IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

Real Fake Audio Detection

Prof.Smita Mulhar, Mansi Dhavan, Nisha Adhav, Neha Kadam, Vedant Jedhe
Department of Computer Engineering
Smt. Kashibai Navale college of Engineering, Vadgaon, Pune, India
smitamulhar_skncoe@sinhgad.edu, mansidhavan64@gmail.com
adhavnisha21@gmail.com, nehakadam1508@gmail.com, vedantjedhe05@gmail.com

Abstract: *The rapid progress of Artificial Intelligence (AI) has enabled the creation of highly realistic synthetic voices, commonly referred to as deepfake audios. These technologies have demonstrated great potential in several creative and assistive fields such as entertainment, speech synthesis for the disabled, and virtual assistants. However, their misuse has raised serious concerns regarding security, privacy, and information authenticity. Malicious individuals can exploit deepfake audio to conduct fraud, impersonate individuals, or spread misinformation, leading to a decline in public trust in digital communications and media. Traditional supervised learning models have shown limited capability in addressing this issue effectively. Such models rely heavily on labeled datasets, which makes them inefficient when faced with unseen or newly emerging manipulation techniques. As the sophistication of deepfake generation methods increases, these conventional models struggle to detect subtle anomalies in manipulated audio, resulting in reduced accuracy and adaptability.To overcome these challenges, the proposed system introduces a Generative Adversarial Network (GAN)- based anomaly detection framework. The system works in an unsupervised manner, focusing on learning the intrinsic patterns and natural characteristics of authentic human speech. By capturing and modeling these genuine audio features, the system can identify any deviation or abnormality present in a test audio sample. Such deviations are then classified as potential deepfake or manipulated audio. The proposed approach significantly enhances robustness, adaptability, and generalization compared to traditional detection systems. Since it learns from real audio rather than depending on pre-labeled fake samples, it becomes capable of detecting both known and novel deepfake audios. Furthermore, the framework's design ensures scalability, making it suitable for real-time audio verification across multiple environments, accents, and languages. This research thus provides a powerful solution to one of the most critical challenges in today's digital era—ensuring the integrity and authenticity of voice data. The integration of GANs with anomaly detection techniques offers a dynamic and intelligent mechanism for distinguishing between real and fake voices. The system not only strengthens digital communication security but also contributes to safeguarding individuals and organizations from audio-based misinformation, identity theft, and fraudulent activities. In summary, the proposed GAN-based deepfake audio detection framework serves as an effective and scalable approach for real-time identification of synthetic voices, thereby enhancing trust and reliability in modern communication systems. The rapid advancement of Artificial Intelligence (AI), particularly in deep learning and speech synthesis, has led to the emergence of highly realistic synthetic audio known as deepfake audio. Leveraging sophisticated techniques such as neural voice cloning and text-to-speech models, these systems can generate human-like speech that is often indistinguishable from genuine audio. While such innovations offer significant benefits in domains like entertainment, assistive technologies for speech- impaired individuals, virtual assistants, and content creation, they also introduce critical challenges related to security, privacy, and trustworthiness. The misuse of deepfake audio technologies has become a growing concern in recent years. Malicious actors can exploit these tools for impersonation attacks, financial fraud, political manipulation, and the spread of misinformation. This increasing threat undermines public confidence in digital media and highlights the urgent need for robust and reliable detection mechanisms. Traditional*

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

supervised learning approaches for deepfake detection rely heavily on large volumes of labeled datasets containing both real and fake samples. However, these methods face limitations in scalability and adaptability, especially when encountering previously unseen or evolving deepfake generation techniques. As a result, their performance degrades in real- world scenarios where new attack patterns continuously emerge. To address these limitations, this research proposes a novel deepfake audio detection framework based on Generative Adversarial Networks (GANs) combined with anomaly detection techniques. Unlike conventional supervised models, the proposed system operates in an unsupervised manner by learning the underlying distribution and intrinsic characteristics of authentic human speech. The GAN architecture, consisting of a generator and a discriminator, is trained exclusively on real audio data to capture subtle temporal and spectral features such as pitch, tone, rhythm, and frequency patterns. During the detection phase, the model evaluates incoming audio samples by comparing them against the learned representation of genuine speech. Any deviation from these learned patterns is identified as an anomaly, which may indicate the presence of manipulated or synthetic audio. This approach enables the system to detect both known and previously unseen deepfake audio with improved accuracy and robustness.

