Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2024, 11(10):70-74

Research Article ISSN: 2394 - 658X

Exploring The Use of Generative AI in Creating Deepfake Content and The Risks it Poses to Data Integrity, Digital Identities, and Security Systems

Archana Todupunuri

Fidelity Information Services, USA

ABSTRACT

Deepfakes are fake audio, video and image contents that are responsible for fraud activities and result in financial loss and reputation damage also. The fraud activities can be preventative by the use of advanced security systems equipped with blockchain and advanced AI algorithms. Deepfake contents are used for entertainment purposes also and it helps learners by offering educational videos also. The regular security audits and technological advancements can be helpful in preventing the fraud attack made by deepfake contents created by generative AI.

Keywords: Deepfake, generative AI, phishing, blockchain

INTRODUCTION

Background

The deepfake content refers to realistic-looking fake videos, images, audio recordings created with the help of artificial intelligence using machine learning algorithms. Deepfakes are often used to spread propaganda or misinformation and they seem to be dangerous as they make it hard to distinguish the real one [1]. Accordingly, deepfakes are growing threat for banking customers and employees as well. The number of fraud attempts increased by 4500 percent yearly in the Philippines followed by the United States, Vietnam and Belgium [2]. It indicates that it has become easy to trick customers and banking staff to commit fraud with the use of deepfakes.

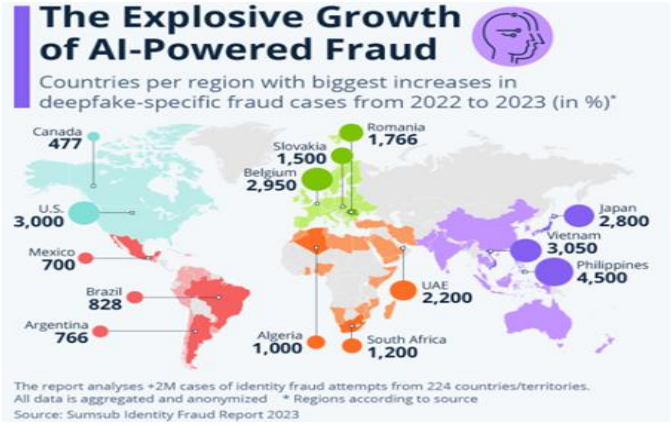

Figure 1: *Explosive growth of deepfake contents in global aspect*

Deepfakes are used by the business firms for marketing and advertising purposes also [3]. In addition to that, deepfake technology also comes with new possibilities in the aspect of filmmaking as it helps in affordable video creation and also in crafting interactive photos of the late celebrities and artists. In another direction, individuals use computer resources and communication devices with harmful intent to impersonate or cheat someone and it can result in three years of imprisonment [4]. These technologies are misapplied for creation of explicit content and spreading false information in recent times.

Deepfake contents are frequently used for malicious purposes such as impersonating the videos and voices of business professionals and deploying them in deceptive contexts [5]. It is quite difficult for the individuals to understand what is real in this context. The generative adversarial networks offer a powerful class related to deep neural networks that are increasingly used for making deepfake content like counterfeit videos and images [5]. The prevalence related to altered videos and images underscores the significance of suitable detection techniques for determining counterfeit and genuine content. This research study explores the significant methods for identifying specific architecture in the aspect of creation of deepfake images.

Research aim

The aim of this research study is to explore the utilisation of Generative AI in creating deepfake content and evaluating the risk related to data integrity, digital identities and security concerns.

Research objectives

- To evaluate the capabilities along with limitations of Generative AI in creating deepfake content
- To examine the impact of deepfake on digital identities, data integrity and security systems
- To analyse the risks of compromising data integrity, digital identities and security systems in the aspect of deepfake contents created by generative AI
- To recommend suitable guidelines for mitigating the challenges and risks related to creation of deepfake contents by generative AI

Research questions

RQ1: How do the generative AI models help in creating deepfake contents for the business organisations?

