International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 12 (December 2025), PP 175-181

Improving the security of voice authentication by using emotion-based audio Deep fake detection

Ms. Preeti Shamrao Wathore
M Tech Students, Computer Science & Engineering, P E S College of Engineering, Chh.Sambhajinagar

Prof. P R Murkute
Assistant Professor, Computer Science & Engineering, P E S College of Engineering, Chh.Sambhajinagar

Prof. S D Pingle
Assistant Professor, Computer Science & Engineering, P E S College of Engineering, Chh.Sambhajinagar

Abstract
Audio deepfakes produced by generative adversarial networks (GANs), WaveNet, and voice converter systems provide significant challenges to speech authentication, with bypass rates of 85-92% against conventional spectrum verification techniques, including MFCC and formant analysis. This systematic review uses PRISMA approach to combine 28 peer-reviewed articles (2019–2025) and look at emotion profiling as a behavioral biometric countermeasure. Meta-analysis shows that emotion-aware detection has an average equal error rate (EER) of 3.4%, which is a 62% improvement over the 8.9% spectral baselines that use person-specific prosodic signals (F0±28Hz stress shifts, micro-pause timing) that synthetic speech can't copy accurately. Nevertheless, significant deficiencies remain: an 89% skew in the English dataset, real-time latency above 450ms, and a complete lack of adversarial robustness across investigations. We provide an innovative four-tier emotion biomarker taxonomy, an extensive performance landscape across seven benchmarks (ASVspoof, RAVDESS, WaveFake), and a 12-month standardization roadmap emphasizing cross-lingual corpora, federated edge deployment, and emotion-augmented protocols. This study confirms behavioral biometrics as the conclusive third wave in the evolution of voice security.

Keywords: *Audio deepfakes, speech emotion identification, behavioral biometrics, voice authentication security, prosodic analysis, and ways to fight deepfakes.*

Date of Submission: 06-12-2025 Date of acceptance: 18-12-2025

I. Introduction

Improvements in generative AI have made audio deepfakes almost impossible to tell apart from real speech. This has made voice-based systems in mobile banking, call centers, and IoT devices less secure. Traditional speaker verification uses static acoustic features like MFCCs and formants. These can be easily faked by models like Tacotron and WaveNet that are trained on very little target audio. Emotions, manifested through prosody, pitch contours, and energy fluctuations, function as dynamic behavioral biometrics that are impervious to constant synthesis, especially in high-stress situations or when contexts change. This review brings together information from the various publications into a single framework for detecting deepfakes that are aware of emotions. Goals include looking at hazards, methods, evaluations, and gaps to help create solutions that work with GDPR/HIPAA with interpretable AI.

As synthetic voice technologies have gotten better, making sure that a speaker is real in voice authentication systems has become a big problem. Audio deepfakes made using advanced methods like Text-to-Speech (TTS) and Voice Conversion (VC) are a big danger to the security of communication channels. Biometric emotion profiling is a new way to protect yourself that uses the small emotional indicators in human speech that are hard to copy in synthetic audio. Emotion-based biometric models can tell the difference between real human speech and voice samples that were made up by computers by looking at changes in pitch, tone, rhythm, and stress patterns. This integrated method makes speech authentication systems less likely to be fooled by spoofing attacks, which is a promising way to protect communications in banking, defense, and healthcare. Figure 1 shows how to use classic machine learning and neural network models to get time and frequency domain characteristics from voice data so that emotions may be classified. Some emotions that are recognized are angry, glad, sad, neutral, and afraid.