Furthermore, the proposed framework demonstrates strong generalization capabilities across diverse datasets, languages, and speaker variations. Its unsupervised nature eliminates the dependency on labeled fake datasets, making it highly scalable and adaptable to emerging threats. The system is also designed to support real-time processing, enabling its deployment in practical applications such as voice authentication systems, call verification platforms, and media forensics.

This research contributes to the development of secure and trustworthy digital communication systems by providing an intelligent and dynamic solution for deepfake audio detection. By integrating GAN-based learning with anomaly detection, the proposed model enhances the reliability of voice-based interactions and protects individuals and organizations from identity theft, fraud, and misinformation.

In conclusion, the proposed GAN-based framework offers an effective, scalable, and future-ready solution for detecting synthetic audio, thereby reinforcing the integrity and authenticity of voice data in the evolving digital landscape.

Keywords: *GAN-based framework*

I. INTRODUCTION

Recent advancements in artificial intelligence (AI), particularly in speech synthesis and generative modeling, have made it possible to produce synthetic voices that are almost indistinguishable from real human speech. These hyper-realistic deepfake audios are created using advanced machine learning models such as Generative Adversarial Networks (GANs) and autoregressive models like WaveNet and Tacotron. While these innovations have significant applications in entertainment, accessibility, and virtual assistance, they have also introduced severe challenges in cybersecurity and information integrity. Deepfake audio can be exploited for fraudulent activities, impersonation, phishing attacks, and the dissemination of misinformation, leading to potential social, economic, and political harm.

Traditional supervised learning-based detection models depend heavily on labeled datasets containing examples of both genuine and fake audio. However, these models often fail to generalize when exposed to new or unseen fake generation techniques. As deepfake synthesis technology evolves rapidly, new types of forgeries continually emerge, rendering static supervised models ineffective. This limitation has motivated the exploration of unsupervised and anomaly detection-based approaches, which focus on learning the intrinsic patterns of genuine human speech and identifying any deviations from these learned norms as potential forgeries.

The proposed system addresses this gap by introducing a GAN- based anomaly detection framework designed to detect deepfake audio in an unsupervised manner. The core idea is to train the model exclusively on real human speech data

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

so that it can learn the underlying distribution of genuine audio features. During inference, if an audio sample significantly deviates from this learned distribution, it is flagged as a potential deepfake. This approach provides a more robust and adaptive solution, capable of detecting both known and novel audio manipulations without requiring explicit examples of fake data during training.

In the proposed methodology, mel-spectrograms and Mel- Frequency Cepstral Coefficients (MFCCs) are used for feature extraction. These representations capture crucial temporal and spectral properties of speech, such as pitch, tone, and timbre, which are often subtly distorted in synthetic audio. By converting raw audio signals into these compact yet information-rich feature spaces, the model can more effectively distinguish authentic human speech patterns from artificial anomalies.

Two specific unsupervised models are utilized—GANomaly and f-AnoGAN. Both models are based on the GAN architecture, which consists of a generator and a discriminator trained in an adversarial setting. GANomaly extends the traditional GAN by incorporating an encoder-decoder structure, enabling it to learn compact latent representations of normal data. During testing, the reconstruction error between the input and the generated output is used as an anomaly score—higher errors indicate greater likelihoods of deepfake audio. Similarly, f-AnoGAN refines this approach by mapping test samples to the latent space and comparing them to the learned manifold of genuine audio.

This framework's strength lies in its ability to generalize across various types of manipulations without prior exposure to fake examples. By modeling the normal speech distribution, the system becomes sensitive to even subtle irregularities introduced during synthesis or manipulation. The use of mel- spectrogram and MFCC features ensures that both time-domain and frequency- domain anomalies are effectively captured.

