Deepfake Technology in Social Engineering: Threats, Detection and Defence Strategies

Tanaka Makamure and Daniel Ogwok
University of Johannesburg, South Africa

222060509@student.uj.ac.za
danielo@uj.ac.za

Abstract: Big developments in the field of Artificial Intelligence (AI) have significantly improved deepfake technology, allowing individuals to create realistic synthetic media. As a result, deepfakes have become a very effective tool for social engineers by allowing them to accurately impersonate colleagues and authority figures to gain the trust of individuals and manipulate them. High-profile cases such as the Arup scam and the UAE fraud case highlight how attackers are leveraging these techniques in real-world scenarios and the devastating financial consequences they bring. In response, several technical detection tools have been developed, many of which rely on deep learning techniques. However, these should not be used in isolation. An effective defence strategy requires organisations to supplement these technological tools with other strategies, such as employee training, awareness campaigns and secure security policies. This creates layers of security, capable of mitigating risks posed by deepfake social engineering.

Keywords: Deepfakes, Social engineering, Exploits, Cybersecurity threats, Generative adversarial networks, Synthetic media

1. Introduction

Advancements in Artificial Intelligence have led to the creation of realistic synthetic media known as deepfakes (Pedersen, et al., 2025). Techniques such as generative adversarial networks (GANs) have enabled deepfake technology to convincingly replicate faces, voices and behaviours, making it very difficult for an individual to distinguish synthetic from authentic media (Yang, et al. 2021). Although such technology can be useful in fields such as entertainment, it can also be used maliciously, particularly in social engineering exploits.

Social engineering, unlike other cyber security exploits, focuses on the exploitation of human vulnerabilities, rather than technical weaknesses, to deceive and manipulate people into revealing private and confidential information (Siddiqi, Pak, & Siddiqi, 2022). Traditional social engineering attacks, such as phishing, rely on gaining human trust to be successful. The integration of deepfake technology into such attacks has the potential of aiding in the acquisition of this human trust, making social engineering attacks more effective.

The convergence of deepfake technology and social engineering presents a concerning threat to cybersecurity. This paper aims to explore how deepfakes have made it easier to manipulate and deceive people, enhancing the effectiveness of social engineering exploits. Furthermore, it reviews the current strategies being employed to detect and defend against such attacks. By taking into consideration the current state of knowledge, this paper also seeks to uncover the challenges that remain in protecting organisations against deepfake-based attacks. The research question being addressed therefore is "How is deepfake technology being used in social engineering attacks, and what strategies exist to detect and defend against them?"

The rest of the paper is structured as follows: Section 2 gives background information on the generation of deepfake media and provides an overview of social engineering tactics. Having explored these key concepts, Section 3 then discusses some case studies where deepfake technology was successfully used to launch social engineering attacks, paying attention to the attack mechanisms used. Section 4 then focuses on the detection tools that are currently being used, from tradition to deep-learning based ones. Finally, Section 5 explores the additional measures that organizations should have in place to complement technical detection strategies and provides a recommendation strategy for companies in the battle against these attacks.

2. Background

Although the introduction has outlined the growing threat presented by deepfakes in social engineering, it is important to understand the background information of these concepts. By first understanding the different types of deepfakes along with how they are generated, and what makes social engineering attacks effective, the convergence between the two fields becomes clearer. This section therefore provides an overview on deepfake technology, social engineering, and the intersection between them.