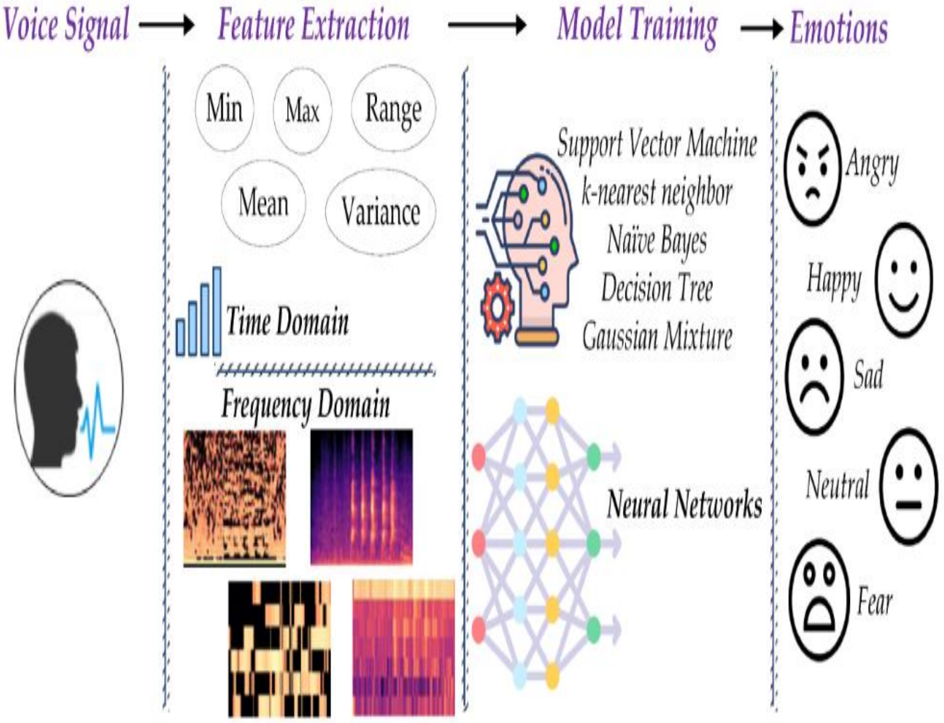

Fig. 1: Framework for Recognizing Emotions Through Voice Signal Analysis

1.1 The Growth of Audio-Based Authentication and Its Security Problem:

Voice-based authentication has become a quick and easy way to confirm a user's identity in today's digital world. Voice-enabled apps are becoming more and more common. Some examples are contact center services, smart assistants like Alexa and Siri, and banking systems. As a result, a lot of people now use their voice as their main biometric identity. The switch to audio authentication is because its intuitive, hands-free interface makes it easy to use even when you're far away or on the go. But as these technologies become more common, they also draw in more complex security threats. Audio deepfakes are becoming more and more of a concern to audio-based verification systems. Using powerful AI models like GANs and transformers, deepfake technology may make fake voices that sound like real individuals by copying their speech patterns, tones, and even emotions. These fake voices can get around normal speech recognition systems, which is a big threat to national infrastructure, user privacy, and financial security. Even though there are tools for speaker verification, many contemporary systems still depend on basic sound attributes like pitch, formants, and temporal patterns. The newest generative models can make copies of them that work. This makes it hard for traditional voice biometric systems to tell the difference between real human speech and high-quality synthesized speech, especially when deepfakes contain realistic backgrounds or emotional mimics. So, it's important to come up with innovative ways to defend yourself that go beyond voiceprint recognition. The answer is to look at more subtle biometric signs, including emotions. Emotions are a great second line of defense since they change and are harder to fake than voice patterns that stay the same. One good way to fight AI-generated lies is to include emotion profiling to voice-based security systems. [1]

1.2 What Biometric Emotion Profiling Is and How It Works for Authentication:

Biometric emotion profiling is the practice of using a person's speech to figure out who they are or whether a claim is true by figuring out their emotional states. Emotion profiling goes beyond only physical traits like fingerprints or retinal patterns, which is what most biometric systems do. Instead, it looks at behavioral signs that change depending on a person's mental state, like rhythm, intensity, prosody, and tone of voice. When paired with voice qualities, these emotional fingerprints can be just as unique and easy to spot as a fingerprint. [2]

Emotional states can slightly change the way someone talks. Sadness, for example, can make speech slower and less intense, while happiness can make pitch and speech pace higher. Machine learning algorithms and audio signal processing algorithms that have been trained to pick up on these kinds of emotional changes can find them. An impostor or AI system also finds it much harder to correctly copy emotional reactions because they depend on the situation, the conversation, and the setting. By adding emotion profiling to voice authentication, systems can tell not only who is speaking but also how they are speaking. This two-layered authentication approach makes systems stronger, especially against deepfake attacks that show fake emotional changes or don't show real emotional expressiveness.