In conclusion, the proposed unsupervised GAN-based anomaly detection system offers a promising solution to the growing threat of deepfake audio. By focusing on learning genuine speech characteristics rather than detecting known fake patterns, this approach enhances resilience, adaptability, and generalization. As deepfake technologies continue to evolve, such anomaly detection frameworks will play a crucial role in safeguarding digital communication, preserving trust, and ensuring audio authenticity across diverse real-world applications.

Deepfake audio technologies, powered by advanced machine learning architectures such as Generative Adversarial Networks (GANs) and autoregressive models like WaveNet and Tacotron, are capable of producing speech with natural tone, pitch, and emotional expression. These developments have opened new opportunities in domains such as entertainment, assistive technologies for individuals with speech impairments, virtual assistants, and automated content generation.

However, alongside these benefits, deepfake audio has introduced significant challenges in the domains of cybersecurity, privacy, and information authenticity. Malicious actors can exploit synthetic voice technologies for impersonation, financial fraud, phishing attacks, and the spread of misinformation. Such misuse can result in severe social, economic, and political consequences, ultimately eroding public trust in digital communication systems. As deepfake generation techniques continue to evolve, detecting such manipulated audio has become increasingly difficult.

Traditional supervised learning-based detection methods rely heavily on labeled datasets that contain examples of both real and fake audio samples. While effective in controlled environments, these models struggle to generalize when confronted with new or unseen deepfake generation techniques. The rapidly evolving nature of deepfake technology renders static supervised models less effective over time, as they fail to adapt to novel manipulation strategies. This limitation has led researchers to explore unsupervised and anomaly detection-based approaches, which focus on learning the intrinsic characteristics of genuine human speech without requiring labeled fake data.

To address these challenges, the proposed system introduces a GAN-based anomaly detection framework for deepfake audio detection. The system operates in an unsupervised manner by training exclusively on authentic human speech data. This allows the model to learn the underlying distribution and natural patterns of real audio. During the inference phase, any deviation from these learned patterns is considered an anomaly and is flagged as a potential deepfake. This approach enhances robustness and enables the detection of both known and previously unseen audio manipulations.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

For effective feature extraction, the system utilizes mel- spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs). These feature representations capture essential temporal and spectral characteristics of speech, including pitch, tone, and timbre. Since synthetic audio often introduces subtle distortions in these properties, transforming raw audio signals into mel-spectrogram and MFCC representations enables the model to identify inconsistencies more effectively.

The framework incorporates two advanced unsupervised models: GANomaly and f-AnoGAN. GANomaly extends the conventional GAN architecture by integrating an encoder-decoder mechanism, which learns compact latent representations of normal audio data. During testing, the model reconstructs the input audio, and the reconstruction error is used as an anomaly score. A higher reconstruction error indicates a greater likelihood of the audio being fake. On the other hand, f-AnoGAN improves anomaly detection by mapping test samples into the latent space and comparing them with the learned distribution of genuine audio, allowing for more precise anomaly identification.

The key strength of this framework lies in its ability to generalize across diverse manipulation techniques without requiring prior exposure to fake samples. By modeling only the distribution of real speech, the system becomes highly sensitive to even minor irregularities introduced during audio synthesis or manipulation. Additionally, the use of both time-domain and frequency-domain features ensures comprehensive analysis of audio signals.

In conclusion, the proposed unsupervised GAN-based anomaly detection system presents a robust, adaptive, and scalable solution for deepfake audio detection. By focusing on learning genuine speech characteristics rather than relying on predefined fake patterns, the system enhances detection accuracy and resilience against evolving threats. This approach plays a vital role in strengthening digital security, preserving trust in communication systems, and ensuring the authenticity of audio data across various real-world applications.

Technique in Vibe Tracking

The Vibe Tracker follows several techniques from Artificial Intelligence, which are

A. Machine Learning

Machine learning (ML) models optimized for vibe detection allow real time analysis of user emotions. Unlike textual data only, a Vibe Tracker considers multimodal datasets including images, videos, and audio to discern emotional trends. For example, ML models can recognize a user's sentiment from their social media text-based posts, while speech recognition models detect changes in speech, such as reducing their speech rates, extended pauses, or uncharacteristic intonation, that may indicate emotional suffering.

