influences electoral processes. For individuals, the risks are unprecedented: deepfakes generated from publicly available personal data can act on their behalf without their knowledge or consent, raising profound concerns for privacy and security [17]. At the same time, the media landscape—long regarded as the primary channel for distributing factual information—is undergoing a dramatic transformation. Journalists and content creators now face the critical challenge of redefining what constitutes credible and trustworthy information in the post-deepfake era. The term "deepfake" itself combines "deep" (from deep learning, DL) and "fake," and is commonly used to describe synthetic media created using DL-based methods or the manipulation of existing media (audio, video, or images). Common examples include fabricated facial images, synthetic voice forgeries, and composite videos that merge both manipulated visuals and audio. Despite the negative connotation of the word "fake," deepfake technologies also have benign or even beneficial applications, such as in creative arts, entertainment, and accessibility. At

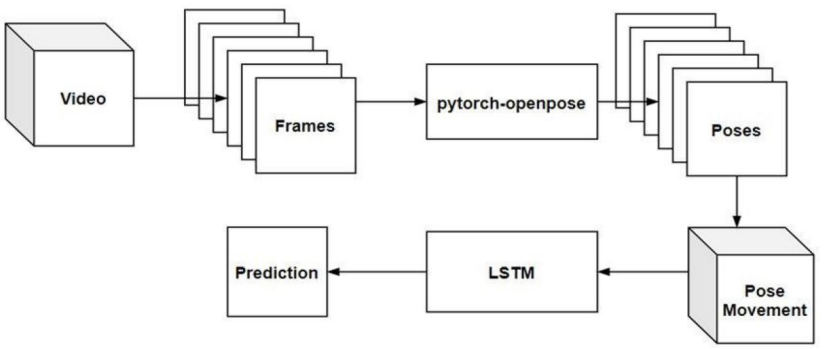

Figure 1

Deepfake detection system.

minimum, two dimensions must be considered: the generation of deepfakes and their detection. Deepfake generation primarily relies on deep learning models, particularly Generative Adversarial Networks (GANs), which employ two neural networks—the generator and the discriminator—that operate in opposition to refine and enhance synthetic content production. This adversarial training continues iteratively until the generator produces media indistinguishable from authentic content [2].

Deepfakes undermine confidence in online information by contesting the legitimacy of digital media. Deepfakes can be used for impersonation in social engineering assaults by imitating a person's voice and look shown in figure 1. By manipulating prominent figures or executives, fake videos might result in financial harm.

2. The Emergence and Proliferation of Large-Scale AI Models

The third decade of the twenty-first century is largely considered to mark a watershed in the history of artificial intelligence, as a result of LM-based generative AI (GenAI) [3]. The impressive ability of foundation models and LM-based GenAI systems including large language models (LLMs), large vision models (LVMs), large audio models (LAMs), and large multimodal models (LMMs) to perceive and imagine human text, images, and audio has led to the rapid advancement of AI-generated

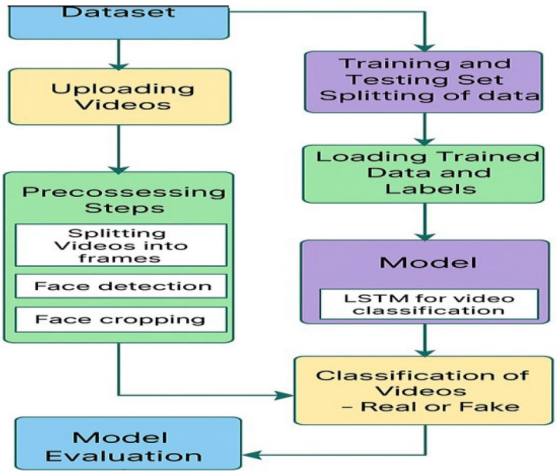

Figure 2

Deep Fake generation and detection

content (AIGC). Here, the development of LM-based GenAI models and their functions are described, as well as the risks of using them. These models are trained as shown in Figure 2 to accommodate the finer aspects of auditory patterns, visual indications and human language. Through context learning, they are able to repeat the style and structure of their training data allowing them to generalize the next word of a sentence or the next pixel of an image or the next waveform of an audio sequence. While their potential is remarkable, real-world case studies also highlight instances of misuse that underscore the technology's risks. Furthermore, the substantial computational demands of training large models pose significant environmental challenges, while excessive reliance on AI may gradually erode human expertise and critical skills.

