Air journal of Artificial Intelligence

ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

EMERGING TECHNIQUES IN DEEPFAKE VOICE DETECTION: METHODS, RISKS, AND ETHICAL IMPLICATIONS

Muhammad Bilal Ikram[1], Ahsan Arif[2], Asad Riaz[*3]

[1,2, *3] *Lecturer, Department of Computer Science, University of Agriculture Lahore.*

[1]bilalikramm20@yahoo.com, [2]ahsankhan67@gmail.com, [*3]asadraiz34@gmail.com

Keywords
Deepfake voice, speech synthesis, voice cloning, Generative Adversarial Networks (GANs), synthetic speech detection, voice authentication, digital trust, ethical implications.

Article History
Received: 03 January 2026
Accepted: 15 March 2026
Published: 31 March 2026

Corresponding Author: *
Asad Riaz

Abstract

Deepfake voice technologies represent a transformative advancement in artificial intelligence, particularly in speech synthesis and voice cloning. Leveraging deep learning architectures such as Generative Adversarial Networks (GANs) and autoencoders, these systems can produce highly realistic synthetic voices that closely resemble human speech. While offering benefits in domains such as accessibility, entertainment, and personalized services, deepfake voices also pose significant risks, including misinformation, identity theft, and cybercrime. This paper examines both the generation methods and detection strategies for synthetic speech, with a focus on neural network–based approaches to voice authentication and deepfake recognition. Furthermore, it discusses the ethical and legal challenges associated with deepfake voice technologies, emphasizing issues of consent, digital trust, and privacy. By critically reviewing recent advancements and proposing a structured framework for detection, this study seeks to contribute to the development of secure, transparent, and resilient solutions against malicious voice manipulation.

INTRODUCTION

Deepfake voice technology represents a rapidly evolving branch of artificial intelligence, focused on synthesizing human-like speech with high fidelity and realism. By employing cutting-edge algorithms such as Generative Adversarial Networks (GANs) and neural voice cloning, this technology [1] can replicate the unique characteristics of a person's voice, including tone, pitch, and accent. While the potential applications are immense—ranging from entertainment and virtual assistants to personalized healthcare—the rapid evolution of deepfake voice technology has introduced significant ethical, legal, and security challenges [2]. One of the most alarming concerns is its misuse in creating fraudulent audio content, such as impersonation in financial scams, spreading misinformation, or breaching voice-based authentication systems [3]. As deepfake voices become increasingly indistinguishable from genuine human speech, it becomes imperative to develop systems capable of detecting and mitigating their impact. Given the growing sophistication of synthetic voice generation, it is essential to examine how deep fake voice technologies are being applied across various domains and the implications they carry [4].

Beyond political deception, deepfake technology has been used for identity fraud and crimes [5]. Financial institutions have reported cases in which fraudsters utilized deepfake-generated voices to impersonate executives, allowing fraudulent transactions that resulted in considerable financial losses. Furthermore, deepfake technology has presented

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

hurdles to law enforcement and forensic analysis [6]. The growing sophistication of AI-powered media manipulation has made it impossible for courts and investigative authorities to authenticate digital evidence, delaying legal proceedings and weakening trust in audio-visual documentation [7]. Legal systems have struggled to keep pace with technological changes, resulting in disparities in how deepfake-related crimes are pursued across jurisdictions. Despite these risks, deepfakes are not always evil. When used appropriately, it has promising applications in a number of fields, including education, accessibility, and the arts [8]. AI-powered content production has opened up new possibilities in filmmaking, allowing for seamless digital character duplication and more storytelling. Deepfake-generated speech synthesis has also given those with speech impairments a voice, increasing accessibility for disabled people [9]. These promising uses demonstrate that, when properly regulated, deepfake technology may be a powerful tool for innovation [10]. Given the potential for both beneficial and harmful applications, advanced detection methods and regulatory systems are becoming increasingly necessary. AI researchers and cybersecurity experts are working on detecting algorithms that can reliably identify tainted media. However, as deepfake technology progresses, so do counter-exploitation measures [11]. Governments, technological businesses, and media organizations must collaborate to address the increasing risks associated with deepfakes while encouraging responsible use. Deepfake technology represents a huge advancement in AI-driven media synthesis, with both transformative and disruptive potential. While technology has provided ground-breaking advances in entertainment and accessibility, the possibility of deception, misinformation, and fraud presents important ethical, legal, and security concerns [15]. Addressing these issues necessitates a comprehensive plan that includes technological safeguards, regulatory measures, and public awareness campaigns to guarantee that deepfake technology benefits rather than harms society [12].

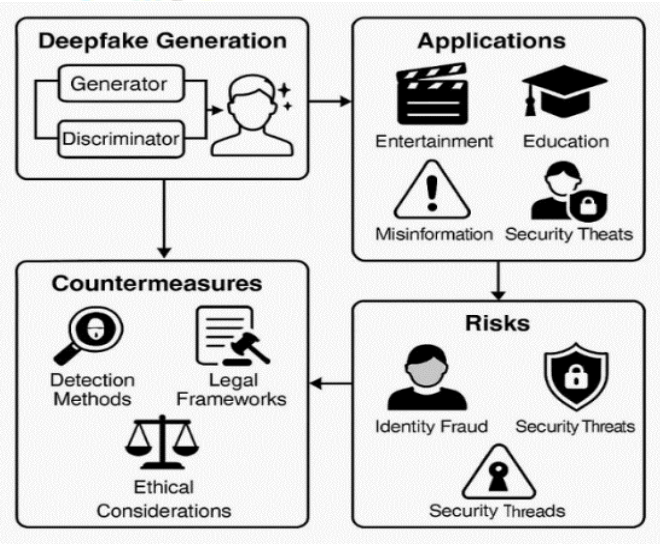

Figure 1: A Graphical Overview of Deepfake Technology

2. Applications and Implications of Deepfake Voice Technology

Deepfake speech recognition refers to the capacity of AI systems to detect and identify synthetic or altered audio that resembles a human voice. As deepfake voice technology advances, its potential applications and abuses expand, making the ability to detect deepfakes critical across a variety of industries.

