Borneo International Journal eISSN 2636-9826; Vol.8 (4); 2025; 56-63
Published by *Majmuah Enterprise*
www.majmuah.com

A Comprehensive Literature Review of Deep Fake Technology and Digital Identity Fraud on Social Media

MUHD AMIRUL NAJHAN SHAMSUDIN, MOHAMAD FADLI ZOLKIPLI
School of Computing, Awang Had Salleh Graduate School, College of Arts and Science, Universiti Utara Malaysia (UUM), 06010 Sintok, Kedah, MALAYSIA
Email : amirul_najhan@ahsgs.uum.edu.my, m.fadli.zolkipli@uum.edu.my

Received: December 20, 2025
Accepted: December 23, 2025
Online Published: December 26, 2025

Abstract

The rapid evolution of Generative Artificial Intelligence (AI) and specifically Generative Adversarial Networks (GANs) has eroded trust in digital media significantly. These technologies allow for the creation of hyper-realistic "deep fakes" including face swaps and voice cloning that pose a severe threat to digital identity verification. When weaponized on social media platforms these synthetic media become potent tools for sophisticated identity fraud, social engineering and biometric spoofing. This paper conducts a comprehensive literature review to analyze the mechanics of these threats and evaluate the efficacy of current forensic detection methodologies. The analysis reveals that existing reactive defenses suffer from critical limitations in generalizability and robustness. Standard Convolutional Neural Networks (CNNs) often fail against the compressed video formats common on social networks. Relying solely on technical detection is proving insufficient as the "arms race" between generation and detection accelerates. Consequently this study concludes that safeguarding digital ecosystems requires a holistic defense strategy. This includes the development of multi-modal detection models, the enforcement of strict legislative reforms criminalizing non-consensual deep fakes and the adoption of proactive content provenance standards.
Keywords: Deep Fakes; Digital Identity; Social Media; Generative AI; Cyber Fraud;

1.0 Introduction

The integrity of digital identity and communication channels faces an unprecedented systemic challenge rooted in the ballistic progression of generative Artificial Intelligence (AI) technologies. This is particularly true for deepfakes (Farid, 2025; Naitali et al., 2023). This phenomenon represents the seamless fabrication of hyper-realistic digital media including images, video and synthesized audio that are increasingly indistinguishable from authentic content to the average human observer (Alanazi et al., 2024; Heidari et al., 2024). The technical foundation of this capability is frequently rooted in sophisticated deep learning (DL) models. Notably these are Generative Adversarial Networks (GANs) which were initially introduced in 2014 (Acim et al., 2025). GANs employ a dual computational architecture. A generator strives to produce synthetic data that perfects mimics real data while a discriminator attempts to identify the forgery (Patel et al., 2023; Duhan & Kajal, 2025). This adversarial process ultimately yields synthetic output of such fidelity that it fundamentally compromises the veracity of the visual and auditory record (Patel et al., 2023; Dhanyalakshmi et al., 2025). What makes this genuinely worrying is not just the technology itself but its accessibility. The severity of this threat is magnified by the democratization of deepfake creation tools. These have become easily accessible to non-experts since gaining widespread notoriety around 2017 (Mehta et al., 2023; Naitali et al., 2023). This accessibility facilitates large-scale malicious use. It enables the sophisticated manipulation of digital content through techniques such as identity swapping, facial expression alteration and complete voice cloning (Dhanyalakshmi et al., 2025; Patel et al., 2023). Critically deepfakes are now leveraged to bypass traditional security and surveillance mechanisms including facial identification and biometric systems. Attackers impersonate authenticated individuals and provide fabricated proof of identity (Duhan & Kajal, 2025; Vijaykumar et al., 2024). This vulnerability has directly fueled an escalation in sophisticated cyber fraud. Recent high-profile cases underscore the acute financial risk. Fraudsters utilize hyper-realistic voice clones or video recreations to deceive corporate finance employees into transferring millions of dollars (Muhly et al., 2025). The global real-time dissemination of this fabricated content is almost exclusively achieved via social media platforms (Atlam et al., 2025; Mustak et al., 2023). The speed and scale inherent to these digital channels amplify marketplace deception and societal harms. This leads to a pervasive crisis of authenticity (Mustak et al., 2023).