II. RELATED WORK

Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching [1] (2025)

Authors: Emma Coletta, Davide Salvi, Viola Neroni, Daniele Ugo Leonzio, and Paolo Bestagini This paper introduces an advanced approach for detecting and localizing speech deepfakes using a Feature Pyramid Matching (FPM) framework. The key idea of this research is to train the model only on real speech data, so that it can easily detect fake or manipulated audio generated by unknown synthesis techniques. Instead of depending on fake data during training, the system learns to understand the genuine characteristics of human speech and uses this knowledge to identify abnormalities. The proposed method is built on a student–teacher architecture. The teacher model is trained first on real speech to learn its deep and rich features at multiple levels. The student model is then trained to mimic the teacher by reproducing its feature maps. During testing, when a fake or manipulated audio sample is given, the student model fails to perfectly match the teacher's representation. This mismatch between the two networks is measured using a Feature Pyramid Matching technique, which captures inconsistencies at multiple feature scales—from low- level spectral patterns to high-level temporal features. These differences are visualized in the form of anomaly maps in the time–frequency domain. The anomaly maps highlight suspicious regions in the audio where possible manipulation or synthesis has occurred. This allows the model not only to classify an audio clip as fake but also to localize where the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

manipulation exists. Such interpretability is an important advantage because it helps human analysts and forensic experts understand why an audio is considered fake. The paper emphasizes robustness and explainability, showing that training only on real audio improves the system's ability to handle unseen deepfake generation methods. The method avoids overfitting to known fake types and generalizes better to new attacks. Extensive experiments demonstrate that this FPM- based anomaly detection approach achieves high accuracy and strong generalization without using any synthetic samples during training.Overall, this work contributes a significant step forward in trustworthy deepfake detection by combining feature pyramid learning, self-supervised comparison, and visual interpretability, making it a powerful and explainable tool for detecting speech manipulation.

One-Class Learning with Adaptive Centroid Shift for Audio Deepfake Detection [2] (2024)
Authors: Hyun Myung Kim, Kangwook Jang, and Hoirin Kim This paper presents a deepfake audio detection technique using a One-Class Learning (OCL) framework with an Adaptive Centroid Shift (ACS) mechanism. The goal is to build a model that can detect fake or manipulated audio using only real, bona fide speech during training. Traditional deepfake detectors rely on both real and fake samples, but this can limit their generalization to unseen attack types. The authors address this limitation by focusing exclusively on real data to learn a compact and adaptive representation of genuine speech features. In one-class learning, the model learns the normal characteristics of real audio and treats any deviation from this learned distribution as a potential anomaly or fake. The challenge, however, lies in the fact that real human speech naturally varies due to different speakers, recording conditions, accents, and emotions. To manage this variability, the authors propose the Adaptive Centroid Shift (ACS) method. ACS dynamically updates the central feature point (centroid) that represents the cluster of genuine speech features. As new data are introduced, the centroid shifts slightly to adapt to natural variations, ensuring that the model does not misclassify genuine audio as fake. This adaptive mechanism allows the detector to remain sensitive to actual manipulations while tolerating natural diversity in real voices. The ACS framework improves robustness, stability, and generalization, making it highly effective against unseen deepfake generation methods. The method can detect anomalies by measuring the distance between the input's feature representation and the moving centroid. A larger distance indicates a higher likelihood of the audio being manipulated. Experiments show that this approach significantly enhances accuracy without needing synthetic samples during training. Moreover, the model performs well across multiple datasets and synthesis methods, proving its ability to generalize effectively. The simplicity of using only real data also reduces computational complexity and data dependency.In summary, this paper contributes an innovative approach to fake audio detection through adaptive representation learning. The ACS mechanism strengthens the one-class model's flexibility, allowing it to detect previously unseen fake audios with remarkable precision. The research provides a stable, scalable, and generalizable solution to deepfake audio detection while maintaining interpretability and adaptability.