2.1 Deepfake Generation

The increasing access to cell phones, the availability of digital media on social media platforms, and advances in the field of machine learning have allowed for the manipulation of pictures, videos, and audio files to generate synthetic but realistic media, referred to as deepfakes (Masood, et al., 2022). The generation of the altered multimedia is done by machine learning algorithms, mainly GANs. Deepfakes fall under two classes, namely visual and audio deepfakes. Visual deepfakes can be classified into four categories namely attribute manipulation, identity swap, entire face synthesis and face reenactment (Akhtar, 2023). Attribute manipulation focuses on altering some facial features from authentic media, producing deepfakes with edited faces. Identity swap, on the other hand, involves the replacement of a person's face in the original video, with a different face, giving the illusion that the other person was in the video. Entire face synthesis involves using GANs to create new face images and placing them in an original video. Face reenactment is the process of replacing the facial expression of individuals in an original video with those of other individuals (Akhtar, 2023). Audio deepfakes are grouped into two categories namely, text to speech (TTS) where a speech is generated using someone else's voice, and voice conversion (VC) which involves the conversion of a speaker's voice to sound like someone else, without changing the words being said (Dagar & Vishwakarma, 2022). Deepfakes have become so sophisticated, with studies indicating that both humans and machine learning algorithms struggle to consistently distinguish them from authentic media (Korshunov & Marcel, 2018).

2.2 Social Engineering

Social engineering attacks occur when cybercriminals manipulate individuals into revealing confidential information or trick them into giving them access to a system (Hijji & Alam, 2021). Therefore, instead of attempting to circumvent firewalls and intrusion detection systems, social engineering exploits human vulnerabilities through techniques such as deception or manipulation to breach security (Wang, Zhu, & Sun, 2021). Such attacks are often successful as studies have shown human beings to be the weakest links when it comes to cybersecurity (Hijji & Alam, 2021). A common social engineering attack is phishing, where an attacker sends a message that is seemingly legitimate, prompting users to follow instructions detailed in the message. One way of conducting a phishing attack is through voice communication, where an attacker tricks users into providing sensitive information via a phone call. This is also known as vishing (Syafitri, et al. 2022). Most social engineering attacks involve an attacker gaining the trust of an unsuspecting user by posing as a legitimate entity.

2.3 Intersection Between Deepfakes and Social Engineering

The intersection between deepfake technology and social engineering has brought about a new class of cyber threats. Instead of just relying on text-based impersonation for social engineering attacks, hackers can now leverage seemingly legitimate audio and video media to gain the trust of unsuspecting users (Ferrara, 2019). Any form of social engineering attack that makes use of video or audio media, such as vishing, can be enhanced by deepfake technology. An example could be the use of deepfake technology to impersonate a manager asking an employee to transfer money into his account, or to provide some confidential information (Westerlund, 2019). This fusion increases the chances of success for attackers as many traditional defence mechanisms often struggle with recognising synthetic fabricated media (Martinek & Bartuzi-Trokielewicz, 2025).

Deepfake technology provides several tools that significantly enhance the strength of social engineering tactics. An example would be the use of voice cloning in vishing and impersonation attacks. These manipulation attacks are heavily dependent on the attacker's ability to pose as a trusted figure and gain the victim's trust. Deepfakes enhance this through the voice synthesis functionality, allowing attackers to convincingly mimic colleagues or executives over the phone. Compared to traditional vishing, where scammers' accents often raise suspicion, audio deepfake technology can mimic accents with high accuracy, reducing the victim's ability to identify the scam (Pedersen, et al., 2025).

Furthermore, combining audio and video deepfake techniques has the advantage of making the manipulated content look more authentic. This could happen when fraudsters use real-time video manipulation tools together with audio synthesis to create a realistic video of the figure being impersonated. A common example would be the use of this combination in video meetings, effectively tricking the victim into believing that they are interacting with their colleague or boss (Muhly, Chizzonic, & Leo, 2025). This amplified attack can be very successful if organisations rely only on visual confirmation to verify requests. To structure the analysis of this

intersection, this paper proposes a taxonomy mapping deepfake method, social engineering technique and corresponding defence strategy, presented in full in Section 5.3.

3. Deepfakes in Social Engineering Attacks

To better understand the application of deepfake technology to enhance social engineering, it is useful to examine real-life case studies where synthetic media was utilised in cyberattacks. These case studies showcase the mechanisms discussed in the previous section in real-world scenarios, demonstrating their effectiveness and the devastating impacts they leave on the victim organisations.