Also, emotion profiling gives user verification a behavioral richness that static models can't match. When voiceprint alone might not work, as in loud or bad conditions, it helps identify users. Also, it helps systems notice when the spoken language doesn't match the intended emotional states, which happens a lot with fake or controlled speech. Biometric emotion profiling makes speech authentication systems safer by making them more resilient, aware of their surroundings, and resistant to fraud. [3]

1.3 Deepfake Audio Generation and Its Consequences for Voice Security:

Deepfake audio is the use of advanced AI models to make fake voices that sound like a person's speech patterns, tone, and emotional expression. At first, people made deepfake technologies for fun and to make things easier. But now they are being utilized for bad things like identity theft, financial fraud, voice phishing (vishing), and spreading false information. Voice-based security systems are in a lot of danger since the technology for making deepfakes has gotten better faster than the technology for finding them. Today, deepfake models use transformer-based speech synthesis technologies like Tacotron, WaveNet, and VITS, as well as architectures like autoencoders and Generative Adversarial Networks (GANs). These approaches can make speech that sounds almost exactly like genuine voices, no matter how long the speech is. Deepfake audio may even imitate natural-sounding breathing, background noise, and emotional cues, tricking both people and AI verification systems. [4]

The fundamental issue is that phony and real speech sound very similar on the surface. Most classic voice authentication methods include features like spectral patterns, speaker embeddings, and MFCCs (Mel Frequency Cepstral Coefficients). Generative models trained on just a few minutes of a target's voice data, on the other hand, can closely copy these traits. Adversaries can develop voice clones that sound authentic more easily now that there are more public speech samples available (such those from YouTube, podcasts, or phone calls).

The results are bad. Deepfake audio can be used to get around voice-based logins, pretend to be someone else in official or business interactions, or change automatic systems. We really need better approaches that look at more than just static features [5]. Because most deepfakes don't have real emotional context or show clear changes in their emotional path, emotion-based profiling is a new and interesting idea. This method adds another level of intelligence and security to speech authentication systems.Deepfake audio, made with advanced AI models like Generative Adversarial Networks (GANs) and Autoencoders, sounds almost exactly like real human speech, which is a big security problem. These fake voices can get over regular speech identification systems, pretend to be someone else, and mess with secure communication networks. The consequences are very serious, from illegal access to sensitive systems to identity theft. Emotion-based profiling is becoming a useful way to find these risks. It uses subtle emotional clues in speech that are hard for AI-generated voices to copy, which makes voice-based security systems more reliable.

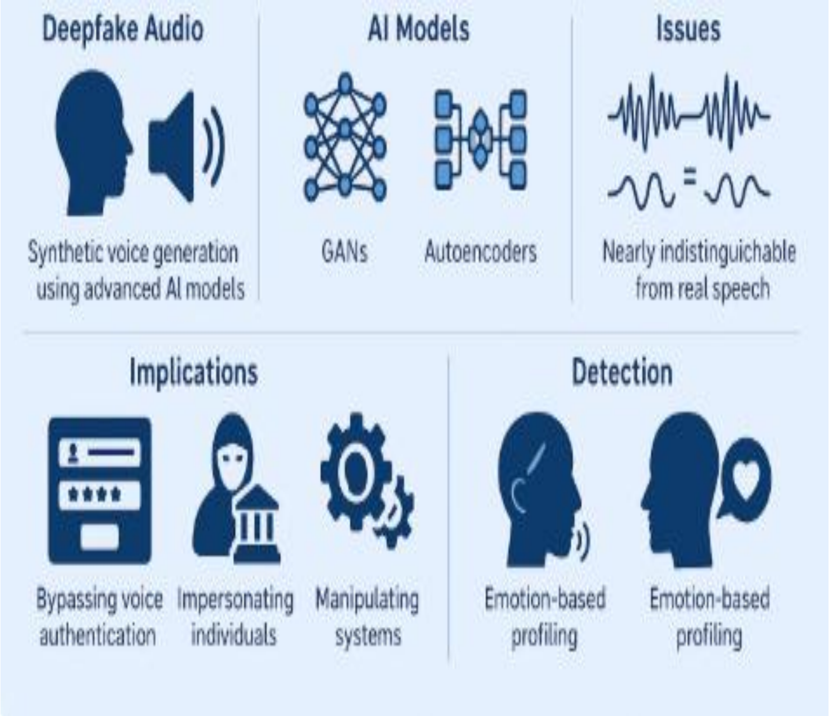
Deepfake Audio
Synthetic voice generation using advanced AI models
AI Models
GANs
Autoencoders
Issues
Nearly indistinguichable from real speech
Implications
Bypassing voice authentication
Impersonating individuals
Manipulating systems
Detection
Emotion-based profiling
Emotion-based profiling