RQ2: What are the risks associated with deepfake content in the context of data integrity, digital identities and security systems also?

RQ3: What kinds of vulnerabilities are exposed by deepfake contents in the aspect of security systems?

RQ4: What are mitigation strategies to mitigate the challenges associated with deepfake contents created by generative AI?

Research scope

The scope of this study is to explore the application of generative AI in the context of creation of deepfake contents and the impact of counterfeit images and videos on individual perspectives. In addition to that, the risks related to deepfake contents are explored with proper information. The exact scope of this study lies in exploring the impact of deepfake contents on business operations and social life and to gain knowledge about risks related to digital identity, data integration and security concerns also. The suitable strategies are discussed to mitigate the challenges associated with deepfake content created by generative AI models.

LITERATURE REVIEW

Altering videos and images with human emotions in the context of deepfake content

Modification of static images is quite simpler than working with videos or moving images in the aspect of deepfake contents. The deepfake videos are typically created with the help of publicly available datasets as human faces often appeared without any meaningful expressions having resemblance with lifeless puppets [6]. The machine learning algorithms are utilised to simulate several human actions like speaking, sobbing, walking and grinning. The advanced generative AI models utilise fusion methods to verify the alterations by contrasting distinct facial regions rather than analysing the whole image [7]. Different kinds of attributes such as facial expression, movements of eyes and hairs are used as random markers to evaluate the changes between real and fake characters [7]. The enhanced algorithms are dedicated towards closely monitoring of the aforesaid regions for proper detection and results in real like image and video content.

Impact of deepfake contents on digital identities, data integrity and security systems

Deepfake poses crucial threats towards digital identities as it helps individuals in identity theft and impersonation also and it results in enhancing the number of criminal activities. Apart from that, deepfakes are also utilised for reputation damage and manipulation of human characters by creating fake images and videos [8]. It basically increases the risk related to social and phishing engineering. The tampering with video, audio as well as image evidence is also considered as threats in the aspect of data integrity. The manipulation of public documents and records along with spreading of fake news are responsible for compromise with sensitive information [8]. Face swapping, voice cloning, manipulation of audio and video contents along with fake image generation are followed as different kinds of deepfake attacks in recent scenarios.

Vulnerabilities exposed by deepfake contents in the aspect of security systems

The deepfakes are enabled by bypassing biometric authentication such as facial recognition. The fake image of an individual has been utilised in this context to match passwords for different locks [9]. Apart from that, manipulation of surveillance footage such as insertion of fake images and videos can cause huge harm towards security systems. Disruption of critical infrastructure operations can be responsible for significant loss for the business firms in this aspect [9]. Tampering with CCTV recordings leads towards compromised access control systems and it results in significant gaps in the aspect of security concerns for the business organisations.

Mitigation strategies towards issues associated with deepfake content

Implementation of AI-powered detection tools can help in distinguishing the real and fake images in this context. The advanced machine learning algorithms are helpful in detecting fraudulent activities through developing the digital watermarking techniques [10]. In addition to that, the use of blockchain techniques for two factor authentication methods can help the business organisation as well as individuals in preventing the fraud activities [11]. Apart from that, improvement in biometric authentication methods can help in identifying the phishing activities in this context. The implementation of advanced threat detection and regular security audits along with testing processes can help the business firms and individuals in mitigating the challenges related to deepfake contents.

Theoretical perspective

Information poverty theory refers to the sociological theory that describes the lack of access towards information as well as inability of the individuals in interpreting and using this in unequal and disadvantaged situations [12]. Information poverty entails the gap of technological skill among the individuals in the society such as they do not have enough resources, skills to understand and utilise the information. In this regard, the individuals have no such advanced technology or technological knowledge so that they can distinguish the real and fake digital contents [13]. The deepfake contents often influenced the ignorant buyers by showing them their favourite actors speaking about products or services. In addition to that, deepfake contents are also utilised for damaging the background of any individuals by disseminating fake news about them. Hence, it can be stated that deepfake contents helps in increasing fraudulent activities in the society with the use of technological knowledge gap among the communities.