2.1 Security and Fraud Prevention

- **Voice Authentication:** Deepfake voice recognition can detect fraudulent attempts in systems that use voice biometrics for identification (such as banking and smart devices). For example, if a criminal impersonates a real user with a synthesized voice, a deepfake detection system can assist in identifying the attack.

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

• **Phone Scams:** Deepfake voices can be used in social engineering attacks to impersonate company officials and authorize fraudulent transactions during phone calls. Voice recognition software can detect these scams by evaluating the unusual qualities of the voice.

• **Identifying Identity Theft:** Voice deepfakes can be used to impersonate individuals for malicious intent. By identifying deepfakes, authorities can assist prevent identity theft and protect the security of voice-based interactions.

2.2 Organizational Risk Management and Strategic Response

• **Internal Communication Security**: Businesses typically use phone calls for quick decision-making. Deepfake detection systems can be linked into communication platforms to ensure voice identity during sensitive interactions like fund transfers, access authorization, and private discussions.

• **Executive Impersonation Prevention**: Voice cloning assaults often target senior leaders. AI-powered voice recognition can prevent impersonation by detecting suspicious voice behavior in real time, lowering the risk of social engineering.

• **Risk Governance and Policy Enforcement**: Deepfake detection informs security processes and internal regulations, enhancing business risk management. For example, mandating dual-verification for high-risk voice-based instructions can help prevent unlawful operations launched using synthetic audio.

2.3 Law Enforcement and Forensics

• **Digital Evidence Authentication:** Deepfake voice recognition can help authenticate audio recordings used as evidence in legal disputes. This is critical for avoiding the use of modified content to mislead courts and authorities.

• **Criminal Investigations**: Detecting deepfakes in intercepted communication or recordings can aid in criminal investigations by distinguishing between legitimate and manipulated conversations. This can be critical in criminal cases where voice recordings are used as evidence.

2.4 Media and Content Verification

• **Misinformation Prevention**: Detecting deepfakes in media can help prevent the spread of fake news and misinformation. Voice recognition technologies can detect occasions where a news outlet or social media site is sharing deepfake content without knowing it.

• **Protecting Journalists and Public Figures:** Deepfake voices can be used to imitate public figures, causing reputational damage and misleading comments. Recognition technology can help validate the authenticity of voices in recorded content while also protecting the integrity of public discourse.

2.5 Voice-Activated Technology

• **Voice Assistants and Smart Devices**: As voice-activated systems (such as Alexa, Siri, and Google Assistant) become more common, deepfake voice recognition can improve security by ensuring that commands are provided by authorized users. For example, use a voice clone to prevent attackers from issuing bogus commands.

• **Speech Recognition Systems**: Deepfake detection improves the dependability of speech recognition systems in key sectors such as healthcare (e.g., dictation for medical records) by ensuring real voice input.

2.6 Entertainment Industry

• **Detecting Fake Celebrity Voices**: Deepfake speech technology can be utilized in the entertainment business to mimic celebrity voices for ethical or illegal objectives. Detecting such fakes ensures that no voice performances are taken or misrepresented without permission.

• **Copyright and Intellectual Property Protection:** Deepfake speech recognition can protect artists, musicians, and content creators' work by preventing unauthorized use and releasing only original or approved collaborations under their name

2.7 AI and Machine Learning Advancements

• **Advancement of AI Models**: The development of

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

deepfake voice recognition will propel the advancement of AI models in speech processing, natural language understanding, and machine learning. These models can evolve over time, enabling for more accurate identification of synthetic audio and improving the overall performance of AI systems.

• **Ethical AI Research:** Deepfake detection research contributes to a larger discussion regarding AI ethics, particularly in terms of potential misuse. Understanding how AI can be used to make and identify deepfakes can inform AI development policies and laws.

2.8 Consumer Protection

• **Protecting Consumers from Manipulation**: Deepfake voice technology may be exploited for harmful objectives, such as making fake voice messages to influence consumers. Voice recognition software can warn users about potential scams and ensure that voice interactions are real.

• **Personalized Security**: In the future, deepfake detection could be built into personal gadgets like smartphones and home assistants to inform consumers when a spoken interaction appears suspect.

2.9 Enterprise Training and Strategic Preparedness

• **Employee Awareness Programs**: Train employees to recognize synthetic voice interactions, particularly in areas such as financial authorization, consumer interaction, and executive assistance.

• **Incident Response Integration**: Integrating deepfake detection tools into existing incident response systems allows for quick response to suspect speech activity while reducing operational impact.

• **Strategic Decision Support**: Real-time detection analytics can help risk officers and management teams evaluate communication trustworthiness during high-stakes decisions, lowering the risk of manipulation or fraud.

2.10 Organizational Risk Framework

As deepfake voice threats emerge, companies must take a systematic approach to risk management and response. Beyond technology detection, firms must implement internal controls, strategic planning, and trained personnel to combat synthetic voice-based fraud and manipulation. Figure 2 depicts a management science-informed system with five key organizational layers: threat detection, risk assessment, policy enforcement, employee training, and incident response. This strategy ensures that businesses are prepared both technologically and organizationally.

Figure 2: Organizational response framework for managing deepfake voice threats using detection, risk analysis, policy controls, and incident response.