3.1 Arup Deepfake Video Scam

One of the most impactful cases of deepfakes being used in social engineering occurred in 2024 when an employee working at Arup, a UK-based company, was scammed into transferring money to the attackers (Lin, 2025). The attack started with a phishing email, where the chief financial officer (CFO) requested transfer of funds. The employee's doubts were wiped when the attackers made use of deepfake technology to impersonate the CFO and other company employees, cloning their faces and voices. This managed to convince the employee that the communication was legitimate, and he proceeded to transfer funds to five different Hong Kong bank accounts (Lin L. S., 2025). In total, this social engineering attack resulted in a financial loss of approximately US$25.6 million (Doraisamy, et al. 2025). This attack demonstrates the dangerous outcomes of bringing together deepfakes and manipulation techniques, in this case, the use of audio and video deepfakes to convince an unsuspecting individual to comply with an urgent request.

3.2 UK Energy Company Scam

A similar attack occurred at a UK-based energy company, where attackers leveraged voice cloning technology to impersonate one of the company's executive official (Farid, 2025). The attack began when a phone call was made to the company's CEO from an individual who was impersonating his superior. Initially, the CEO had no reason to doubt the call, as the deepfake voice cloning convincingly replicated the superior's voice as well as his German accent (Pedersen, et al., 2025). Believing this call to be genuine, the CEO was tricked into authorising a payment transfer of approximately US$ 243,000 to a Hungarian supplier. The attackers subsequently requested additional funds, but the CEO grew suspicious after noticing no reimbursements had been made. Investigations carried out of the case concluded that the attackers had used a commercially available voice-generating software to impersonate the superior (de Rancourt-Raymond & Smaili, 2022). This case is an example of how voice cloning on its own is powerful enough to catch individuals off-guard and exploit their trust, manipulating them into doing what exactly the attacker wants.

3.3 UAE Bank Scam

A third high-profile deepfake social engineering attack occurred in 2020 at a United Arab Emirates (UAE) bank. To execute this attack, cybercriminals used voice cloning technology to mimic the voice of a company's director and proceeded to make a call to a bank branch manager asking him to authorise a series of transfers (Farid, 2025). This request for funds was accompanied by emails and supplementary documents from an alleged lawyer, adding credibility to the request. Since the branch manager was familiar with the director's voice, and was additionally reassured by the supporting documents, he had no reason to suspect foul play, therefore he proceeded with approving the transactions (Lin L. S., 2025). In total, approximately US$35 million was transferred to the attackers through various international accounts, and a portion of it was traced to the United States. Investigators who were assigned to this case concluded that the attackers had used speech samples from public sources or hijacked communications to clone the director's voice. This case is regarded as one of the biggest financial fraud attacks utilising deepfake technology (Lin L. S., 2025). This social engineering attack demonstrates how traditional phishing methods such as email attacks are significantly enhanced using deepfakes, making the process of gaining human trust, and ultimately manipulating them much easier.

3.4 Almost Successful Financial Scams

Not all attempts to use deepfakes in social engineering have been successful, but they still demonstrate the harmful potential this convergence possesses. In 2020, the advertising company WPP became the target of a deepfake scam, where attackers leveraged deepfake audio cloning to impersonate the CEO (Zhang, et al. 2025). The attack was initiated when the fraudsters used a public image of the CEO to create a WhatsApp account and proceeded to arrange a virtual meeting with a senior executive. They then made use of the CEO's voice and a public footage to impersonate the CEO during the meeting, and they requested for a funds transfer (Muhly, Chizzonic, & Leo, 2025). A similar attack targeted a Swiss financial institution in 2024, where

cybercriminals used deepfakes to mimic the company's chief financial officer (CFO). However, the financial clerk managed to recognise inconsistencies in the fake CFO behaviour and dress code, making the attack unsuccessful (Muhly, Chizzonic, & Leo, 2025). While both scam attempts were unsuccessful and no money was transferred, it highlights how audio and video deepfake technology can be combined, to reinforce trust and credibility.

4. Deepfake Detection

The case studies discussed in the previous section demonstrate how deepfake technology can be applied to boost the success of social engineering attacks, resulting in severe financial consequences. From this information, it becomes apparent that this convergence presents a real threat, and therefore there is need for countermeasures to detect and protect organisations against such attacks. This section examines the current technology being used by detection tools in the battle against deepfakes, with the aim of evaluating their effectiveness and highlighting the limitations that need to be addressed in defending against these amplified attacks.

