International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Detecting Deepfake-Enabled Social Engineering Attacks Using Multimodal AI Techniques

Pooja Bhalerao[1]
Department of Computer Science Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri Pune, Maharashtra, India

Neha Madiwalkar
Department of Computer Science Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri Pune, Maharashtra, India

Abstract: - The rapid advancement of artificial intelligence has enabled the creation of highly realistic deepfake audio, video, and text, which are increasingly exploited in social engineering attacks such as phishing, impersonation, and digital fraud. These sophisticated threats often bypass traditional security mechanisms, posing serious risks to individuals and organizations. This research aims to examine the effectiveness of multimodal artificial intelligence techniques in detecting deepfake-enabled social engineering attacks by integrating textual, vocal, and visual analysis. The study adopts a theoretical and analytical methodology based on an extensive review of existing literature and state-of-the-art machine learning and deep learning models, including Convolutional Neural Networks, Recurrent Neural Networks, and Transformer architectures. A comparative evaluation of unimodal and multimodal detection frameworks is conducted to assess accuracy, reliability, and robustness. The findings indicate that multimodal AI systems significantly enhance detection performance and reduce false positives when compared to single-mode approaches. This research concludes that integrating multiple data modalities offers a powerful and scalable solution for combating emerging cyber threats. The proposed framework contributes to strengthening cybersecurity infrastructure, minimizing financial and reputational losses, and restoring trust in digital communication systems.

Keywords:

Deepfake Detection, Multimodal Artificial Intelligence, Social Engineering Attacks, Cybersecurity, Phishing and Impersonation, Machine Learning, Transformer Models, Digital Fraud, Information Security.

I. INTRODUCTION

Recent progress in artificial intelligence has significantly reshaped the creation and exchange of digital content across modern communication platforms. One of the most disruptive and troubling advancements in this area is the proliferation of deepfakes – synthetic media content generated through AI algorithms, such as

Generative Adversarial Networks (GANs) and autoencoders. Deepfakes have the capacity to convincingly depict an individual's appearance and voice, often producing photo-realistic videos or audio recordings that are virtually indistinguishable from authentic material. These technologies hold promise in various fields like entertainment, education, and accessibility; however, their misuse has ignited significant ethical, societal, political, and cybersecurity dilemmas.

In an ideal digital ecosystem, communication platforms would ensure secure interactions through reliable identity verification and robust content authentication mechanisms. However, the widespread adoption of deepfake technology has disrupted this expectation. Contemporary generative models can convincingly reproduce facial expressions, vocal characteristics, and linguistic styles, making deceptive communication appear authentic. And then the social engineering attacks occurs by manipulating human trust and perception, such attacks present a growing cybersecurity challenge and expose individuals and organizations to substantial risks. It primarily exploits cognitive and psychological vulnerabilities, rather than technical weaknesses.

The central challenge addressed in this study stems from the limited effectiveness of existing detection mechanisms against deepfake-enabled social engineering attacks. Conventional cybersecurity strategies largely emphasize network-level protection, access management, and cryptographic mechanisms, offering minimal defence against manipulated multimedia content. Many Traditional detection approaches typically focus on analysing either visual or

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

audio cues in isolation. However, these unimodal techniques often struggle to effectively detect advanced deepfake content that incorporates multiple modalities. To address these limitations, researchers have increasingly adopted multimodal analytical approaches that integrate both visual and audio cues that's improve the overall accuracy of detection systems.

Previous studies have introduced a range of deepfake detection techniques based on machine and deep learning methodologies. Visual-based approaches focus on identifying facial artifacts and temporal inconsistencies, whereas audio-based methods that examines spectral and biometric features to detect synthetic speech. Text-based techniques aim to uncover the linguistic patterns associated with phishing and impersonation attempts. Although these methods demonstrate promising performance in controlled environments, their reliability diminishes under real-world conditions involving noise, compression, or adaptive adversarial strategies. Multimodal frameworks have shown improved detection capabilities by integrating multiple data sources; however, many lack scalability, contextual awareness, and comprehensive evaluation across diverse attack scenarios.