The public increasingly struggles to distinguish credible information from disinformation. This actively erodes social trust and endangers democratic processes (Heidari et al., 2024; Ahmed, 2021; Darma et al., 2025). This systematic breakdown of trust is further exploited by malicious actors who employ the "liar's dividend." Factual content can be plausibly denied as merely another deepfake creation (Ahmed, 2021). The cybersecurity and media forensics

communities have engaged in an intensive arms race against deepfake generation in response to this rapidly evolving threat environment (Patel et al., 2023; Naitali et al., 2023). However existing detection systems face substantial limitations despite considerable research employing Deep Learning models like Convolutional Neural Networks (CNNs). Specifically concerns remain regarding generalizability and robustness against unseen manipulation techniques (Duhan & Kajal, 2025; Heidari et al., 2024). The constant evolution of generative models often outpaces the efficacy of reactive detection methods. This renders them obsolete shortly after deployment (Mustak et al., 2023; Naitali et al., 2023). Thus relying solely on technological detection is demonstrably insufficient for achieving the integrity and security standards required for modern digital environments. Addressing this requires a multi-faceted strategy encompassing technical, organizational and regulatory interventions (Muhly et al., 2025; Naitali et al., 2023). Therefore this paper undertakes a comprehensive literature review to meticulously analyze the critical issues presented by advanced deepfake technology. The primary objectives are to systematically detail the underlying mechanisms of deepfake generation and the corresponding threats posed to digital identity and social media security. It also aims to critically evaluate the performance, challenges and architectural limitations of current deep learning-based detection models. Finally it proposes a rigorous framework of integrated technical and policy safeguards essential for mitigating the pervasive societal and cybersecurity risks associated with deepfake proliferation.

2.0 Literature Review

The rapid advancement of artificial intelligence (AI) has instigated a paradigm shift in multimedia forensics. This redefines the challenge of content authenticity in the digital ecosystem (Farid, 2025; Naitali et al., 2023). Deepfakes are defined as hyper-realistic fabricated media including images, video and synthesized audio created using deep learning (DL) algorithms. They constitute a foundational threat to information security and societal trust (Alanazi et al., 2024; Altuncu et al., 2024; Naitali et al., 2023). This section systematically reviews the technological progression of deepfake creation. It analyzes the associated cyber fraud vectors and critically evaluates the current state of deepfake detection and its inherent limitations.

2.1. The Evolution of Deep Fake Technology

The capability to generate deceptive media content is rooted in the computational evolution of generative AI (Acim et al., 2025). Various methods previously allowed for media manipulation but the modern deepfake phenomenon accelerated dramatically following the introduction of Generative Adversarial Networks (GANs) in 2014 (Acim et al., 2025; Patel et al., 2023). The technical architecture of GANs employs a dual computational process designed to continuously improve the realism of synthetic content through adversarial competition (Duhan & Kajal, 2025). This process involves two core neural network components which are the generator (G) and the discriminator (D) (Naitali et al., 2023). The function of the generator is to produce synthetic digital data that faithfully mimics the probability distribution of real data samples (Duhan & Kajal, 2025; Patel et al., 2023). Concurrently the discriminator attempts to verify the authenticity of the material. It distinguishes the fake content produced by the generator from real training data (Duhan & Kajal, 2025). The system operates as a minimax two-player game. The generator strives to maximize the probability of the discriminator making an error (Patel et al., 2023). This allows it to learn to produce increasingly realistic and difficult-to-detect output.

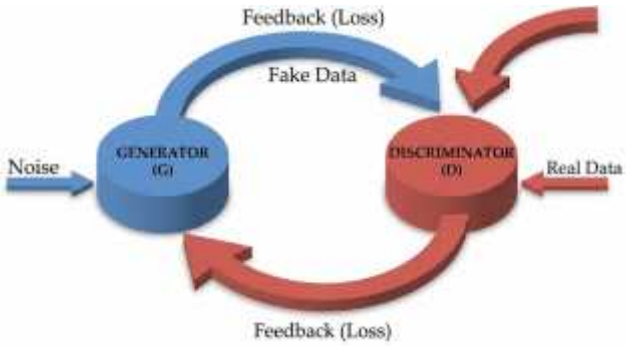

Figure 1: *The adversarial training cycle of a GAN. The Generator creates synthetic data from noise while the Discriminator evaluates it against real data, creating a feedback loop that progressively improves the realism of the deep fake.*