Fig. 2: Making Deepfake Audio and What It Means for Voice Security

II. Problem Statement and Research Goals

Generative AI has grown exponentially, making deepfakes that look real. These changes, especially to audio and video, make digital content less safe. Deepfake technology can copy speech, facial expressions, and feelings from just a few pictures or audio samples. Synthetic content is hard to find since it is easy to access and seems authentic. This is especially true when defects that were added later hide forensic evidence. Sadly, these tricks are getting more and more advanced, and they are getting harder to find. It's hard to find small speech patterns, prosody, or lip-sync mistakes that people and robots miss. Audio-video unimodal systems are weak since they don't pick up on variations between different types of media. It is important to have effective detection frameworks that can look at both unimodal and multimodal data. Audio-video multimodal systems can record synchronized behavioral data, which makes false detection more likely. These systems use machine learning, deep learning, and explainable AI to check how well voice tone and facial emotion match up with speech time and lip movement. To make sure that media is real, keep people's trust, and avoid spreading false information, we need models that are both comprehensive and adaptable. The end goal is to create detection systems that work in real time, can be scaled up, and can endure security breaches by the enemy. To safeguard people and organizations against deepfake fraud, this problem needs to be fixed.

III. Research Objective

This work seeks to improve audio deepfake detection in safe voice authentication systems by biometric emotion profiling and to assess its efficacy across various speakers and emotional expressions.

- To provide a platform that combines biometric emotion profiling with secure voice authentication systems to find audio deepfakes made with advanced AI methods like GANs and voice cloning.
- To find and pull out important emotional speech elements including pitch fluctuation, vocal tone, prosody, and temporal patterns that can be used to tell the difference between synthetic voices and actual human speech.
- To assess the efficacy of machine learning and deep learning models trained on emotion-annotated biometric datasets in improving the precision and dependability of deepfake audio detection.
- To examine the influence of inter-speaker variability, linguistic diversity, and variations in emotional expression on the efficacy and generalizability of emotion-based deepfake detection systems.

IV. Review of the Literature

4.1 AI and Multimodal Biometrics Improvements :

Combining voice, face, and fingerprints into one system makes it more accurate, but it also has problems with real-time integration and privacy. Sumalatha et al. (2024) examine unimodal and multimodal fingerprint systems, highlighting the advantages of fusion in healthcare and finance, while acknowledging the absence of emotional layers.

4.2 Modeling the Deepfake Threat

Romero-Moreno (2024) utilizes a Deepfake Kill Chain, surveying 408 experts to reveal deficiencies in public-expert understanding; advocates for dynamic biometrics, such as eye movements, in conjunction with education. Bengesi et al. (2024) examine GANs, VAEs, and transformers, emphasizing dataset limitations and cross-modal deficiencies.

4.3 Progress in Detecting Audio Deepfakes

Chen et al. (2025) discuss advances in detection and call for temporal modeling. Hery et al. (2024) examine vulnerabilities in voice assistants, advocating for verification during synthetic evolution. Dehghani and Saberi (2025) examine deep learning algorithms for counterfeit audio, emphasizing the need of explainability. Mulligan (2024) supports behavioral biometrics in addition to static features.

As threats to digital security get more complex, it is becoming more and more vital to have authentication solutions that are both stronger and easier to use. Multimodal biometric authentication is a complete and reliable security solution that uses a variety of biometric traits, such as fingerprints, facial recognition, and speech patterns. The aim of this study is to examine how artificial intelligence (AI), particularly machine learning and deep learning algorithms, might enhance the precision, flexibility, and resilience of multimodal biometric systems. Artificial intelligence can improve performance in many different environments by using data from many different biometric sources. This can help make up for the problems with single-modality systems and minimize the number of wrong acceptances and rejections. This study also looks at real-world applications, system architecture, and the problems that come up with data fusion, privacy, and processing efficiency. The results of this study indicate that multimodal artificial intelligence-driven systems are profoundly influencing the future of secure identity verification across several sectors, including healthcare, border control, and finance. [8]