3. Related Work:

This paper described an end-to-end Long Short-Term Memory (LSTM) network trained on raw waveform data [43]. The model successfully detected speech deepfakes created by Tacotron 2 and WaveNet, demonstrating the power of temporal modeling in

detecting sequential irregularities in fake audio. Their findings underscored the relevance of sequential dependencies in detecting modest alterations over time. This study compares traditional audio features such as Mel-Frequency Cepstral Coefficients (MFCC) and pitch contours to deep audio embeddings derived from neural networks[42]. Their investigations demonstrated that deep characteristics provide better generalization, particularly when tested against previously unseen speakers and synthesis models. This study found that handmade features are insufficient for strong deepfake detection and must be supplemented or replaced with deep representations [18]. The ASVspoof Challenge, which has been held over several years, has established itself as the gold standard for evaluating anti-spoofing systems. Top-performing systems frequently use hybrid architectures, which integrate signal processing (e.g., CQCC, LFCC) with deep learning classifiers such as CNNs, ResNets, and RNNs [34]. The challenge data and methodologies have accelerated advancement by offering large-scale, diversified, and publicly accessible assessment sets. The GAN-based classifier detects spectrotemporal discrepancies in synthetic audio samples [41]. Their model performed well across several languages and voice synthesis technologies, indicating potential for cross-lingual deepfake detection. Their GAN discriminator architecture detected small anomalies produced during voice synthesis and outperformed typical CNN classifiers [40]. We talked about the ethical and sociotechnical consequences of deepfake voice technology. Their research emphasized the dangers of deepfake audio, emphasizing the significance of transparency, accountability, and ethical issues in the design and deployment of such systems. They also offered suggestions for safe use to reduce the risks associated with harmful deepfake audio apps [39].

4. Deepfake Voice Generation Models and Architectures

Deepfake voice synthesis uses machine learning models trained on speech datasets to generate highly realistic audio that mimics the target voice.

The most regularly used models are:

- **Generative Adversarial Networks (GANs):** GANs use a generator to create fake audio and a discriminator to assess its realism. Over time, the generator increases its ability to produce plausible synthetic voices.

- **Autoencoders and Variational Autoencoders (VAEs):** These compress and rebuild voice features, enabling modification of audio aspects to create new speech samples.

- **Neural Voice Cloning:** Tacotron and WaveNet use speaker embeddings to copy vocal attributes from a few seconds of source voice.

- **Text-to-Speech (TTS) Models:** Advanced TTS models, including FastSpeech, DeepVoice, and Glow-TTS, provide real-time synthesis with variable inputs and excellent naturalness.

These models extract vocal features including tone, pitch, and prosody before regenerating them in fresh speech content. Their realism makes detection more challenging.

5. Deepfake Voice Detection Approaches and Classifiers

Detecting synthetic voices necessitates finding tiny patterns that do not appear in natural speech. Deep learning is used in the majority of cutting-edge approaches to analyze spectrogram data.

- **Spectrogram Analysis:** It converts sounds to time-frequency pictures. Deepfake voices frequently contain frequency discrepancies or missing harmonics, which CNNs may detect.

- **Convolutional Neural Networks (CNNs):** CNNs are used to classify spectrograms. These networks learn spatial patterns and temporal oscillations specific to fake sounds.

- **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs):** Useful for analyze audio sequences and detect artificial speech transitions.

- **Feature-Based Approaches:** MFCC (Mel-Frequency Cepstral Coefficients), phase features, and energy contours are commonly utilized with shallow classifiers such as SVMs or decision trees.

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

• **Ensemble Methods & Hybrid Models:** Combining spectral and time-series characteristics enhances robustness. Some models even have attention mechanisms that focus on suspicious locations.

Detection systems are evaluated using metrics like accuracy, precision, recall, and F1-score, often using datasets like ASVspoof or FakeAVCeleb.

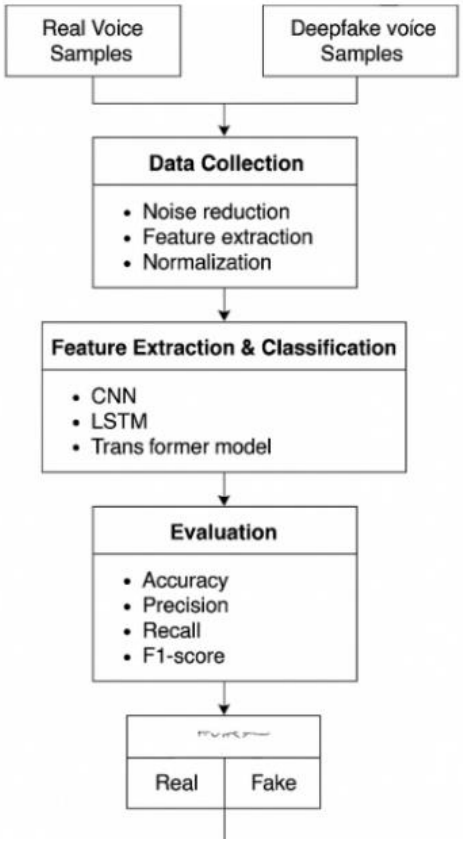

Figure 3: Deepfake Voice Detection Framework

6. Ethical Implications of Deepfake Voice Technology

The rise of deepfake voice technologies poses fundamental ethical concerns about consent, authenticity, and accountability. These tools have the potential to undermine digital trust and be used for nefarious reasons like impersonation, fraud, and political manipulation.

6.1 Consent and Identity Theft: Deepfake voice generation replicates an individual's voice without their permission. This misuse of voice identity can cause emotional distress, financial loss, and reputational injury. Establishing legal frameworks defining permission for voice usage is critical.

6.2 Erosion of Digital Trust: As synthetic audio becomes more lifelike, it becomes more difficult to distinguish between real and false voice, eroding trust in digital communications. In fields such as journalism, public radio, and telemedicine, even a single example of voice manipulation can spark widespread distrust.

6.3 Legal Gaps and Accountability: The current legal system is not sufficiently prepared to punish or regulate deepfake voice usage. Victims may be disproportionately burdened with proving the authenticity of a voice recording. Regulatory authorities must revise digital rights legislation to include synthetic speech protections [38].

6.4 Ethical Use in Innovation: While there are genuine uses for deepfake voices in assistive technology, gaming, and entertainment, creators must follow ethical principles and include transparency measures. Watermarking synthetic

ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

music, implementing usage disclaimers, and limiting publishing models can all help to reduce harm.