The deepfake domain transitioned rapidly from theoretical academia to widespread public accessibility building on this powerful generative capability (Acim et al., 2025). The proliferation of user-friendly open-source tools such as DeepFaceLab, FaceApp and Reface democratized the creation of sophisticated manipulations by 2017. This moved the capability beyond the reach of technical experts alone (Acim et al., 2025; Mehta et al., 2023; Naitali et al., 2023). Modern deepfake generation techniques extend beyond simple image manipulation to encompass entire face synthesis, identity swapping, attribute manipulation and expression swapping. This enables the complete replacement or alteration of an individual's digital identity in video, image or audio formats (Dhanyalakshmi et al., 2025; Patel et al., 2023; Tolosana et al., 2020). This continuous refinement in generation algorithms ensures that the output is often indistinguishable from authentic material to the naked eye (Heidari et al., 2024; Mustak et al., 2023).

2.2 Attack Vectors: Digital Identity Fraud on Social Media

The threat profile of deepfakes is amplified significantly by their ability to be effortlessly disseminated across global social media platforms (Atlam et al., 2025; Mustak et al., 2023; Naitali et al., 2023). Generative AI enhances the arsenal of cybercriminals across multiple vectors. It transforms simple scams into industrial-scale attacks characterized by hyper-realistic content and advanced targeting (Falade, 2023; Schmitt & Flechais, 2024).

Figure 2: *Taxonomy of Deep Fake attack vectors on social media platforms.*

A primary area of concern is the vulnerability of authentication systems. Deepfakes are actively weaponized to bypass conventional security and surveillance mechanisms particularly facial identification or biometric systems. These are used in processes such as Know Your Customer (KYC) or two-factor authentication (Duhan & Kajal, 2025; Vijaykumar et al., 2024; Mustak et al., 2023). Malicious actors utilize deepfakes to perpetrate identity fraud by impersonating individuals and providing fabricated proof of identity. This activity ranks among the top five methods of financial deception (Muhly et al., 2025). Deepfakes pose a critical threat via sophisticated social engineering (SE) and cyber fraud against organizations beyond bypassing technical security (Falade, 2023; Schmitt & Flechais, 2024). Generative AI provides the capability for highly realistic content creation and personalization. This enables attackers to convincingly impersonate senior executives or trusted partners (Falade, 2023; Muhly et al., 2025). Documented cases illustrate the severity of this risk. A 2019 incident involved an AI-generated voice clone used to trick a UK energy CEO into transferring $243,000 (Muhly et al., 2025). More recently in 2024 fraudsters utilized video deepfake recreations of a CFO and other staff during a malicious video conference call. This resulted in a staggering $25 million loss for a multinational firm (Muhly et al., 2025). This proliferation of convincing forged media undermines communication security. Dedicated messaging and video call platforms are now essential organizational defenses (Muhly et al., 2025). Furthermore the widespread propagation of synthetic media on social platforms poses profound reputational and societal risks (Mustak et al., 2023). Deepfakes are employed to create misinformation and propaganda. They disrupt public discourse and sway election results which can jeopardize national peace and stability (Alanazi et al., 2024; Darma et al., 2025; Heidari et al., 2024). The most insidious consequence is the exploitation of the "liar's dividend." The very existence of deepfake technology allows individuals including political figures to plausibly deny the authenticity of genuine factual video or audio evidence as merely another synthetic creation (Ahmed, 2021; Darma et al., 2025).

2.3 The Defense Landscape: Detection and Limitations

Research into deepfake detection has become an intensive arms race in response to the escalating threat (Naitali et al., 2023; Patel et al., 2023). Detection techniques generally operate reactively. They focus on identifying the residual traces or inconsistencies left by the generation process (Patel et al., 2023). These methods are typically categorized into image/video detection, audio spoofing detection and multimodal approaches (Patel et al., 2023). Deep Learning models are central to current defense strategies. Convolutional Neural Networks (CNNs) are the most predominant technique utilized for detecting forged images and videos. They are prized for their ability to autonomously extract spatial

characteristics including minute pixel-level discrepancies and artifacts created during manipulation (Duhan & Kajal, 2025; Heidari et al., 2024; Patel et al., 2023).