METHODOLOGY

The Methodology chapter describes the research procedure for designing methods that are reliable and correct for fulfilling research objectives. In this regard, research philosophy, strategy, approach, data collection along with analysis processes were maintained to design methods for this research study. Research philosophy helps in understanding the belief regarding data collection about a phenomenon. “Interpretivism philosophy”was maintained in this research study to obtain necessary information regarding deepfake contents and its impact on business operations. “Interpretivism philosophy” focuses on human factors for creating in depth knowledge regarding subject matter [14]. In addition to this, “Inductive approach” was used to interpret the significant consequences related to deepfake content and its impact on business operations. Researchers accumulate empirical data to enhance suitable concepts on the basis of previously collected data [15]. Empirical data was analysed in this research study to analyse the impacts of deepfake contents on its benefits in marketing context with the help of “Inductive approach” by examining necessary information on filmmaking, marketing and advertising. Hence, “Inductive approach” along with “interpretivism philosophy” were maintained to obtain the perception of individuals on utilisation of deepfake content in the marketing and advertising. “Qualitative strategy” has been maintained to integrate in-depth exploration of every aspect like individual perspectives after observing deepfake contents, difficulties in distinguishing real scenarios and fraudulent activities through counterfeit videos and images. In addition to that, “secondary data collection” method was maintained due to incorporating cost effectiveness while accumulating relevant data that addresses research questions. Secondary data can be obtained in a cheaper way in comparison with primary data [16]. “Google Scholar” has been used as a secondary database for obtaining necessary information from authentic journal articles. Apart from that, thematic analysis was integrated to interpret meaningful insights in the form of themes by examining accumulated information related to deepfake contents created by Generative AI. Themes were established on the basis of research objectives through identifying dataset patterns. Therefore, the study integrated “secondary data collection” and “Thematic analysis” methods for fulfilling research requirements.

IMPACT AND IMPLICATION

Theme 1: Consequences of deepfake attacks

The deepfake contents are used by the attackers in phishing activities that lead towards financial loss and damage [17]. The fake voice and audio contents are observed as a national security threat in this context. The damage in reputation and loss of trust for an individual can be possible through spreading fake contents in the social media platform by creating deepfakes with the help of generative AI. It is basically used in the time of political election to damage the reputation of the election candidates [17]. In addition to that, the violations of privacy as well as data breaches also refers to financial and reputation loss. It is responsible for massive social unrest and instability also.

Theme 2: Impact of deepfake contents on social life

The deepfake contents have a positive impact on the society as it helps in entertainment and creativity by offering significant contents for the audiences. The educational content along with simulations are helpful for the learners. In another direction, it helps in journalism by enhancing the concept of storytelling. Hence, it can be stated that deepfake contents created by generative AI have the potential for social changes and activism also [18]. The negative impact of deepfake content is observed in the aspect of spreading disinformation and misinformation

regarding a particular subject matter. The erosion of trust in media as well as in institutions results in depression and bad mental health of the target individual. It is also responsible for social division. The increased stress and anxiety lead towards degraded social life in this context. The normalisation of harmful behaviours is considered as a significant concern for social life in this context.

Theme 3: Risks associated with network and system vulnerabilities

The phishing attacks are quite common in the aspect of deepfake contents created by generative AI as it is responsible for security breaches in the reputed business organisations and results in huge financial loss also. The dissemination of malware and ransomware activities are significant concerns in the context of cyber securities. The compromise with the security system basically results in loss of valuable data for the business organisations [19]. The Distributed Denial of Service (DDoS) attacks are the instances of such deepfake contents, while the system algorithm fails to detect the anomalies and results in significant security breaches in this scenario. Hence, it can be stated that deepfakes are responsible for network as well as system vulnerabilities and it leads to degradation of security infrastructure of business organisations and public sector organisations also.