4.1 Audio Deepfake Detection

Multiple approaches to audio deepfake detection have been explored over the years. One such approach involves the extraction of features from an audio file, then applying machine learning algorithms to detect whether the voice has been tampered with (Stroebel, et al. 2023). The use of algorithms such as GMM (Gaussian Mixture Model) and LCNN has been proposed for such detections. The findings indicate that these models perform well on fully synthesised audio files, however, they tend to struggle with partially synthesised ones.

A more effective approach uses deep learning algorithms, such as neural networks, to learn patterns that emerge in fabricated audio files, and use those patterns in classification (Shaaban, Yildirim, & Alguttar, 2023). This can be achieved using CNNs to analyse features from audio mel spectrograms, which can be used to differentiate altered audio files from genuine ones (Stroebel, et al. 2023). Several experiments have shown that the CNN-based models outperformed traditional GMM-based ones in distinguishing spoofed audio files from authentic files. For example, an analysis carried out on the ASVspoof2019 dataset revealed that CNN models achieved a lower Error Equal Rate (ERR) of 0.59 compared to 6.4 achieved by GMM models. The lower ERR indicates the superiority of CNN models in detection accuracy (Salih, et al., 2025). However, even though these models perform well in a controlled environment, they often struggle with generalisation, limiting their overall effectiveness in real world situations.

Table 1: Comparison of audio deepfake detection methods

Method	Strengths	Weaknesses
GMM/LCNN	Strong on fully synthesised audios	Performance drops on partial audio
CNN + mel spectrograms	Better performance than GMM/LCNN	Poor generalisation

4.2 Traditional Image Deepfake Detection Methods

Traditional deepfake detection tools, such as Error Level Analysis (ELA), rely on looking for unnatural facial movements, or imperfections within the fabricated videos such as speech consistencies, shadows as well as lack of breathing (Taeb & Chi, 2022). If such a tool detects one or more of these inconsistencies, that video will be flagged as fabricated. However, the introduction of AI algorithms such as convolutional neural networks (CNNs) and GANs has significantly improved the quality of deep-fakes, eliminating these imperfections. Therefore, these traditional tools are becoming ineffective, particularly against deep learning based deepfakes (Taeb & Chi, 2022).

4.3 Deep Learning-based Detection

A better approach to deepfake detection is to utilise deep learning algorithms such as CNNs and deep neural networks (DNNs) (Stroebel, et al. 2023). These deep learning techniques are classified into two categories namely holistic and feature-based categories.

Holistic approaches to deepfake detection, such as support vector machines (SVM) and CNN classifiers, focus on the entire face, to detect irregularities that can be used to differentiate authentic images from fabricated ones (Taeb & Chi, 2022). An experiment carried out by (Mallet, et al., 2023) demonstrated the effectiveness of a CNN model in identifying synthetic images, achieving a detection accuracy of 88.33% on a dataset of real and

fake faces. However, despite this good performance, holistic approaches face generalisation challenges. An analysis by (Raza, et al., 2026) has shown that CNN-based models suffer a performance decline of about 15% when evaluated on different datasets. This suggests that even though holistic approaches show promising results in controlled environments, their effectiveness in real-world situations is limited.

Feature-based methods, on the other hand, extract features that are unique to deepfakes, and these are then used to classify real and fabricated media. These features are grouped into biometric, media and model classes. Biometric features are human features, such as eye and lip movement, head pose and eyebrows, that can be used in the distinction of deepfake and genuine videos. Media features focus on information like noise and inconsistencies between frames in media, and this information can be used to identify videos that have been tampered with (Zhang, et al. 2025). Model features, unlike media features, are based on the model used in the deepfake generation and refer to specific fingerprints left behind by the models, particularly GAN models, in the generated media. Deep learning models can scan for the presence of these fingerprints to help in the identification of fake media (Zhang, et al. 2025) (Taeb & Chi, 2022). However, the weakness of fingerprint detection stems from the fact that models trained to detect GAN-based deepfakes struggle with diffusion-based deepfakes, because these introduce entirely different artifact patterns (Singh & Dhumane, 2025).