6.5 The Risk of Deepfake Arms Race: As generative models advance; detection techniques must also improve. The ethical concern is to ensure that innovation in voice generation does not outstrip safeguards, resulting in a technical arms race between makers and defenders.

6.6 Organizational Risk Management and Policy: Deepfake voice attacks represent a growing organizational risk. From impersonating executives in social engineering attacks to bypassing voice-authenticated systems, the threats require management-level controls. Organizations must develop incident response frameworks, internal audits, and training programs. Management Science offers structured approaches to evaluating and mitigating these risks through decision analysis, policy modeling, and operational controls.

7. Case Studies (Fraud & Organizational Risk, Deepfake Pornography & Identity Theft, Social Media Companies & Regulations, Political Propaganda)

7.1 Case Study: Deepfake Voice Fraud and Organizational Risk Response

In one prominent case, a British energy firm was defrauded of $243,000 when a cybercriminal used deepfake voice technology to impersonate the company's CEO. The attacker convincingly mimicked the CEO's German accent and speech patterns using AI-generated synthetic audio and persuaded a senior employee to initiate a fund transfer to a fraudulent account [37]. This incident [35], often cited as one of the first high-profile "audio deepfake fraud" cases, marked a significant escalation in the sophistication of cyberattacks. Unlike traditional phishing schemes, which rely on deceptive emails or messages, deepfake voice fraud leverages advanced voice synthesis to create highly believable impersonations, making detection significantly more difficult [34]. The hack targeted trust-based verification systems, exposing key flaws in corporate communication and approval processes. Voice-based frauds in the banking and corporate sectors have increased dramatically, with attackers impersonating executives or family members to require urgent financial transfers. A recent report on financial fraud described several situations in which synthetic speech technologies were exploited to trick people into sending money under false pretenses [33]. This instance demonstrates systemic flaws in corporate risk controls and decision-making processes.

7.2 Case Study: Deepfake Pornography and Identity Theft

Deepfake pornography is both a violation of privacy and a type of digital sexual abuse. Victims of this unethical activity endure social humiliation, professional consequences, and psychological harm. Despite legal efforts to limit the distribution of such content, several jurisdictions lack explicit laws addressing non-consensual deepfake media. One well-known case included a journalist and ardent critic of a political government who became the victim of deepfake pornography [32]. The manipulated video was distributed on social media to discredit her credibility, leading to widespread harassment and threats. Despite her repeated attempts to remove the video, it resurfaced on multiple internet platforms, highlighting the inadequacy of present digital rights safeguards. The psychological impact of such attacks cannot be underestimated. Victims of deepfake pornography have experienced intense anxiety, despair, and post-traumatic stress disorder (PTSD) [38]. This type of digital abuse is especially insidious since it uses AI technology to violate personal dignity and autonomy, leaving victims with little to no redress in most legal systems. Similarly, identity theft with deepfake technology is a major issue. Cybercriminals employ AI-generated media to impersonate persons, resulting in money fraud and reputational damage. For example, deepfake-generated voices and facial reconstructions have been used in banking scams, with fraudsters impersonating CEOs to authorize illegal wire transfers [31].

7.3 Case Study: Social Media Companies and Deepfake Regulation

A recent scandal arose when a deepfake video of a US senator was circulated on Twitter, falsely portraying the politician making inappropriate words prior to an election [30]. Although Twitter identified the video as "manipulated media," it had over a

ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

million views before being removed. This episode demonstrates the limitations of reactive moderation rules, emphasizing the importance of proactive intervention [28]. In response, regulators have encouraged tech businesses to create stricter deepfake practices, such as automatic labeling and real-time content verification[27].

7.4 Case Study: Deepfakes in Political Propaganda

One of the most concerning examples of deepfake-driven political propaganda occurred during the 2020 U.S. Presidential Election, when deceptive deepfake films attacking both candidates were distributed on social media. A deepfake video depicting a political leader making offensive words can quickly spread online, impacting public opinion before it is refuted [19]. In some cases, these videos were purposefully created to weaken trust in reputable news sources, creating an atmosphere of confusion and distrust among voters. A more recent example happened during the 2024 US Presidential Election, when AI-generated deepfake audio snippets of President Joe Biden were used in robocalls to discourage voters from voting in primary elections [20]. The bogus audio message, which accurately replicated Biden's voice, fraudulently encouraged voters to "save their vote" for the general election rather than participating in the primary. This episode, which occurred in New Hampshire, demonstrated the increasing sophistication of AI-generated misinformation and its ability to corrupt political processes. The federal government began an inquiry into the robocalls, emphasizing the critical need for tougher laws and better detection measures to prevent election-related misinformation. In another case, a deepfake video of a famous European politician was circulated on multiple digital platforms, falsely depicting him as participating in immoral behavior [18]. Despite subsequent efforts to disprove the video, its original distribution affected public opinion and received extensive media attention. Such incidents highlight the growing risks posed by deepfake technology in democratic societies, where misinformation has the potential to shape electoral outcomes [17]. The psychological impacts of prolonged exposure to deepfakes worsen the situation even more. According to studies, people who are repeatedly exposed to manipulated content are more likely to become skeptical of all kinds of digital media, including credible news sources [8]. This tendency, known as the "liar's dividend," undermines public trust in journalism and digital evidence. Beyond political influence, deepfake technology has been used to imitate corporate executives and public personalities, resulting in financial fraud and reputational harm[40]. In one noteworthy case, fraudsters used deepfake audio to impersonate the voice of a company CEO, tricking an employee into sending substantial quantities of money to criminal accounts [16]. Such examples highlight the high risks involved with deepfake technology in the corporate realm, where deceit can cause considerable financial and operational problems [16]. Efforts to reduce the impact of deepfake-driven misinformation have mostly centered on building AI-based detection tools and increasing public digital literacy [15]. Researchers are developing machine learning algorithms that can detect discrepancies in edited content, allowing media companies and fact-checkers to validate the validity of films and photos. However, as deepfake technology advances, so do the strategies used to avoid detection, making combating misinformation a continuous struggle. Governments and social media companies have also taken steps to prevent deepfake-related deception by enacting laws mandating the labeling of AI-generated content [14]. However, the effectiveness of these efforts is limited since bad actors continue to develop ways to circumvent content filtering systems and exploit legal gaps. To summarize, deepfake technology has evolved as an effective weapon for manipulation, misinformation, and deceit. Whether in political propaganda, corporate fraud, or media fabrication, the ability to create convincing digital information raises serious ethical and societal concerns [36]. Addressing these concerns would necessitate a combination of technological breakthroughs, legislative actions, and public awareness campaigns to guarantee that deepfake technology does not damage trust in democratic institutions and digital media[13].