Complementary DL architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are often integrated with CNNs. This allows the analysis of temporal features crucial for detecting irregularities across video frames (Dhanyalakshmi et al., 2025; Patel et al., 2023). Detection clues exploit multiple levels of forensic analysis.

i. Artifact/Trace Detection: These methods focus on technical imperfections such as convolutional traces and bi-granularity artifacts from GAN up-sampling. They also look for blurred edges and blending errors at the facial feature boundaries (Duhan & Kajal, 2025; Naitali et al., 2023).
ii. Physical/Physiological Discrepancies: This category leverages inconsistencies visible to forensic tools but often unnoticeable to humans. Examples include abnormal eye blinking patterns, unnatural head pose directions or discrepancies in biological signals like heartbeat (Naitali et al., 2023; Patel et al., 2023).
iii. Multimodal Approaches: Hybrid systems combine analysis across visual and auditory domains to identify inconsistencies. These include misalignment between phoneme (speech sounds) and viseme (mouth movements) or lip-sync errors (Dhanyalakshmi et al., 2025; Patel et al., 2023).

Some proactive countermeasures suggest utilizing techniques like digital watermarking or blockchain provenance tracking to verify content authenticity from the source though these are underrepresented in literature (Naitali et al., 2023). Despite these advancements current detection systems face severe operational limitations and critical gaps. The most persistent challenge is the acute lack of generalizability (Patel et al., 2023; Heidari et al., 2024). Models trained specifically to detect deepfakes generated by one method frequently exhibit substantial performance degradation when confronted with content created by different or novel generative tools (Patel et al., 2023; Duhan & Kajal, 2025; Heidari et al., 2024). This failure in cross-dataset evaluation means that detection efficacy achieved in controlled environments often fails to meet the rigorous requirements of real-world deployment (Heidari et al., 2024). Conversely practical application is undermined by the crucial challenge of robustness against compression and low data quality (Patel et al., 2023). Social media platforms routinely compress video and image files. This inadvertently removes or minimizes the minute artifacts that DL detection models rely upon (Patel et al., 2023; Duhan & Kajal, 2025). The performance of current models drops significantly when applied to this lower-quality compressed content ubiquitous on the internet (Patel et al., 2023). Furthermore the evolution of generative techniques constantly introduces sophisticated anti-forensic measures. Adversarial perturbations are designed specifically to exploit detection algorithm vulnerabilities (Naitali et al., 2023; Mustak et al., 2023; Patel et al., 2023). This forces detection systems into a continuous cycle of updates and obsolescence. These limitations underscore the necessity for continued research focused not merely on incremental accuracy improvements but on developing sophisticated models with superior generalization capability (Duhan & Kajal, 2025; Patel et al., 2023)

3.0 Discussion and Critical Analysis

Synthesizing the literature gathered in this review reveals a somewhat alarming reality. The current ecosystem of digital identity verification is facing a systemic crisis that technical solutions alone seem ill-equipped to handle. The technological advancements in Generative AI are impressive but they have created a security environment defined by profound instability. The discussion that follows breaks this crisis down into two distinct but interrelated dimensions. First we examine the technical "arms race." We argue that the inherent asymmetry between generation and detection makes reactive defense a losing strategy. Second we shift focus to the human element. We analyze how deepfakes exploit psychological vulnerabilities and erode the fundamental trust mechanisms that underpin our social and economic interactions.

3.1 The Asymmetry of the "Arms Race"

The contemporary cybersecurity environment is characterized by a pervasive and escalating threat from deepfake technology. This is underpinned by an undeniable asymmetry between offensive and defensive capabilities. The offensive technology is primarily driven by sophisticated deep learning (DL) models such as Generative Adversarial Networks (GANs). It continues its rapid and almost ballistic trajectory in terms of realism, ease of use and overall accessibility (Acim et al., 2025; Farid, 2025). This relentless progression enables malicious actors to generate hyper-realistic fabricated media including images, video and synthesized audio. These are increasingly indistinguishable from authentic content and fundamentally compromise the integrity of digital communication (Heidari et al., 2024; Naitali et al., 2023). The efficiency with which new generative models emerge ensures that the creation of forged content consistently outpaces the development of reactive detection solutions (Mustak et al., 2023; Naitali et al., 2023).