Theme 4: Mitigation procedures to resolve the issues related to deepfakes

The education and awareness campaigns can be helpful in spreading awareness about deepfake contents among individuals. The business sectors and public sector organisations need to be encouraged about implementation of responsible AI setup to detect the anomalies in the early stage. Advanced detection and removal of deepfake contents can be useful in this regard and development of AI-powered fact-checking systems is quite significant in this scenario [20]. The blockchain-based authentication is highly recommended in this aspect to prevent such fraud activities. Investing in interdisciplinary research and collaboration activities can help in mitigating the deepfakes related issues in upcoming days.

CONCLUSION

It has been observed that use of deepfake contents increased drastically in the digital era and it is used for creating fake videos, images and audios for marketing and advertising purposes across the global market. The deepfakes are very difficult to distinguish from real content as they are quite similar with real ones and various human patterns such as speaking, walking and watching patterns are copied properly to make it like real one. Deepfakes are also used in fraud activities such as bypassing of security passwords using facial recognition and synthesis of text allows the attackers to proceed with criminal activities. In addition, the deepfake contents are also used for damaging images of the individuals by spreading fake contents about individual images and videos. Apart from that, the business organisations are affected by deepfake contents as the consumers are influenced by fake videos and their buying behaviour is affected by this. The regular security audits and implementation of advanced technologies can help in preventing the fraud activities made by deepfake contents. The advanced AI-detection mechanisms are helpful enough in detecting the anomalies in this context. Regular checking of CCTV footage can be helpful in detecting phishing activities in this context. The implementation of blockchain applications is helpful in this context as blockchain algorithms are enabled to detect the phishing activities in this scenario.

FUTURE DIRECTION

The future research would focus on specific areas of deepfake contents created by generative AI to enhance information quality in certain contexts like the way individuals or business organisations compromise with business operations because of fraud video, audio and images. For example, future research could be explored on the impact of deepfake content created by generative AI in response towards filmmaking, advertising accuracy. In another direction, the researchers can emphasise on personalised advertising strategies for leveraging deepfake contents to minimise adverse effects on business firms. The importance of exploring deepfake contents within developing countries could be conducted to obtain knowledge about both positive and negative aspects of deepfake contents. The future direction of this study would be focused on developing advanced strategies to distinguish the real and fake contents with ease.

REFERENCES

[1]. M. E. Myers, “Propaganda, Fake News, and Deepfaking,” Understanding Media Psychology, pp. 161–181, Sep. 2021, doi: https://doi.org/10.4324/9781003055648-8.

[2]. F. Zandt, “Infographic: How Dangerous are Deepfakes and Other AI-Powered Fraud?,” Statista Daily Data, Mar. 13, 2024. https://www.statista.com/chart/31901/countries-per-region-with-biggest-increases-in-deepfake-specific-fraud-cases/

[3]. L. Whittaker, K. Letheren, and R. Mulcahy, “The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing,” Australasian Marketing Journal, vol. 29, no. 3, p. 183933492199947, Mar. 2021, doi: https://doi.org/10.1177/1839334921999479.