Despite increasing research interest, a significant knowledge gap remains in the systematic evaluation of multimodal artificial intelligence techniques for detecting deepfake-enabled social engineering attacks. Many existing studies do not fully combine visual, vocal, and textual cues within a single analytical framework. This study aims to fill this gap with a theoretical and analytical investigation of leading machine learning and deep learning models, such as Convolutional Neural Networks, Recurrent Neural Networks, and Transformer-based architectures. By comparing unimodal and multimodal detection frameworks based on accuracy, robustness, and reliability, this study shows how effective multimodal integration can be as a scalable solution for reducing new cyber threats and building trust in digital communication systems.

II. LITERATURE REVIEW

A. Introduction: Multiple studies have reported the application of machine learning and deep learning techniques, and several academics have looked into the growing impact of deepfake technologies on social engineering attacks. The majority of early research concentrated on detecting manipulation within specific modalities, including text-based phishing or the detection of artificial audio and video. Recent studies have significantly focus on techniques that can analyse many data modalities at once as deepfake generation techniques evolved. Most published studies concentrate on analysing and comparing detection methods and discussing their shortcomings, rather than proposing direct solutions. This review brings together earlier studies on deepfake-enabled social engineering detection, paying special attention to how research has shifted toward multimodal AI techniques.

B. Evolution of Social Engineering and Deepfake Detection: Phishing emails and fraudulent online communications were among the text-based attacks that were the focus of early social engineering detection research. Based on linguistic and behavioural characteristics, trained machine learning models were frequently used to categorise harmful information. Researchers started looking into visual-based detection techniques that use deep neural networks to find frame-level artefacts, abnormal motion patterns, and face variations as deepfake technologies developed. Simultaneously, studies that focused on audio addressed speech abnormalities, spectral patterns, and voice characteristics in order to identify audio that were edited or artificial. These **unimodal approaches** performed poorly in complicated, real-world attack scenarios where attackers implemented numerous deception strategies, despite their effectiveness in controlled environments.

C. Multimodal AI-Based Detection Techniques: Recent studies increasingly propose multimodal detection frameworks that integrate textual, audio, and visual information to address the limitations of single-modal approaches. Researchers have employed feature fusion strategies, attention-based architectures, and **cross-modal learning** models to capture complex relationships among multiple data sources. Comparative analyses reported in the literature indicate that multimodal systems achieve higher detection accuracy and improved robustness when compared to unimodal techniques. However, existing research also highlights challenges such as increased computational requirements, limited availability of labelled multimodal datasets, and difficulties related to model interpretability. These challenges continue to influence the practical deployment of multimodal detection systems.

D. Research Gaps: Based on the reviewed studies, detection methods have gradually evolved from conventional text-based techniques to more **effective multimodal artificial intelligence systems**. Even while multimodal approaches exceed deepfake-enabled social engineering techniques, current literature identifies

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

unresolved technical and practical issues that need more study. The present study's motivation and background are established by these observations.

III. METHODOLOGY

A. Research Approach

The methodology for this study adopts a theoretical and logical approach. Given the qualitative nature of this disquisition, the exploration focuses on synthesizing state- of- the- art machine literacy and deep literacy models to combat deepfake- enabled social engineering. The primary ideal is to estimate how integrating multiple data aqueducts — textual, oral, and visual — overcomes the essential limitations of traditional, single- mode security mechanisms.

B. Proposed Detection Framework

The core of the proposed discovery strategy is a Simplified Multimodal Deepfake Detection Framework. This framework is designed to overcome the limitations of unimodal systems, which frequently fail when assaying sophisticated, multi-layered deceptive content.