This inherent asymmetry is acutely evident when assessing the efficacy of detection algorithms in real-world environments. This is particularly true against content distributed via social media platforms. Current detection strategies predominantly rely on identifying subtle residual traces or minute pixel-level inconsistencies introduced during the deepfake generation process (Duhan & Kajal, 2025; Patel et al., 2023). However social media platforms universally employ compression and post-processing mechanisms on uploaded video and image files (Patel et al., 2023). This compression inherently minimizes or entirely eliminates the microscopic artifacts upon which deep learning detection models depend for forensic analysis (Duhan & Kajal, 2025).Consequently models demonstrating high accuracy under controlled and uncompressed laboratory conditions experience immediate and substantial performance degradation when confronted with the low-quality compressed content ubiquitous across the internet (Duhan & Kajal, 2025; Patel et al., 2023). This failure in robustness suggests a losing battle for defenders. Furthermore detection models suffer from a critical lack of generalization capability. Models trained to identify deepfakes generated by one specific method often exhibit poor performance against media produced by novel or different generative tools (Duhan & Kajal, 2025; Heidari et al., 2024). This architectural fragility combined with the constant introduction of sophisticated anti-forensic measures by attackers traps the defense ecosystem in a continuous and reactive cycle. Detection systems are rendered potentially obsolete shortly after their deployment (Mustak et al., 2023; Naitali et al., 2023; Patel et al., 2023).

3.2 The Human Factor: Vulnerability of Trust

The success of deepfake-enabled cyber fraud hinges critically on the exploitation of fundamental human psychological vulnerabilities beyond the technological imbalance. This is a primary tenet of social engineering (Falade, 2023; Schmitt & Flechais, 2024). Deepfakes serve as a force multiplier for impersonation attacks. They create scenarios that leverage strong emotional or contextual trust signals (Muhly et al., 2025). Attackers skillfully employ hyper-realistic voice clones or video recreations to impersonate authoritative figures, senior executives or even close family members. This effectively bypasses logical scrutiny (Falade, 2023; Muhly et al., 2025). For older adults the exploitation of intergenerational affection or high trust in authority figures can override cultivated awareness. This renders them susceptible to schemes that compel urgent financial transfers or sensitive disclosures (Zhai et al., 2025). This dynamic exacerbates a pervasive erosion of digital trust. It confirms that the maxim "seeing is believing" is now fatally undermined (Mustak et al., 2023; Zhai et al., 2025). The core problem is not merely that human perception fails to distinguish high-fidelity synthetic media. Indeed studies show that distinguishing real from AI-generated faces is near-chance for the average person (Farid, 2025; Ahmed, 2021). Rather the content instills uncertainty regarding the veracity of all digital information. This systematic breakdown of trust creates the "liar's dividend." Factual content and especially inconvenient truths can be dismissed as mere deepfake fabrications which compounds societal disinformation risks (Ahmed, 2021; Farid, 2025). Consequently advanced generative technology may falter slightly against a dedicated anti-forensic algorithm. However its capacity to trigger emotional urgency and exploit inherent psychological trust mechanisms ensures that the human target remains the most vulnerable point of failure in the entire defense apparatus (Falade, 2023).

4.0 Conclusion and Future Recommendations

Having traversed the technical environment of Generative Adversarial Networks (GANs) and the increasingly sophisticated ecosystem of digital identity fraud we arrive at a sobering realization. The challenge of deepfakes is no longer merely a technical glitch to be patched but a systemic crisis of authenticity. The evidence reviewed in the preceding sections suggests that we are standing at a critical juncture. The democratization of forgery tools is rapidly outpacing our capacity to detect them. Therefore this final section aims to synthesize these findings into a coherent verdict. Perhaps more importantly it lays out a pragmatic multi-layered defense strategy that acknowledges the limitations of technology while empowering the stakeholders enforcement on platforms, policymakers and users.

4.1 Summary of Findings

The literature unequivocally confirms that deepfake technology constitutes an immediate and systemic threat to the integrity and authenticity of digital identity across global communication networks (Falade, 2023; Farid, 2025; Mustak et al., 2023). This is underpinned by the relentless progression of generative Artificial Intelligence (AI). The exponential rise of sophisticated forgery tools such as Generative Adversarial Networks (GANs) and large language models (LLMs) allows for the fabrication of media that profoundly undermines public trust in the visual and auditory record disseminated via social media platforms (Acim et al., 2025; Heidari et al., 2024). Critically current cybersecurity measures are demonstrably losing the technical arms race against content generation. They rely heavily on reactive deep learning detection models (Duhan & Kajal, 2025; Naitali et al., 2023). This asymmetry is exacerbated by practical constraints unique to the social media environment. Existing detection models exhibit fragile generalizability and suffer catastrophic performance degradation when confronted with compressed video formats and low-quality digital artifacts

ubiquitous in real-world circulation (Duhan & Kajal, 2025; Patel et al., 2023). This reliance on solely reactive forensic analysis operates in a constantly shifting adversarial landscape. This confirms that a fundamental and multidisciplinary overhaul of defense strategies is mandatory for organizations and policymakers alike (Muhly et al., 2025).