3. Techniques for Deepfake Detection

Deepfakes undermine confidence in online information by contesting the legitimacy of digital media. Advances in AI, particularly in Generative Adversarial Networks (GANs), have led to a significant expansion in the ability to create deepfakes. Deepfake detection has emerged as a critical area within cybersecurity due to the growing misuse of artificial intelligence and machine learning to create highly realistic synthetic media, such as videos, images, or audio. The increase in deepfake content poses significant threats, from disinformation campaigns to identity theft and financial fraud. Researchers have observed a rapid escalation in deepfake technology's sophistication, which challenges traditional detection methods. Real-world cases where deepfakes have been employed in political manipulation, social engineering, and financial scams underscore the urgency of developing robust detection systems [6].

4. Proposed Algorithm and Results

Six cybersecurity and AI specialists participated in semi- structured interviews that provided important new information about the rapidly changing field of deepfake technology and its implications for cybersecurity. All experts agreed that deepfakes have rapidly evolved from specialized, unreachable tools to very sophisticated, widely accessible technology, greatly raising the risks involved.

Algorithm: Deepfake Detection Using AI & ML

Input: Digital media (image or video)

Output: Classification as "Deepfake" or "Authentic"

Step 1: Data Preprocessing. Input Acquisition, Accept an image or video frame sequence. Frame Extraction (for videos), Extract frames at regular intervals. Normalization & Resizing, resize frames to a fixed dimension, Apply histogram equalization and noise reduction.

Step 2: Feature Extraction: Facial Landmark Detection, Use Dlib or OpenCV to extract facial key points. Frequency Analysis use Discrete Fourier Transform (DFT) to detect unnatural pixel distributions. Blink Rate & Lip Sync Analysis. Track eye blinks and mouth movements for natural behavior patterns. Texture Analysis use Local Binary Patterns (LBP) to detect synthetic artifacts.

Step 3: Model-Based Classification: Feature Encoding, Convert extracted features into vectors. Deep Learning Model (CNN/RNN/Transformer-based), Use a pre-trained deepfake detection model like XceptionNet, EfficientNet, or Vision Transformer (ViT). Prediction & Confidence Score and model outputs a probability score.

Step 4: Post-Processing & Decision Making : Aggregation of Frame-Wise Results. Use majority voting or temporal consistency analysis. Threshold-Based Classification : If the average probability exceeds a predefined threshold (e.g., 70%), classify as Deepfake; otherwise, Authentic.

Step 5: Output & Report Generation: Display Results, Show classification results with confidence scores. Log and Save Metadata, Store analysis results for forensic purposes.

Final Output: "Deepfake" (if probability > threshold), "Authentic" (if probability ≤ threshold).

They emphasized the serious cybersecurity risks, such as financial fraud, identity theft, and disinformation campaigns, and the growing challenge of identifying increasingly complex deepfakes described in figure 3. An "arms race" between deepfake generation and detection technologies was revealed by the study, and experts emphasized the urgent need for significant investment in AI and machine learning capabilities to create reliable detection methods.

Lightweight detection systems that can function in real-time without compromising accuracy should be the main focus of future research. Real-

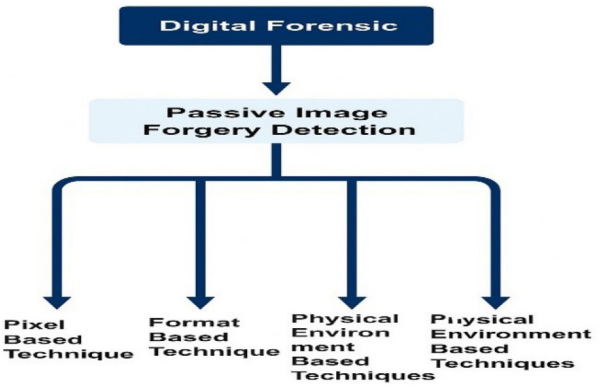

Figure 3

Overview of image forgery detection.

time deepfake detection systems might be developed by refining network designs or investigating novel machine learning paradigms (like quantum computing) [11].