Table 2: Comparison of deep learning-based detection approaches

Approach	How it works	Strengths	Weaknesses
Holistic (SVM and CNN)	Full face analysis	High detection accuracy on known dataset	Poor generalization
Feature-based (Biometric, media and model)	Extracts deepfake-specific features media	Captures movement inconsistences	Struggle with diffusion-based deepfakes.

4.4 Limitations

Although the above-mentioned detection methods show promising results, they face several limitations. Experiments on both machine learning and deep learning models have revealed that these detection methods struggle to generalise. Therefore, while they perform well on a single dataset, the detection accuracy drops when the models are evaluated on unseen deepfakes. An experiment by (Singh & Dhumane, 2025) has highlighted this performance drop, where a CNN-based model that achieved detection accuracy of 90% on one dataset only managed a 60% accuracy on a new unseen dataset.

Another significant limitation of existing detection tools is that of scalability. State-of-the-art detection models require significant computing resources including premium GPUs and large memory capacity (Singh & Dhumane, 2025). As a result, real-time deepfake detection remains a challenge, especially for smaller organisations, with limited computer infrastructure.

Finally, there is a limitation of partial media manipulation. As discussed in Section 4.1, GMM and LCNN based classifiers produce good performance on fully synthesised audio, however, their performance drops when evaluating partially synthesised samples (Stroebel, et al., 2023). This suggests that attacks that blend authentic and spoofed media have a higher chance of evading detection. Table 3 presents a summary of limitations faced by each detection approach.

Table 3: Summary of limitations across detection approaches

Detection approach	Limitations
Traditional image detection methods (ELA)	Obsolete against CNN-generated deepfakes
GMM and LCNN (audio detection)	Poor performance on partial media
Audio CNN + mel spectrogram	Poor generalisation
Holistic (CNN and SVM)	Performance drop across datasets (15% drop)
Feature-based approach	Struggle against diffusion-based deepfakes Computationally heavy

Overall, these limitations highlight that no single detection method offers complete protection against deepfake-powered social engineering.

5. Defence Strategies

While the detection tools discussed in the previous section provide a layer of defence against deepfakes, they do not always flag every manipulation attempt. Therefore, detection tools on their own cannot fully mitigate the threats possessed by deepfake social engineering exploits. Consequently, organisations should have other defence mechanisms in place to reduce the probability of successful attacks. This section discusses human-centric strategies along with organisational safeguards that can be utilised in shielding companies from these attacks. It then concludes by suggesting a recommended strategy organisations can employ, by structuring their defence in layers, adopting a defence-in-depth approach to security.

5.1 Human-centric Strategies

Humans have been identified as the weakest link when it comes to cybersecurity (Hijji & Alam, 2021), therefore, in addition to having automated detection tools, it is also necessary to ensure that employees are made aware of the threat posed by deepfake technology. One way of achieving this is by organizing training and awareness campaigns focused on equipping individuals with the knowledge required to detect altered media (Lin L. S., 2025). A study conducted in Bangladesh demonstrates the effectiveness of awareness campaigns in the battle against deepfakes. This study was conducted in two phases, with the results from phase one indicating that 65% of the sample group was totally unaware of deepfakes. However, after being made aware of this concept, phase two results were promising, with 70% of the group feeling confident that their deepfake detection abilities had improved (Ahmed, et al. 2021).

5.2 Organisational Safeguards

Organisations can also take a proactive approach in the battle against deepfake scams by putting in place controls and safeguards to assist in the detection and defence against attacks. By defining strong and secure policies, companies ensure that employees have an established framework to follow when dealing with requests (Muhly, Chizzonic, & Leo, 2025). A practical example would be having a policy outlining secure communication tools and platforms organisational employees should use, that way, any request made from a non-approved platform is immediately shot down (Pedersen, et al., 2025). Furthermore, organisations should outline reporting procedures to follow when employees face deepfake scams (Pedersen, et al., 2025).