[4]. W. Huang, W. Tang, H. Jiang, J. Luo, and Y. Zhang, "Stop Deceiving! An effective Defense Scheme against Voice Impersonation Attacks on Smart Devices," IEEE Internet of Things Journal, pp. 1–1, 2021, doi: https://doi.org/10.1109/jiot.2021.3110588.
[5]. S. Alanazi and S. Asif, "Exploring deepfake technology: creation, consequences and countermeasures," Human-Intelligent Systems Integration, Sep. 2024, doi: https://doi.org/10.1007/s42454-024-00054-8.
[6]. M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," Proceedings of the National Academy of Sciences, vol. 119, no. 1, Dec. 2021, doi: https://doi.org/10.1073/pnas.2110013119.
[7]. R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez, "DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation," Engineering Applications of Artificial Intelligence, vol. 110, p. 104673, Apr. 2022, doi: https://doi.org/10.1016/j.engappai.2022.104673.
[8]. L. K. Seng, N. Mamat, H. Abas, and W. N. H. W. Ali, "AI Integrity Solutions for Deepfake Identification and Prevention," Open International Journal of Informatics, vol. 12, no. 1, pp. 35–46, Jun. 2024, doi: https://doi.org/10.11113/oiji2024.12n1.297.
[9]. E. Nowroozi, S. Seyedshoari, M. Mohammadi, and AlirezaJolfaei, "Impact of Media Forensics and Deepfake in Society," Springer eBooks, pp. 387–410, Jan. 2022, doi: https://doi.org/10.1007/978-3-031-10706-1_18.
[10]. N. Choi and H. Kim, "DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in Blockchain Environment," Applied Sciences, vol. 13, no. 4, p. 2122, Feb. 2023, doi: https://doi.org/10.3390/app13042122.
[11]. ÁngelFernándezGambín, AnisYazidi, A. Vasilakos, H. Haugerud, and YoucefDjenouri, "Deepfakes: current and future trends," Artificial Intelligence Review, vol. 57, no. 3, Feb. 2024, doi: https://doi.org/10.1007/s10462-023-10679-x.
[12]. W. Matli and M. Ngoepe, "Extending information poverty theory to better understand the digital access and inequalities among young people who are not in education, employment or training in South Africa," Higher Education, Skills and Work-Based Learning, vol. ahead-of-print, no. ahead-of-print, Sep. 2021, doi: https://doi.org/10.1108/heswbl-05-2020-0107.
[13]. W. Matli, "Extending the theory of information poverty to deepfake technology," International Journal of Information Management Data Insights, vol. 4, no. 2, pp. 100286–100286, Sep. 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100286.
[14]. Alharahsheh, H.H. and Pius, A., (2020). A review of key paradigms: Positivism VS interpretivism. Global Academic Journal of Humanities and Social Sciences, 2(3), pp.39-43.
[15]. Okoli, C., (2023). Inductive, abductive and deductive theorising. International Journal of Management Concepts and Philosophy, 16(3), pp.302-316.
[16]. Taherdoost, H., (2021). Data collection methods and tools for research; a step-by-step guide to choose data collection technique for academic and business research projects. International Journal of Academic Research in Management (IJARM), 10(1), pp.10-38.
[17]. Ajegbile, N.M.D., Aderonke, J., Cosmos, C., Tamunobarafiri, G. and Abdul, N.S. (2024). Integrating business analytics in healthcare: Enhancing patient outcomes through data-driven decision making. World Journal of Biology Pharmacy and Health Sciences, 19(1), pp.243–250.
[18]. Audrey de Rancourt-Raymond and N. Smaili, "The unethical use of deepfakes," Journal of Financial Crime, vol. 30, no. 4, pp. 1066–1077, May 2022, doi: https://doi.org/10.1108/jfc-04-2022-0090.
[19]. J. T. Hancock and J. N. Bailenson, "The Social Impact of Deepfakes," Cyberpsychology, Behavior, and Social Networking, vol. 24, no. 3, pp. 149–152, Mar. 2021, doi: https://doi.org/10.1089/cyber.2021.29208.jth.
[20]. T. Hwung and M. Zolkipli, "Hacking Techniques and Future Trend: Social Engineering (Phishing) and Network Attacks (DOS/DDOS)," International Journal of Advances in Engineering and Management (IJAEM), vol. 5, p. 434, 2023, doi: https://doi.org/10.35629/5252-0507434444.
[21]. M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, Mitigations, and Opportunities," Journal of Business Research, vol. 154, Jan. 2023, doi: https://doi.org/10.1016/j.jbusres.2022.113368.