Multi-modal detection, which combines audio, video, and other information, can improve detection accuracy overall. The comparison of visual and aural information, known as cross-modal discrepancies, has demonstrated potential for increasing detection rates shown in figure 4.

The efficiency of the suggested integrated framework depends on how well its elements interact together in its critical to identify deepfakes quickly. However, detection techniques might need to advance in specialisation as deepfake production technology advances, which could result in an arms race between detection and creation skills. Cross-platform

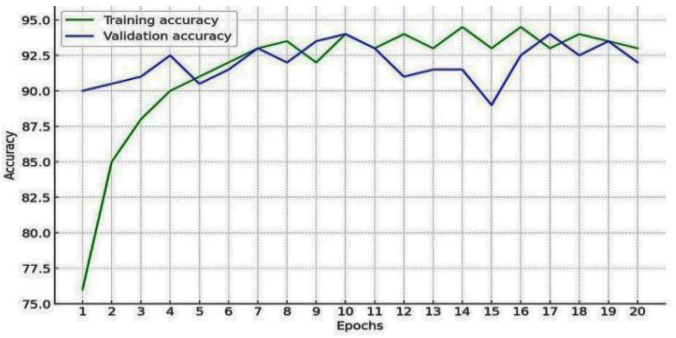

Figure 4

Training and validation accuracy.

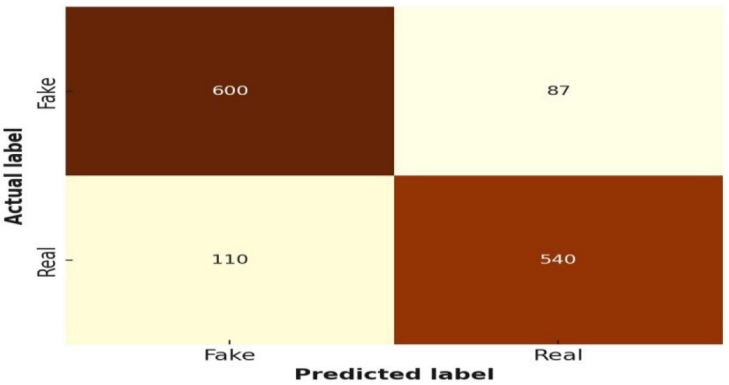

Figure 5

Deepfake Detection Matrix.

tactics highlight the necessity of a concerted effort to combat false information. Initiatives of this kind must be scalable and flexible enough to quickly adjust to emerging disinformation tactics.

The willingness of the international community to accept and apply harmonized standards as well as the enforcement of policy measures will be key factors in determining their efficacy shown in figure 5. The practicality of educational programs might be one of the most sustained protection mechanisms against misinformation, to enhance the ability of the population to differentiate between relevant and unreliable information.

5. Conclusion and future work

The research study underscores the importance of addressing the escalating threat of deepfakes through a comprehensive, multimodal approach that includes public awareness initiatives, legal regulations, and ongoing technological advancements. It emphasizes the critical role of continuous innovation in AI-driven detection systems to keep pace with the rapid development of deepfake technology. Effective identification and mitigation of deepfake risks rely on sophisticated tools and methods. The study also highlights the necessity of establishing robust legal frameworks to regulate the creation, use, and dissemination of deepfakes, ensuring a balance between protecting individual rights and preventing societal harm. Public awareness is equally essential, as educating people about deepfakes can empower them to identify and counter false

information. Such guidelines should find some kind of harmony between defending individual privileges, like opportunity of articulation, and forestalling cultural mischief brought about by the spread of deception and pernicious deepfakes. Public mindfulness assumes a similarly essential part in fighting the danger, as teaching people about the presence and risks of deepfakes can engage them to evaluate and distinguish misleading or controlled content basically.