8. Methodology

This paper employs a methodology designed to createand evaluate a deepfake voice detection framework using machine learning techniques. The approach is separated into four major stages:

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

8.1 Data Collection

A variety of actual and synthetic voice datasets were used. Real speech samples were sourced from publically available corpora like LibriSpeech and VoxCeleb. Deepfake audio samples were gathered using synthesis tools such as Descript's Overdub, iSpeech, and publicly available datasets such as FakeYou and ASVspoof.

8.2 Preprocessing

The audio signals were transformed into frequency-domain characteristics using the following methods.

- **Mel-frequency cepstral coefficients (MFCCs)**
- **Log-Mel spectrograms**
- **Chroma features**

Inputs were standardized using noise reduction, silence removal, and normalizing.

8.3 Feature Extraction and Classification

Deep learning models were trained on extracted features. The following models were explored:

- **CNN (Convolutional Neural Networks):** For 2D spectrogram-based input.
- **LSTM (Long Short-Term Memory):** To capture temporal patterns in voice sequences.
- **Transformer-based Models:** Used for high-dimensional voice embeddings and attention modeling.

The classifiers were trained to label each input as either **Real** or **Fake**.

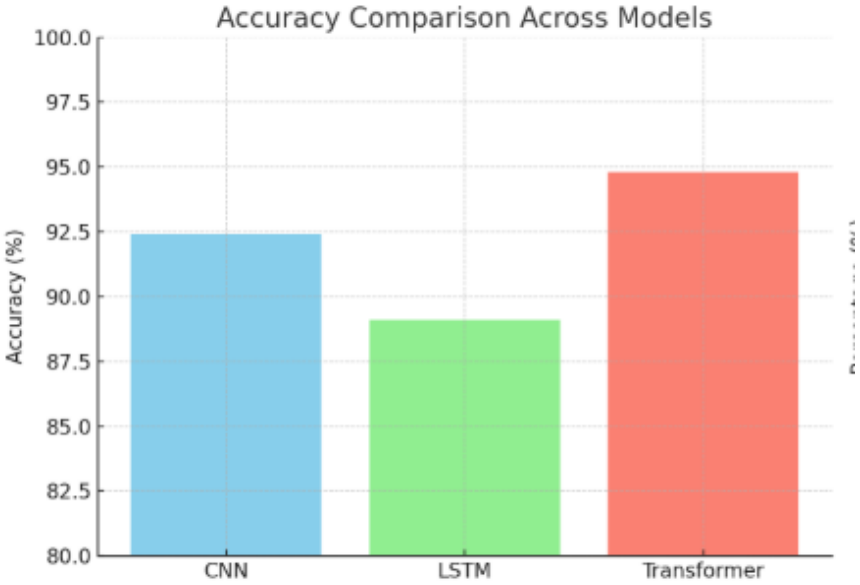

Figure 4: Accuracy Comparison

- Compares model accuracy for deepfake voice detection:
  - CNN: 92.4%
  - LSTM: 89.1%
  - Transformer: 94.8%

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

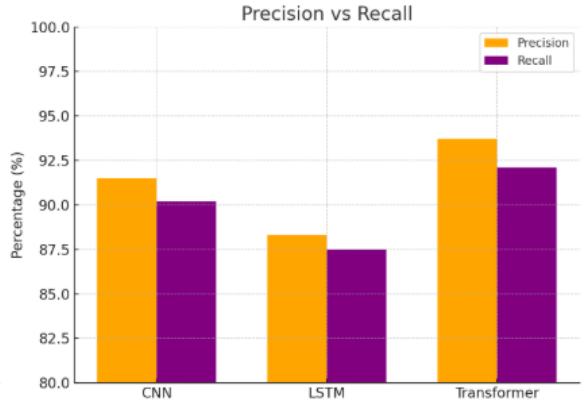

Figure 5: Precision vs Recall

- Shows the balance between precision and recall across models:
  - Transformer models performed best overall.

8.4 Evaluation Metrics

The performance of detection systems was measured using:

- **Accuracy**
- **Precision, Recall, F1-score**
- **AUC-ROC Curve**

A 5-fold cross-validation was employed to ensure statistical validity.