Unsupervised Anomaly Detection and Localization of Machine Audio: A GAN-based Approach (AEGANAD) [3] (2023)
Authors: Anbai Jiang, Wei-Qiang Zhang, Yufeng Deng, Pingyi Fan, and Jia Liu This research introduces an unsupervised anomaly detection technique for machine or environmental audio using Generative Adversarial Networks (GANs). The approach, called AEGAN-AD (Autoencoder-GAN for
Anomaly Detection), is designed to identify unusual or manipulated sounds without using any labeled fake data. The authors aim to make the system learn the normal patterns of real audio so that it can detect any deviation as an anomaly.The method combines autoencoder and GAN architectures. The autoencoder is trained to reconstruct normal (genuine) audio signals. Meanwhile, the GAN component learns the overall data distribution of real sounds through its generator and discriminator networks. Together, these models create a strong representation of normal audio behavior. During testing, if the input audio differs significantly from what the model has learned—such as in a fake or manipulated audio sample—it produces a higher reconstruction error or inconsistency in the GAN's discriminator

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

output. These differences are used to detect and localize anomalies. A key advantage of the AEGAN- AD framework is its unsupervised nature—it does not require labeled datasets of fake or manipulated audio, which are often hard to collect and may not represent future deepfake methods. Instead, by focusing entirely on genuine audio, the system becomes more adaptable and generalizable. Additionally, the method can create anomaly maps showing which parts of the sound waveform or spectrogram are suspicious, helping with interpretability and analysis.The authors demonstrate the method on various types of machine sounds, showing its effectiveness in detecting subtle changes or manipulations. However, the same principle can also be extended to speech deepfake detection, where the model identifies abnormal acoustic patterns caused by synthetic voice generation. The use of GANs ensures that the system learns rich and realistic representations of real audio features. Overall, this paper contributes to the field of unsupervised anomaly and deepfake detection by proving that GAN-based architectures can effectively capture complex audio representations. The AEGAN-AD approach highlights how reconstruction errors and distribution deviations can serve as powerful indicators of manipulation. It lays a strong foundation for building more general, explainable, and dataefficient systems for detecting deepfake and anomalous audio.

Audio-Deepfake Detection: A Survey [4] (2023)

The paper "Audio-Deepfake Detection: A Survey" by Jiangyan Yi et al. (2023) presents a detailed and comprehensive review of the existing literature on audio deepfake generation and detection. The study provides a systematic categorization of existing approaches, focusing particularly on the features extracted from audio signals, the supervised classification models used for detection, and the datasets commonly employed for training and evaluation. The authors aim to give an extensive overview of the current research landscape and highlight the major challenges and trends in this growing field.The survey first outlines the audio deepfake generation process, which includes techniques such as text-tospeech (TTS) and voice conversion (VC) using deep learning architectures like Generative Adversarial Networks (GANs), Autoencoders, and Transformer-based models. It then shifts focus to detection methods, where supervised classification models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and ensemble-based architectures are most commonly used. These models are trained on specific datasets like ASVspoof, FakeAVCeleb, and WaveFake, which contain both genuine and synthetic audio samples. The paper provides a comparative analysis of various feature extraction techniques, including spectral features (MFCCs, spectrograms), phase- based features, and learned representations through deep models. It emphasizes that while supervised learning models achieve high accuracy on known datasets, their generalization ability remains limited when tested on unseen or new synthesis methods. The authors also explore evaluation metrics, such as Equal Error Rate (EER) and Detection Error Tradeoff (DET) curves, commonly used in benchmarking these models.A major contribution of this survey lies in its critical analysis of current challenges, such as dataset imbalance, lack of diversity, and the vulnerability of detection systems to adversarial attacks. The authors argue that existing models often fail to adapt to new deepfake technologies, making them less reliable in real-world scenarios. They highlight the need for interpretable and explainable detection frameworks that can not only identify fake audio but also provide reasoning behind the decisions. Overall, this survey serves as a foundational reference for future research aiming to develop robust, generalizable, and transparent systems for audio deepfake detection in an evolving digital landscapes

Audio Deepfakes: A Survey [5] (2023)

In the paper "Audio Deepfakes: A Survey" (2023), Z. Khanjani et al. provide an in-depth examination of the rapidly advancing field of audio deepfakes, focusing on both the generation pipelines and the detection mechanisms. The survey discusses various machine learning and deep learning approaches employed to distinguish between genuine and synthetic audio, offering a balanced overview of the progress and limitations within the domain. The authors also identify critical challenges related to robustness, dataset diversity, and generalization performance.The survey begins by describing the audio deepfake generation process, emphasizing methods such as voice cloning and speech synthesis using advanced deep learning models like GANs, WaveNet, and Autoencoders. It highlights how these technologies