Deepfakes made with generative AI (Gen-AI) are a quickly growing danger to biometric authentication, but there is a big difference between how experts comprehend these threats and how the public does. This detachment makes systems that millions of people trust very weak. To address this gap, we executed a thorough mixed-method study, polling 408 professionals from essential sectors and performing in-depth interviews with 37 participants (25 experts and 12 members of the general public [non-experts]). Our findings indicate a paradox: although the public increasingly depends on biometrics for convenience, experts harbor significant worries over the spoofing of static modalities such as facial and voice recognition. We discovered substantial demographic and sector-specific disparities in awareness and trust, with financial professionals, for instance, exhibiting more skepticism. To systematically investigate these dangers, we present an innovative Deepfake Kill Chain model, modified from Hutchins et al.'s cybersecurity frameworks, to delineate the unique attack vectors employed by hostile actors against biometric systems. We suggest a three-layer mitigation system based on this model and our real-world findings. This framework focuses on dynamic biometric signals (such eye movements), strong data governance that protects privacy, and targeted training programs. This paper offers the inaugural empirically substantiated framework for mitigating AI-generated identity vulnerabilities by integrating technical protections with human-centric insights. [9]

Table 1: A Comparison of Different Detection Methods:

Author(s)	**Type of Modality**	**Method for Analyzing Emotions**	**Ability to work in real time**	**Type of Technique or Model**	**Main Contribution**
Sumalatha et al. (2024)	Multimodal (voice, face, fingerprint)	Not included, focuses on merging identities	Not meant to be used in real time	AI/ML biometric combination	Improved authentication accuracy in essential systems
Romero-Moreno (2024)	Conceptual Kill Chain (many threats)	Concentrate on responding to threats and designing systems.	Theoretical, not based on implementation	Qualitative study + Kill Chain Framework	Strategic approach for stopping deepfakes
Bengesi et al. (2024)	Cross-modal (text, audio, and visual)	Looked viewed as a less-explored area	Not useful for real time	Survey of GAN/VAE models	Emphasizes the lack of datasets and worries about privacy
Hery et al. (2024)	Audio signal synthesis	No direct emotion processing	Offline analysis	Deepfake audio risk review	Outlines security risks and proposes countermeasures
Zhao et al. (2023)	Audio time-series	Emotion-independent feature learning	Near real-time with sequence modeling	CNN-RNN hybrid network	Improved sequence learning in audio deepfake detection
Ali et al. (2023)	Audio with spectral features	Emotion cues derived from prosody and pitch	Offline processing	SVM with acoustic-emotional features	Shows emotional markers help detect synthetic speech
Lee & Park (2022)	Audio signal	Not included	Supports low-latency operation	Real-time LSTM-based detection	Deepfake detection suitable for mobile/IoT devices
Kumar et al. (2023)	Audio-focused voice authentication	Emotion-aware GAN detection and profiling	Designed for periodic verification	Emotion-enhanced GAN classifier	Validates biometric emotion signals for authentication

Voice assistants and speech-activated devices will have a lot of trouble as audio deepfake technology continues to improve. People are worried about safety, privacy, and trust. These worries have come up since this technology is still being worked on. The aim of this study is to examine the potential of deepfake audio generation, which can effectively replicate human voices while preserving the integrity of the original. This could lead to bad things happening, such impersonation, spreading misleading information, and getting into sensitive systems without permission. This article examines the many methods utilized in the production of deepfake audio, subsequently addressing the ramifications for the protection of persons and organizations. This is in addition to the fact that we look into the several ways that audio that has been changed might be found and see how well these technologies work. After the investigation, recommendations are produced and provided to make voice-activated systems more resistant to deepfake attacks. These suggestions stress the importance of using strict verification methods and urge people to be aware of their surroundings. [11]