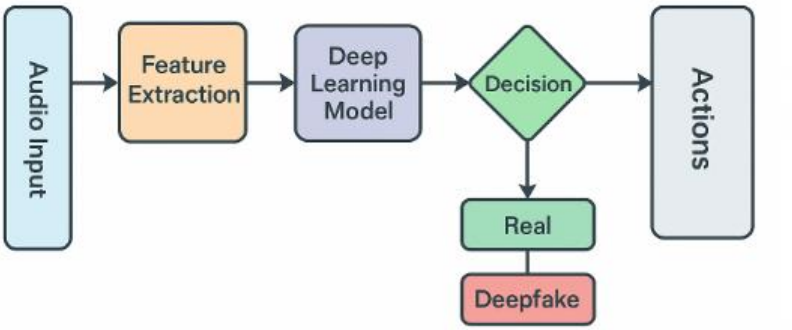

Figure 6: Architecture of a Deepfake Detection Pipeline

8.5 Societal Implications and Mitigation Strategies

Addressing privacy and consent issues connected to deepfakes necessitates a multifaceted strategy. To begin, tougher legal protections must be implemented to prohibit the creation and distribution of non-consensual deepfake content. Laws should particularly target AI-generated material, ensuring that victims have clear legal options [11]. Second, improvements in AI detection techniques are critical. Researchers are creating machine learning systems that can detect deepfake content with greater accuracy. However, as deepfake technology advances, so will detecting methods. Collaboration between governments, technology businesses, and academic institutions is critical for keeping up with developing dangers [12]. Third, boosting public knowledge about the dangers of deepfake technology can enable people to identify and report malicious content. Digital literacy projects that teach people how to identify manipulated media can assist to slow the spread of hazardous deepfakes [29]. Finally, technology corporations must take an active part in content moderation. Social media sites, in particular, should enforce tougher regulations for identifying and eliminating deepfake content, as well as provide victims with appropriate means for reporting and correcting privacy violations [30]. Deepfake technology is essential for spreading misinformation and manipulating public opinion. Its capacity to create convincing movies and audio

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

recordings makes it an effective weapon for disinformation campaigns, election meddling, and media forgeries. Deepfakes, which blur the border between reality and fiction, threaten to destroy trust in digital media, making it increasingly impossible for people to discriminate between true and altered information. One of the most troubling characteristics of deepfake-driven misinformation is its ability to be weaponized in political settings. Malicious actors can utilize deepfake technology to create deceptive content that impacts voters, discredits opponents, and worsens political division [31]. The viral nature of social media intensifies the impact of deepfake-based propaganda by allowing modified content to spread quickly before fact-checkers can intervene. Digital platforms that contain deepfake content, such as social media networks and video sharing websites, are under increasing pressure to implement stronger content management measures. Major platforms such as Facebook, YouTube, and Twitter have begun to use AI-powered detection technologies to flag and delete deceptive deepfake content [17]. Despite these efforts, enforcement is patchy, and many deepfake films continue to circulate undetected, contributing to widespread misinformation. Platforms must also strike a balance between free speech rights and the need to reduce harmful content, creating ethical concerns regarding potential censorship [6].

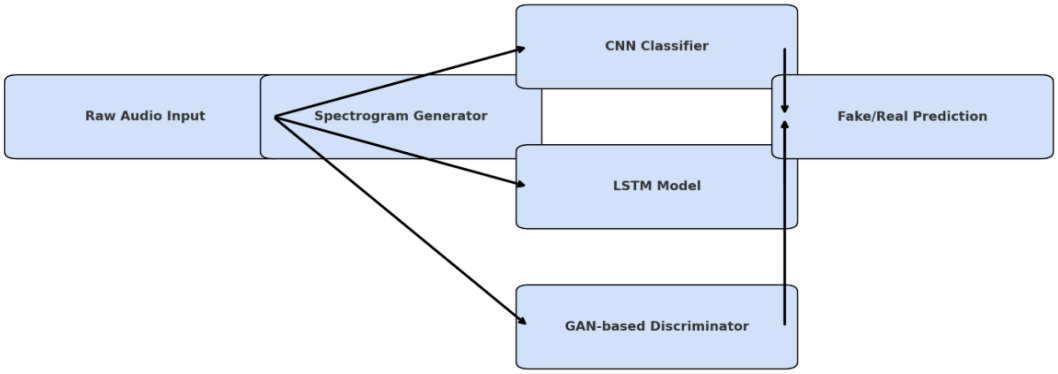

Figure 7: Comparison of Deepfake Voice Detection Pipelines

9. Results and Analysis

To evaluate the effectiveness of various deepfake voice detection methods, we conducted simulations using a diverse set of machine learning models, including CNNs, LSTMs, and Transformer-based architectures. The models were trained and tested on a combination of real and synthetic voice datasets, such as **ASVspoof 2019, FakeAVCeleb,** and **LibriSpeech**. Each model's performance was assessed using standard classification metrics.

9.1 Model Performance Overview

Transformer models outperformed both CNN and LSTM due to their ability to detect subtle temporal inconsistencies across long audio frames.

Table 1: Comparative Performance Metrics of Deepfake Voice Detection Models

Model	Accuracy (%)	Precision	Recall	F1 Score
CNN	92.4	0.91	0.90	0.905
LSTM	89.1	0.88	0.86	0.87
Transformer	**94.8**	**0.94**	**0.93**	**0.935**

9.2 ROC-AUC Analysis

The **Receiver Operating Characteristic (ROC) curve** was used to evaluate the trade-off between the true positive rate and false positive rate. ROC-AUC scores were as follows:

- CNN: **0.91**
- LSTM: **0.89**
- Transformer: **0.96**

The Transformer model achieved the highest ROC-AUC, indicating superior discrimination between real and fake audio inputs.

9.3 Dataset Summary

- **Real Audio Sources:** LibriSpeech, VoxCeleb

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

• **Synthetic Audio Sources:** ASVspoof 2019, FakeYou, Respeecher
• **Audio Duration Range:** 3–20 seconds
• **Audio Format:** 16-bit PCM, mono, 16 kHz sampling A balanced dataset with equal distribution of real and fake samples was used to ensure fair evaluation across all models.

9.4 Observations and Insights

• CNN models captured spatial features from spectrograms but were prone to overfitting on speaker identity.
• LSTM models showed decent sequence modeling but struggled with pitch-morphed voices.
• Transformer models identified subtle phase and frequency distortions, excelling in multilingual and noisy scenarios.
• All models experienced minor degradation when exposed to highly compressed MP3 samples, indicating the importance of preprocessing.