5.3 Recommended Approach

Given the complexity of deepfake-fuelled social engineering attacks, having a single mitigation strategy cannot guarantee complete protection. Case studies discussed in section 3 demonstrate that attackers use different deepfake methods and social engineering techniques, each requiring a unique response. This paper therefore proposes a taxonomy that maps the attack method and social engineering technique to the appropriate defence strategy.

Table 4: Taxonomy of Deepfake Social Engineering Attacks and Defences

Deepfake method	Social Engineering technique	Real-world example	Defence strategy
Audio only	Vishing	UK energy and UAE bank scams	CNN mel spectrogram analysis. Call verification policy. Employee training.
Video and Audio	Visual and audio impersonation	Arup scam, WPP attempt	Holistic CNN. Feature-based analysis. Secure communication policies.
Audio and Documents	Phishing enhancements	UAE bank scam	Email filtering. Audio deepfake detection. Multi-channel verification policies.

As the table illustrates, organisations should adopt a defence-in-depth strategy, combining technical detection tools, human awareness and training, along with secure organisational policies. For example, a vishing attack

requires a combination of audio CNN detection and strict call verification policies, ensuring that even if the detection tool fails, a trained employee following the policy acts as a secondary defence. The key

advantage of this multi-layer approach to security is that even if one layer is bypassed, such as an employee being tricked, the organisation will have additional safeguards, preventing a single point of failure.

6. Conclusion

Deepfake technology has been significantly enhanced by breakthroughs in the field of artificial intelligence, enabling the creation of convincing and realistic synthetic audio and video media. This enhancement has negatively revolutionised deepfake social engineering attacks, allowing cybercriminals to exploit individual trust with ease, putting organisations at risk of deepfake fraud. This paper introduced key concepts in the fields of deepfakes and social engineering by exploring the different ways in which deepfakes are created and introducing the key factors that allow social engineering exploits to be successful. Then it proceeded to examine how the two fields can be brought together, describing how deepfake technology can be utilised in social engineering attacks such as vishing and video conference scams, by impersonating an authoritative figure. Several high-profile case studies were additionally explored, further reinforcing the intersection of the two fields and the devastating consequences it has on the victims. The discussion then shifted to the possible solutions by examining the current detection tools that are available, ranging from early forensic tools to deep-learning based approaches. Additionally, the paper considered necessary supplementary defence mechanisms, such as training and awareness campaigns for employees and secure organisational policies, which can be used in conjunction with technical detection tools, effectively implementing a defence-in-depth approach to securing organisations. Finally, this paper proposes a taxonomy mapping attack methods to the appropriate defence strategies, offering a practical framework for organisations.

The objective of this paper was to investigate the role of deepfake technology in social engineering by discussing the risks posed and the detection and defence measures that can be employed. This objective has been met through case study explorations and a thorough examination of the current state of detection and defence mechanisms. Although total mitigation of this threat remains a challenge due to huge leaps in the field of AI, this paper identifies a layered approach to security as the best course of action against deepfake social engineering.

Ethics declaration: This study did not require any ethical clearance. The research was conducted using information from public databases and no collection of personal data was done.

AI declaration: No AI tools were used in the writing of this paper.

References

Ahmed, M. F. B., Miah, M. S. U., Bhowmik, A. & Sulaiman, J. B., 2021. *Awareness to Deepfake: A resistance mechanism to Deepfake.* s.l., IEEE, pp. 1-5.

Akhtar, Z., 2023. Deepfakes Generation and Detection: A Short Survey. *Journal of Imaging,* 9(1), p. 18.

Dagar, D. & Vishwakarma, D. K., 2022. A literature review and perspectives in deepfakes: generation, detection, and applications. *International Journal of Multimedia Information Retrieval,* 11(3), pp. 219-289.

de Rancourt-Raymond, A. & Smaili, N., 2022. The unethical use of deepfakes. *Journal of Financial Crime,* 30(4), p. 1066–1077.

Doraisamy, V., Ab Rahman Muton, N., Rasalingam, R. & Ab Malik, A., 2025. Cybersecurity Challenges and Solutions in the Metaverse: A Critical Review of Threats, Risks, and Technologies. *PaperASIA,* 41(4b), p. 267–276.