4.2 Recommendations for Stakeholders

Mitigating the pervasive threat posed by hyper-realistic deepfakes requires an immediate shift toward a coordinated multi-layered defense. This must incorporate robust technical mandates, decisive legislative frameworks and targeted user education.

Figure 3: *A proposed holistic defense framework against deep fake-driven identity fraud.*

For Social Media Platforms

i. Mandatory Content Provenance.
   - Platforms must transition from reactive content removal to enforcing Mandatory Content Credentials and Digital Watermarking standards at the point of upload (Farid, 2025; Naitali et al., 2023). This proactive approach embeds cryptographically secured metadata to verify the origin and subsequent modifications of digital media.

ii. Enhance Multi-Modal Detection.
   - Platforms should prioritize foundational investment in advanced multi-modal detection pipelines. These integrated systems must jointly analyze audio, video and lip-synchronization data to exploit subtle yet consistent inconsistencies that betray synthetic content. This provides superior detection robustness over single-modality models (Dhanyalakshmi et al., 2025; Patel et al., 2023).

For Policymakers

i. Specific Criminal Legislation.
   - Legislatures must urgently implement specific statutes that explicitly define and penalize the malicious misuse of deepfake technology. This moves beyond the ambiguity of existing fraud or defamation laws (Alanazi et al., 2024; Darma et al., 2025). This legislation should explicitly criminalize the creation and intentional dissemination of non-consensual synthetic media.

ii. Intermediary Accountability.
   - Legal frameworks must expand liability beyond the content creator to include platform intermediaries that negligently facilitate the rapid and mass proliferation of harmful deepfake content. This forces them to adopt a proactive governance model for digital integrity (Darma et al., 2025).

For Users

i. Advanced Digital Hygiene.
   - Users must elevate their Digital Hygiene protocols. This involves adopting a critical and sceptical mindset or the "pause-reflect-act" cycle when encountering high-pressure or emotionally manipulative content (Falade, 2023; Zhai et al., 2025).

ii. Biometric Data Control.
   - Individuals must drastically limit the public sharing of personal biometric data such as high-resolution images and voice snippets. This data serves as the critical training material that fuels deepfake generation engines (Farid, 2025; Zhai et al., 2025).

iii. Mandate MFA.
   - Adopting robust Multi-Factor Authentication (MFA) across all digital accounts is non-negotiable. It acts as a vital organizational defense layer that bypasses the effectiveness of deepfakes used in biometric spoofing attacks (Falade, 2023; Vijaykumar et al., 2024).

4.3 Limitations and Future Research

This comprehensive review is necessarily bound by a prevailing limitation. There is an inherent time lag between the continuous emergence of novel generative AI models and the availability of their subsequent peer-reviewed academic forensic analysis (Acim et al., 2025; Muhly et al., 2025). Consequently conclusions regarding the effectiveness of detection against the latest commercial-grade deepfakes remain tentatively constrained. Future research must immediately focus on three critical pathways to address this rapidly evolving threat. Firstly engineering highly generalizable and robust detection models capable of maintaining high performance against previously unseen synthesis methods and the pervasive artifact erosion caused by social media compression remains paramount (Heidari et al., 2024; Patel et al., 2023). Secondly there is an urgent need to develop real-time deepfake detection algorithms specifically optimized for resource-constrained environments such as smartphones and edge network devices (Duhan & Kajal, 2025; Naitali et al., 2023). Finally interdisciplinary studies exploring the psychological and cognitive vulnerabilities that permit highly personalized AI-driven social engineering attacks to succeed must be rigorously pursued to fortify the human defense layer (Ahmed, 2021; Schmitt & Flechais, 2024).

Acknowledgement

The authors extend their gratitude to all members of the School of Computing for their invaluable contributions to this study. This research was conducted under the Cybersecurity in Social Media Project and supported by Universiti Utara Malaysia (UUM).