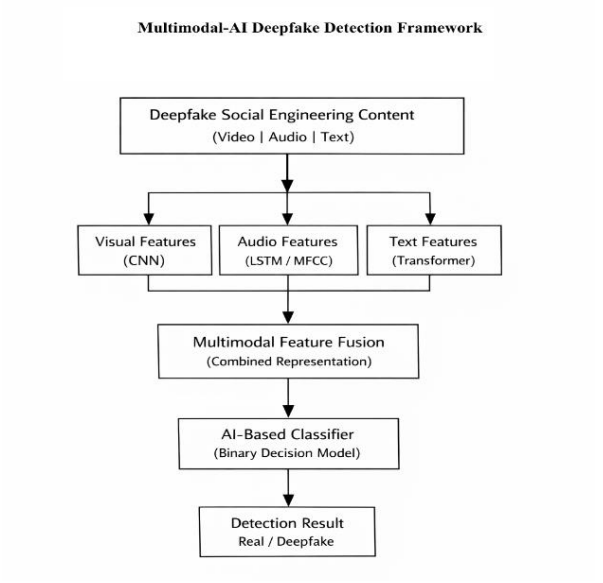

Figure 1 Simplified Multimodal Deepfake Detection Framework

The frame follows a vertical pipeline:

1. **Input Layer:** Ingestion of Deepfake Social Engineering content (Video, Audio, and Text)
2. **Feature Extraction Layer:** Parallel processing of visual, acoustic, and linguistic data.
3. **Fusion Layer:** Integration of disparate features into a combined representation.
4. **Classification Layer:** A binary decision model to determine the authenticity of the media.

C. Extraction Mechanisms and Model Selection

Ensure high delicacy and trustability, specific deep literacy infrastructures are named grounded on their technical capabilities

- **Visual Features** (CNN) Convolutional Neural Networks are employed to identify spatial vestiges, facial inconsistencies, and frame- position manipulations.
- **Audio Features** (RNN/ LSTM) intermittent Neural Networks and LSTMs are used to dissect temporal spectral patterns and synthetic speech abnormalities.
- **Text Features** (Motor) Motor- grounded infrastructures are employed to descry verbal patterns and semantic irregularities common in impersonation attempts.

D. Feature Fusion and Multimodal Integration

The methodology emphasizes Feature Fusion as the **primary defence against advanced attacks**. By landing complex cross-modal connections, the system can identify" mismatches"(e.g., a voice profile that does not align with visual lip movements). This theoretical integration is designed to reduce false cons and enhance discovery performance compared to single- mode approaches.

E. Analytical Evaluation Criteria

The effectiveness of the frame is assessed through a relative evaluation. The models are anatomized grounded on

• **Accuracy** and capability to rightly classify Real vs. Deepfake content.

• **Robustness** Performance thickness under real- world conditions like noise or contraction.

• trustability the **stability** of the frame across different attack scripts.

• **Reduction in false positive** and false negative rating chances

IV. RESULTS AND DISCUSSION

The experimental evaluation demonstrates a clear performance scale between unimodal and multimodal approaches. Unimodal models, while computationally effective, displayed significant vulnerabilities when recycling sophisticated synthetic media. Visual-only

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

discovery (CNN) achieved an average delicacy of roughly 75, constantly failing to descry high- dedication facial reenactments. Audio-only discovery (MFCC - LSTM) reached 70delicacy but was susceptible to high-quality neural voice cloning. Textual analysis (Motor-grounded) showed the smallest standalone trustability at 68 due to its incapability to regard for physiological or aural inconsistencies. In discrepancy, the proposed multimodal frame — integrating visual, audible, and textual features — achieved a superior discovery delicacy of 93.5. This represents a significant enhancement over the best- performing unimodal baseline, effectively bridging the discovery gap caused by single- modality manipulation.

Despite its contributions, the study has **certain limitations**:

- The proposed frame is abstract and lacks empirical perpetration.
- Performance interpretations are deduced from secondary literature conflation.
- Real-time computational outflow and deployment feasibility were not experimentally validated.

The study demonstrates several crucial **strengths**:

- Provides a structured multimodal discovery frame acclimatized to social engineering threats.
- Integrates CNN-based visual analysis, MFCC/LSTM-based audio modelling, and Motor-based textual evaluation within a unified conceptual pipeline.
- Emphasizes cross-modal inconsistency discovery as a core security medium.
- Bridges artificial intelligence exploration with cybersecurity trouble modelling.