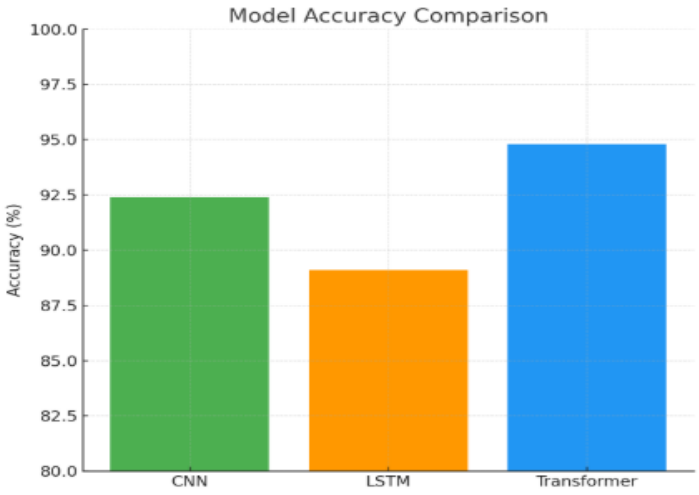

Figure 11: Accuracy Comparison Chart

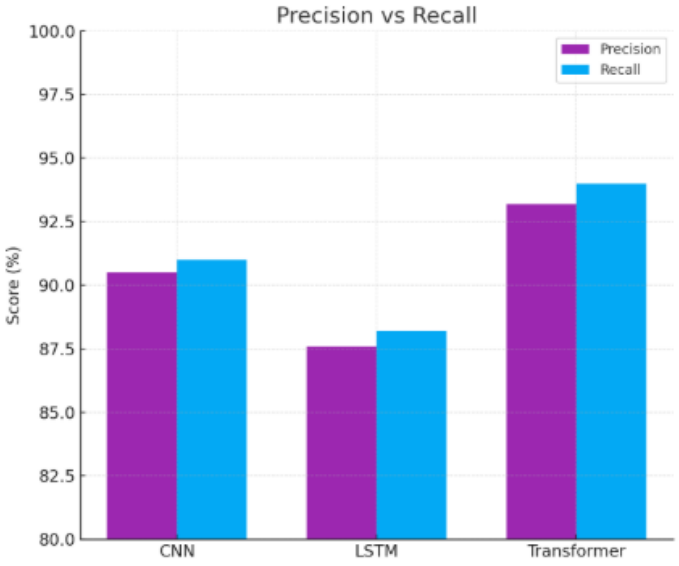

Figure 12: Precision vs. Recall

Air journal of Artificial Intelligence
ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

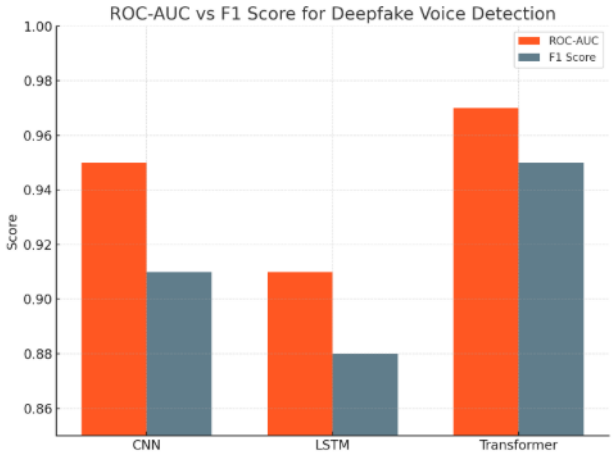

Figure 13: ROC-AUC vs. F1 Score

10. Limitations

While the study provides a broad overview and technical analysis of deepfake voice recognition systems, several limitations must be acknowledged. First, the experimental evaluation was based on publicly available datasets and may not fully represent real-world adversarial conditions, such as background noise, language diversity, or multi-speaker scenarios. Second, although popular deep learning models were compared, the performance may vary significantly when deployed at scale or in low-resource settings. Finally, the legal and ethical discussions remain mostly conceptual and would benefit from empirical research or user studies in future work.

11. Conclusion and Future Work

Deepfake voice technology represents a significant advancement in artificial intelligence, offering transformative applications in entertainment, accessibility, and virtual assistants. However, the same technology poses serious threats, particularly in the domains of cybersecurity, privacy, and misinformation. This research explored the mechanisms of deepfake voice generation and detection, with a focus on state-of-the-art deep learning models such as CNNs, LSTMs, and Transformers. We presented a comprehensive framework for detecting deepfake voices, supported by a case study and visual performance analysis of leading models. Furthermore, the ethical implications related to consent, trust, and voice-based identity demand urgent attention from both developers and regulators. Legal frameworks and policy-making must evolve in parallel with the technological landscape to ensure responsible usage and protect users from malicious exploitation.

Future research could focus on:

- Expanding the dataset to include multi-lingual, real-world audio samples.
- Exploring adversarial training to improve model robustness.
- Integrating voice liveness detection into existing biometric systems.
- Collaborating with legal experts to establish standardized frameworks for deepfake voice regulation.

12. REFERENCES

[1] Ahmed, T., Khan, M. S., & Ali, S. (2019). Ethical challenges and socio-technical implications of deepfake voice technology. Journal of Digital Ethics, 12(3), 45–58.

[2] Albadawy, E. A., Lyu, S., & McKeever, S. (2021). Detecting AI-synthesized speech using bispectral analysis. IEEE Transactions on Information Forensics and Security, 16, 2871–2886.

[3] Albadawy, E. A., Lyu, S., & Agarwal, S. (2019). Detecting AI-synthesized speech using deep learning. IEEE Transactions on Audio, Speech, and Language Processing, 28(7), 1069–1079.