References

Acim, B., Boukhlif, M., Ouhnni, H., Kharmoum, N., & Ziti, S. (2025). A decade of deepfake research in the generative AI era, 2014–2024: A bibliometric analysis. *Publications*, *13*(4), 50. https://doi.org/10.3390/publications13040050

Ahmed, S. (2021). Navigating the maze: Deepfakes, cognitive ability and social media news skepticism. *New Media & Society*. https://doi.org/10.1177/14614448211019198

Alanazi, S., Asif, S., & Moulitsas, I. (2024). Examining the societal impact and legislative requirements of deepfake technology: A comprehensive study. *International Journal of Social Science and Humanity*, *14*(2), 58–64. https://doi.org/10.18178/ijssh.2024.14.2.1194

Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets and a meta-review. *Frontiers in Big Data*, *7*, 1400024. https://doi.org/10.3389/fdata.2024.1400024

Atlam, E.-S., Almaliki, M., Elmarhomy, G., Almars, A. M., Elsiddieg, A. M. A., & ElAgamy, R. (2025). SLM-DFS: A systematic literature map of deepfake spread on social media. *Alexandria Engineering Journal*, *111*, 446–455. https://doi.org/10.1016/j.aej.2024.10.076

Darma, M., Gisymar, N. A., Purwadi, A., Zubaedah, P. A., & Judijanto, L. (2025). Legal implications of deepfake technology misuse in digital content on social media. *Science of Law*, *2025*(3), 98–103. https://doi.org/10.55284/eqazc148

Dhanyalakshmi, R., Stoian, G., Danciulescu, D., & Hemanth, D. J. (2025). A survey on face-swapping methods for identity manipulation in deepfake applications. *IET Image Processing*, *19*, e70132. https://doi.org/10.1049/ipr2.70132

Duhan, K., & Kajal, A. (2025). A comparative analysis of deep learning based approaches for deepfake identification. *Procedia Computer Science*, *259*, 482–493. https://doi.org/10.1016/j.procs.2025.03.350

Falade, P. V. (2023). Decoding the threat landscape: ChatGPT, FraudGPT and WormGPT in social engineering attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(5), 185–198. https://doi.org/10.32628/CSEIT2390533

Farid, H. (2025). Mitigating the harms of manipulated media: Confronting deepfakes and digital deception. *PNAS Nexus*, *4*(7), pgaf194. https://doi.org/10.1093/pnasnexus/pgaf194

Galyashina, E. I., & Nikishin, V. D. (2022). The protection of megascience projects from deepfake technologies threats: Information law aspects. *Journal of Physics: Conference Series*, *2210*, 012007. https://doi.org/10.1088/1742-6596/2210/1/012007

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery*, *14*(2), e1520. https://doi.org/10.1002/widm.1520

Mehta, P., Jagatap, G., Gallagher, K., Timmerman, B., Deb, P., Garg, S., Greenstadt, R., & Dolan-Gavitt, B. (2023). Can deepfakes be created on a whim? In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)* (pp. 123–126). ACM. https://doi.org/10.1145/3543873.3587581

Muhly, F., Chizzonic, E., & Leo, P. (2025). AI-deepfake scams and the importance of a holistic communication security strategy. *International Cybersecurity Law Review*, *6*, 53–61. https://doi.org/10.1365/s43439-025-00143-7

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations and opportunities. *Journal of Business Research*, *154*, 113368. https://doi.org/10.1016/j.jbusres.2022.113368

Naitali, A., Ridouani, M., Salahdine, F., & Kaabouch, N. (2023). Deepfake attacks: Generation, detection, datasets, challenges and research directions. *Computers*, *12*(10), 216. https://doi.org/10.3390/computers12100216

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3342107

Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, *57*, 324. https://doi.org/10.1007/s10462-024-10973-2

Vijaykumar, R., Purnapatra, S., Plesh, R., Imtiaz, M., Hou, D., & Schuckers, S. (2024). Deepfake attacks on biometric recognition: Evaluation of resistance to injection attacks. *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 404–410). IEEE. https://doi.org/10.1109/UEMCON62879.2024.10754732

Zhai, Y., Xue, X., Guo, Z., Jin, T., Diao, Y., & Jeung, J. (2025). Hear us, then protect us: Navigating deepfake scams and safeguard interventions with older adults through participatory design. In *CHI Conference on Human Factors in Computing Systems (CHI '25)* (pp. 1–15). ACM. https://doi.org/10.1145/3706598.3714423