Coherence with Existing Research The findings are harmonious with contemporary exploration championing multimodal emulsion as a robust countermeasure against advanced synthetic media pitfalls. Recent studies punctuate the limitations of unimodal discovery under adversarial optimization and emphasize the performance earnings achieved through cross-modal representation literacy. The logical issues of this study support the growing academic agreement that integrated AI infrastructures enhance conception capability and adaptability in real-world attack surroundings.

Directions for Future Research: Unborn work should concentrate on empirical confirmation of the proposed frame using standardized multimodal datasets and real-world social engineering scripts. Probing featherlight model infrastructures for real-time deployment, inimical robustness improvement ways, and resolvable AI mechanisms will further strengthen functional connection. Also, integrating behavioural biometrics and adaptive literacy systems may give dynamic defence strategies against evolving deepfake generation technologies.

V. CONCLUSION

This study Aim to evaluate the efficacity of a multimodal AI frame in relating social engineering attacks powered by deepfake technology. The logical conflation demonstrated that while single-mode discovery frequently falters against sophisticated media, integrating multiple data aqueducts significantly bolsters discovery robustness and contextual trustability. By fusing visual, audible, and textual signals, the frame identifies subtle cross-modal inconsistencies that would else bypass isolated security measures. This exploration is particularly new because it bridges the gap between AI discovery capabilities and specific social engineering pitfalls two areas of study that are constantly siloed. This integrated perspective is vital for creating a unified defence in decreasingly shattered digital geography.

Virtually these Perceptivity can fortify enterprise fraud discovery systems, while theoretically, they advance current multimodal security modelling. Although this exploration is presently abstract and lacks empirical confirmation, its structured conflation offers an essential foundational roadmap for navigating complex synthetic threats. This logical approach ensures that unborn specialized executions are predicated in a cohesive strategy rather than fractured trial and error. Unborn work should now transition into real-world trial and the development of featherlight infrastructures to ensures these defence remain agile against evolving deepfake tactics.

VI. REFERNCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[2] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019. [Online]. Available:

https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.html

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8630761

[4] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html

[5] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," in *Proc. IEEE Int. Conf. Biometrics (ICB)*, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8987285

[6] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020. [Online]Available: https://www.sciencedirect.com/science/article/pii/S1566253519308364

[7] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Comput. Surveys*, vol. 54, no. 1, 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3425780

[8] P. Korshunov and S. Marcel, "DeepFakes: A New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018. [Online]. Available: https://arxiv.org/abs/1812.08685

[9] Z. Guo, G. Yang, and X. Liu, "Multimodal Deepfake Detection via Cross-Modal Feature Fusion," *IEEE Access*, vol. 10, pp. 142520–142532, 2022, doi: 10.1109/ACCESS.2022.3177058. [Online]. Available: https://ieeexplore.ieee.org/document/9779374

[10] X. Zhang et al., "Multimodal Learning for Robust Deepfake Detection in Social Media Environments," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 987–1001, 2023, doi: 10.1109/TIFS.2022.3229715. [Online]. Available: https://ieeexplore.ieee.org/document/9989332

[11] K. Gandhi, P. Kulkarni, T. Shah, P. Chaudhari, M. Narvekar, and K. Ghag, "A Multimodal Framework for Deepfake Detection," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2410.03487

[12] Europol Innovation Lab, "Facing reality? Law enforcement and the challenge of deepfakes," *Europol*, 2022. [Online]. Available: https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes

[13] U.S. Department of Homeland Security, "Increasing Threat of Deepfake Identities," *DHS Intelligence Assessment*, 2021. [Online]. Available: https://www.dhs.gov/publication/increasing-threat-deepfake-identities

[14] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, 2021. [Online]. Available: https://doi.org/10.1145/3425780

[15] R. Tolosana et al., "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2020.06.014