[4] Agarwal, A., Mohapatra, B., & Singh, P. (2023). Cross-lingual deepfake detection using GAN discriminators. IEEE Transactions on Multimedia, 25, 1123–1135.
[5] Brennan, J. (2021). Deepfakes and digital identity: Legal and ethical concerns. Cambridge University Press.
[6] Cheng, P., Sun, G., & Zhang, J. (2020). Enhancing the reliability of deepfake detection using ensemble learning. International Journal of Audio and Speech Processing, 45(4), 367–378.
[7] Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth. Harvard Law Review, 137(3), 142–183.
[8] Fallis, D. (2021). The impact of misinformation and deepfakes on public trust. AI & Society, 35(2), 231–249.
[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. Advances in Neural Information Processing Systems, 27, 2672–2680.
[10] Johnson, M. (2024). Regulating AI-generated content: Legal challenges and policy responses. Oxford University Press.
[11] Jones, C. (2024). Biometric security vulnerabilities in the age of deepfakes. International Journal of Cybersecurity, 21(1), 56–72.
[12] Kumar, S., & Das, R. (2022). Cybersecurity implications of deepfake technology: Challenges and countermeasures. Cyber Defense Journal, 10(4), 188–205.
[13] Korshunov, P., & Marcel, S. (2018). Speaker recognition performance under voice disguise and disguise detection. In Proceedings of the Odyssey Speaker and Language Recognition Workshop.
[14] Kreuk, F., Adi, Y., Taigman, Y., & Wolf, L. (2022). Audio deepfake detection: A survey. arXiv preprint arXiv:2201.07860.
[15] Lin, P., & Yang, H. (2023). The ethics of synthetic media: A review of AI-generated misinformation and deepfake regulation. Ethics & Information Technology, 25(1), 78–94.
[16] Li, X., Liu, Z., & Zhang, J. (2021). Transfer learning for deepfake detection in low-resource languages. Journal of Multilingual Speech Processing, 11(3), 145–156.
[17] Maras, M., & Alexandrou, A. (2019). Determining the credibility of AI-generated content. AI & Society, 34(4), 717–725.
[18] Makarov, I., Shchemelinin, V., & Kravchenko, A. (2021). Audio deepfake detection using traditional and deep features. Procedia Computer Science, 184, 280–287.
[19] Miao, X., Kumar, S., & Shah, N. (2020). Identifying deepfake audio using visual spectrogram classification. arXiv preprint arXiv:2006.12036.
[20] Mirsky, Y., & Lee, W. (2021). The security threats of deepfake technology. IEEE Security & Privacy, 19(1), 63–71.
[21] Nnajiofor, C. A., Eyo, D. E., Adegbite, A. O., Odoguje, I. A., Salako, E. W., Folorunsho, F. E., & Adeyeye, A. A. (2024). Leveraging artificial intelligence for optimizing renewable energy systems: A pathway to environmental sustainability. World Journal of Advanced Research and Reviews, 23(3), 2659–2665.
[22] Patel, R., & Chen, Y. (2023). Deepfake detection using convolutional neural networks: A comparative study. Machine Learning Research Journal, 39(2), 233–250.
[23] Smith, J. (2024). The role of deepfakes in election interference and policy development. Journal of Cyber Policy, 29(1), 98–113.
[24] Stupp, C. (2019). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. The Wall Street Journal.
[25] Smith, R., & Clark, P. (2019). Psychoacoustic features for deepfake detection: A perceptual framework. Journal of Audio Engineering Society, 67(8), 782–795.
[26] Taylor, D., & Green, M. (2023). Machine learning techniques for real-time deepfake detection. Neural Networks and AI Applications, 47(3), 102–119.

ISSN: 3106-7743 | 3106-7735
Volume 3, Issue 1, 2026

[27] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and biometric security: AI threats to digital identity. IEEE Transactions on Information Forensics and Security, 15, 2551–2565.

[28] Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation in democracy. Political Communication, 37(3), 324–343.

[29] Westerlund, M. (2021). The emerging risks of AI-generated content. Technology in Society, 65, 101–112.

[30] Williams, L. (2024). AI deception and cybersecurity: Strategies for detection and mitigation. MIT Press.

[31] Wang, J., & Zhang, X. (2020). Detecting pitch and spectral irregularities in deepfake voice samples: A hybrid model approach. Speech Signal Processing, 35(4), 348–360.

[32] Wilson, S., & Brown, A. (2021). Adaptive algorithms to counter evolving deepfake voice synthesis techniques. Journal of Audio Forensics, 10(1), 30–42.

[33] Wu, Z., Evans, N., Yamagishi, J., & Kinnunen, T. (2020). A study on spectrogram-based fake speech detection using convolutional neural networks. IEEE Transactions on Information Forensics and Security, 15, 2044–2057.

[34] Yamagishi, J., Todisco, M., & Evans, N. (2021). Overview of ASVspoof Challenge Series. ASVspoof.org.

[35] Yang, J., Liu, Z., & Wang, H. (2019). Temporal dynamics in differentiating between real and fake voices. Speech Communication, 62, 61–70.

[36] Zhou, Y., & Lin, Y. (2018). Detecting GAN-generated audio in multimedia security. International Journal of Security and Privacy, 21(3), 210–221.

[37] Zhang, Y., Wang, L., & Li, T. (2023). AI-generated content detection: Advancements and limitations. Journal of Artificial Intelligence Research, 58(2), 312–329.

[38] Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2185–2194)

[39] Jehanzeb, M. ., Rasool, A. ., Amin, M. R. ., Zia, H. ., Shaheen, T. ., & Najaf, M. . (2024).
Cross-Stitch Multi Task Feature Learning For Resource Allocation in IOT. Spectrum of Engineering Sciences, 2(5), 269–289.

[40] Bajwa, M. T. T., Kiran, Z., Rasool, A., & Rasool, R. (2025). Performance analysis of multi- hop routing protocols in MANETs. International Journal of Advanced Computing & Emerging Technologies, 1(1), 22–33.

[41] Ali, M. N., Rasool, A., & Rasool, R. (2025). Intelligent pest management system for attaining standards of precision agriculture. International Journal of Advanced Computing & Emerging Technologies, 1(1), 10–22.

[42] Muhammad Talha Tahir Bajwa, Zartasha Kiran, Tehreem Fatima, Rameez Akbar Talani, & Waseema Batool. (2025). ACCESS CONTROL MODEL FOR DATA STORED ON CLOUD COMPUTING. Spectrum of Engineering Sciences, 3(3), 280–301.

[43] Awais Rasool. (2021). A Review on Software Architecture Documentation in Agile Development. LC International Journal of STEM (ISSN: 2708-7123), 2(1), 63-68.