Farid, H., 2025. Mitigating the harms of manipulated media: Confronting deepfakes and digital deception. *PNAS Nexus,* 4(7), p. pgaf194.

Ferrara, E., 2019. The history of digital spam. *Communications of the ACM,* 62(8), p. 82–91.

Hijji, M. & Alam, G., 2021. A Multivocal Literature Review on Growing Social Engineering Based Cyber-Attacks/Threats During the COVID-19 Pandemic: Challenges and Prospective Solutions. *IEEE Access,* Volume 9, p. 7152–7169.

Korshunov, P. & Marcel, S., 2018. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *CoRR,* Volume abs/1812.08685.

Lin, L. S., 2025. Examining the Role of Deepfake Technology in Organized Fraud: Legal, Security, and Governance Challenges. *Frontiers in Law,* Volume 4, pp. 6-17.

Lin, L. S. F., 2025. Organisational Challenges in US Law Enforcement's Response to AI-Driven Cybercrime and Deepfake Fraud. *Laws,* 14(4), p. Article 46.

Mallet, J., Pryor, L., Dave, R. & Vanamala, M., 2023. *Deepfake Detection Analyzing Hybrid Dataset Utilizing CNN and SVM.* New York, Association for Computing Machinery, pp. 7-11.

Martinek, A. & Bartuzi-Trokielewicz, E., 2025. *Detecting deepfakes and false ads through analysis of text and social engineering techniques.* Abu Dhabi, UAE, Association for Computational Linguistics, p. 8432–8448.

Masood, M. et al., 2022. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence,* 53(4), pp. 3974-4026.

Muhly, F., Chizzonic, E. & Leo, P., 2025. AI-deepfake scams and the importance of a holistic communication security strategy. *International Cybersecurity Law Review,* 6(1), pp. 53-61.

Pedersen, K. T. et al., 2025. Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments. *Journal of Cybersecurity and Privacy,* 5(2).

Raza, A. et al., 2026. A Comprehensive Review of Deepfake Detection Techniques: From Traditional Machine Learning to Advanced Deep Learning Architectures. *AI,* 7(2), p. 68.

Salih, A. et al., 2025. Deepfake Audio Detection in Voice Authentication: A Spectral and CNN-Based Comprehensive Review. *Engineering, Technology & Applied Science Research,* Volume 15, pp. 29824-29832.

Shaaban, O. A., Yildirim, R. & Alguttar, A. A., 2023. Audio Deepfake Approaches. *IEEE Access,* Volume 11, pp. 132652-132682.

Siddiqi, M. A., Pak, W. & Siddiqi, M. A., 2022. A Study on the Psychology of Social Engineering-Based Cyberattacks and Existing Countermeasures. *Applied Sciences,* 12(12), p. 6042.

Singh, S. & Dhumane, A., 2025. Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. *MethodsX,* Volume 15, p. 103632.

Stroebel, L. et al., 2023. A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology,* pp. 83-113.

Syafitri, W. et al., 2022. Social Engineering Attacks Prevention: A Systematic Literature Review. *IEEE Access,* Volume 10, p. 39325–39343.

Taeb, M. & Chi, H., 2022. Comparison of Deepfake Detection Techniques through Deep Learning. *Journal of Cybersecurity and Privacy,* 2(1), pp. 89-106.

Wang, Z., Zhu, H. & Sun, L., 2021. Social Engineering in Cybersecurity: Effect Mechanisms, Human Vulnerabilities and Attack Methods. *IEEE Access,* Volume 9, p. 11895–11910.

Westerlund, M., 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review,* 9(11), pp. 40-53.

Yang, C., Ding, L., Chen, Y. & Li, H., 2021. *Defending against GAN-based DeepFake Attacks via Transformation-aware Adversarial Faces.* s.l., IEEE, pp. 1-8.

Zhang, B., Cui, H., Nguyen, V. & Whitty, M., 2025. Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. *Sensors,* 25(7), p. 1989.