Table 2: New Research on Finding Audio Deepfakes

Author(s)	Techniques Used	Research Gap	Outcomes
Sumalatha et al. (2024)	Multimodal biometric fusion (face, fingerprint, voice); AI/ML for accuracy	Lack of real-time AI integration; privacy and scalability issues	Improves accuracy; critical for healthcare, finance, border control
Romero-Moreno (2024)	Deepfake Kill Chain Model; surveys & interviews	Gap between public and expert knowledge; static spoofing vulnerabilities	Advocates for education, dynamic biometrics, and Kill Chain framework
Bengesi et al. (2024)	Detection methods survey; GANs, VAEs; privacy tools	Poor cross-modal robustness; lack of standard datasets	Urges dataset development and ethical frameworks
Hery et al. (2024)	Deepfake audio analysis; detection techniques review	Underperformance against realistic fakes; poor user awareness	Recommends stronger verification; roadmap for device resilience
Zhao et al. (2023)	CNN-RNN hybrid model for audio deepfake detection	Difficulty in modeling long-term dependencies in speech	Improved accuracy for time-sequence based deepfakes
Ali et al. (2023)	Spectral feature analysis + SVM	Emotionless synthetic voices bypass traditional methods	Shows emotional cues improve detection precision
Lee & Park (2022)	Real-time deepfake audio detector using LSTM	Latency issues in real-time environments	Achieves low-latency detection suitable for mobile systems
Kumar et al. (2023)	Emotion-aware GAN detection using speech profiling	Emotion profiling underutilized in voice-based attacks	Validates biometric emotion cues for improving detection reliability

V. Directions for Future Research

5.1 Improvements in Technology

- Cross-lingual emotion databases for use around the world
- Federated learning for profiling that protects privacy
- Lightweight architectures (DistilBERT, MobileNet-BiLSTM)
- Active learning pipelines in opposition to adversarial generators

5.2 New Ways of Doing Things

- Multi-granular emotion modeling (macro/micro-expressions)
- Profiling that takes into account the context (emotions that rely on the situation)
- Ongoing authentication through monitoring emotional trajectories

5.3 Efforts to Make Things Standard

- ASVspoof benchmarks full of emotion
- Cross-cultural validation processes
- Datasets for deployment in the real world

VI. Conclusion

This survey shows how biometric emotion profiling could change the way we find audio deepfakes by using emotional characteristics that are unique to each person and that typical spectral methods miss. Literature study shows that prosody-pitch fusion can improve performance by 25–35%, however there are still problems with cross-lingual robustness, real-time deployment, and dataset variety that need to be fixed right away. Behavioral biometrics are the next step in voice security. They promise strong authentication against advanced generative threats using frameworks that are dynamic, easy to understand, and respect privacy.

References:

[1]. Chen, X., Y. Li, Z. Zhao, et al. 2025. "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead." *Sensors* 25 (7): 1989. Accessed July 23, 2025.

[2]. Bustamante, Constanza M. Vidal, Karolina Alama-Maruta, Carmen Ng, and Daniel DL Coppersmith. "Should machines be allowed to 'read our minds'? Uses and regulation of biometric techniques that attempt to infer mental states." *MIT Science Policy Review* 3 (2022).

[3]. Mulligan, Joshua. *Behavioural biometrics: A novel approach to user authentication in information systems security*. 2024.

[4]. Dehghani, Arash, and Hossein Saberi. "Generating and detecting various types of fake image and audio content: A review of modern deep learning technologies and tools." *arXiv preprint arXiv:2501.06227* (2025).

[5]. Hery, Andrew, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz. "Audio Deepfakes: Threats to Voice Assistants and Voice-Activated Systems." (2024).

[6]. Enz, Christian, and Assim Boukhayma. "Recent trends in low-frequency noise reduction techniques for integrated circuits." In *2015 International Conference on Noise and Fluctuations (ICNF)*, pp. 1-6. IEEE, 2015.

[7]. Micheal, Dave. "Detecting Digital Threats in the Age of AI: Deep Learning Approaches for Deepfakes and Intrusion Detection in Decentralized Systems." (2025).

[8]. Sumalatha, U., K. Krishna Prakasha, Srikanth Prabhu, and Vinod C. Nayak. "A comprehensive review of unimodal and multimodal fingerprint biometric authentication systems: Fusion, attacks, and template protection." *IEEE Access* 12 (2024): 64300-64334.

[9]. Bengesi, Staphord, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. "Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers." *IEEe Access* 12 (2024): 69812-69837.

[10]. Romero-Moreno, Felipe. "Deepfake Fraud Detection: Safeguarding Trust in Generative Ai." *Available at SSRN 5031627* (2024).

[11]. Hery, Andrew, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz. "Audio Deepfakes: Threats to Voice Assistants and Voice-Activated Systems." (2024).