References

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," arXiv preprint arXiv:2303.04226 (2023).

[2] W. Shin and M. O. Lwin, "Parental mediation of children's digital media use in high digital penetration countries: Perspectives from Singapore and Australia," *Asian J. Commun.*, vol. 32, no. 4, pp. 309–326 (2022).

[3] A. Tiwari and R. M. Sharma, "Potent cloud services utilization with efficient revised rough set optimization service parameters," in *Proc. Int. Conf. on Advances in Info. Comm. Tech. & Computing*, pp. 1–7 (Aug. 2016).

[4] N. Amezaga and J. Hajek, "Availability of voice deepfake technology and its impact for good and evil," in *Proc. 23rd Annu. Conf. on Info. Tech. Education*, pp. 23–28 (2022).

[5] A. Tiwari and R. M. Sharma, "A Skywatch on the Challenging Gradual Progression of Scheduling in Cloud Computing," in *Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proc. of MARC 2018*, Singapore: Springer Singapore, pp. 531–541 (2019).

[6] D. Silverman, K. Kaltenthaler, and M. Dagher, "Seeing is disbelieving: The depths and limits of factual misinformation in war," *Int. Stud. Q.*, vol. 65, no. 3, pp. 798–810 (2021).

[7] A. Tiwari and R. Garg, "Reservation System for Cloud Computing Resources (RSCC): Immediate Reservation of the Computing Mechanism," *Int. J. Cloud Appl. Comput. (IJCAC)*, vol. 12, no. 1, pp. 1–22 (2022).

[8] M. T. Ahvanooey, Q. Li, X. Zhu, M. Alazab, and J. Zhang, "ANiTW: A novel intelligent text watermarking technique for forensic identifica-

tion of spurious information on social media," *Computers & Security*, vol. 90, p. 101702 (2020).

[9] A. Tiwari and R. M. Sharma, "Realm Towards Service Optimization in Fog Computing," in *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing*, IGI Global, pp. 1530–1563 (2021).

[10] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions," in *Proc. 2023 CHI Conf. on Human Factors in Computing Systems*, pp. 1–20 (2023).

[11] A. Tiwari and R. Garg, "Adaptive ontology-based IoT resource provisioning in computing systems," *Int. J. Semantic Web Inf. Syst. (IJSWIS)*, vol. 18, no. 1, pp. 1–18 (2022).

[12] J. Fletcher, "Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance," *Theatre J.*, vol. 70, no. 4, pp. 455–471 (2018).

[13] D. Dudeja, S. M. Sabharwal, Y. Ganganwar, M. Singhal, N. Goyal, and A. Tiwari, "Sales-Based Models for Resource Management and Scheduling in Artificial Intelligence Systems," *Eng. Proc.*, vol. 59, no. 1, p. 43 (2023).

[14] K. Chakraborty, S. Singhal, N. Arya, M. Tiwari, A. Khatoon, and A. Tiwari, "Multilevel Cloud Resource Scheduling Approach Based on Image-Based Signature System," in *Proc. 5th Int. Conf. on Info. Management & Machine Intelligence*, pp. 1–7 (Nov. 2023).

[15] S. Kumar, P. K. Srivastava, G. K. Srivastava, P. Singhal, D. Singh, and D. Goyal, "Chaos based image encryption security in cloud computing," *J. Discrete Math. Sci. Cryptogr.*, vol. 25, no. 4, pp. 1041–1051 (2022).

[16] K. Sarvesh, D. K. Kumar, G. A. Kumar, V. Shikha, K. Vinay, and M. Udit, "Detection of recurring vulnerabilities in computing services," *J. Discrete Math. Sci. Cryptogr.*, Art. no. 2072432 (2022).

[17] S. Kumar, P. K. Srivastava, A. K. Pal, V. P. Mishra, P. Singhal, G. K. Srivastava, and U. Mamodiya, "Protecting location privacy in cloud services," *J. Discrete Math. Sci. Cryptogr.*, vol. 25, no. 4, pp. 1053–1062 (2022).

Received February, 2025