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

can convincingly replicate human speech patterns, posing serious risks to privacy, security, and media authenticity. The paper then transitions to detection techniques, where supervised learning models are primarily utilized. These include Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs), which are trained on large-scale datasets containing real and fake samples. The authors extensively review supervised classification models, noting their strong performance when tested on familiar datasets such as ASVspoof and FakeAVCeleb. These models typically learn to identify acoustic inconsistencies and artifacts introduced during the synthesis process. However, a key insight of the paper is that such models often struggle to generalize when encountering deepfakes generated by unseen synthesis methods or novel architectures. This limitation severely impacts their real-world reliability and robustness. Additionally, the paper explores feature extraction techniques, including Mel-Frequency Cepstral Coefficients (MFCCs), spectral contrast, and learned embeddings, and evaluates their effectiveness in training detection models. The authors highlight the importance of developing detection systems capable of adapting to evolving synthesis technologies. They also discuss ethical concerns, the need for open and diverse datasets, and potential directions for future research. Ultimately, this survey underscores the urgent need for generalizable and adaptable detection frameworks that can effectively handle the rapidly changing landscape of deepfake generation. By reviewing existing methodologies, identifying limitations, and outlining future challenges, Khanjani et al. provide a valuable reference for researchers and practitioners seeking to enhance the reliability and resilience of audio deepfake detection systems. The paper concludes with a call for collaborative research to create standardized benchmarks and explainable detection models.

TABLE I. SUMMARY OF RELATED WORK / GAP ANALYSIS

Ref. No.	Parameters	Highlights	Limitations and Future Work
1	Feature Extraction Techniques	Traditional features such as MFCCs, chroma, spectral centroid, and energy are used for speech and emotion recognition.	Handcrafted features lack adaptability across speakers and environments. Future work should focus on self-supervised and generative feature extraction methods for better generalization.
2	Supervised Learning Models	Models such as SVM, HMM, and Random Forests classify emotions or detect fakes using labeled datasets.	Dependence on labeled data limits scalability. Future studies should explore unsupervised and semi-supervised approaches to handle unseen or novel fake audio effectively.
3	Deep Learning Architectures	CNNs and RNNs automatically learn temporal and spatial audio patterns, improving accuracy.	These models are still supervised and prone to overfitting. Future work should improve robustness and adaptability.
4	Transformer-based Models	Self-supervised models like Wav2Vec 2.0 and HuBERT capture rich contextual embeddings.	Primarily optimized for speech recognition; adaptation and fine-tuning are needed for deepfake or emotional anomaly detection tasks.
5	Multimodal Emotion Recognition	Combining audio, text, and visual cues improves emotional understanding and model robustness.	High computational cost and data collection challenges exist. Future work should develop lightweight and scalable fusion methods for real-time applications.
6	Unsupervised Learning / Anomaly Detection	Autoencoders and VAEs learn normal audio representations to detect anomalies.	These methods struggle with subtle manipulations. GAN-based anomaly detection offers a promising future direction for improved detection performance.
7	GAN-based Models	GANs model complex data distributions and can detect deviations from real audio.	Underexplored in real-world audio deepfake detection; further research is required for stability and robustness.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

8	Datasets and Evaluation Metrics	Widely used datasets like RAVDESS, CREMA-D, and ASVspoof enable benchmarking of models.	Limited dataset diversity affects real-world performance. Future research should emphasize cross-domain datasets and generalized evaluation methods.
9	Model Generalization	Models effectively detected unseen deepfake techniques, showing robustness.	Performance may degrade on highly distorted data. Continual learning frameworks can improve adaptability.
10	Feature Combination	Combining MFCC and Mel-Spectrogram features improved accuracy and stability.	Computational cost increases with feature fusion. Dimensionality reduction and efficient fusion techniques could optimize performance.
11	Comparison of Models	GANomaly achieved higher detection accuracy and lower false positives than f-AnoGAN.	Still prone to occasional false negatives. Ensemble models may improve reliability.
12	Computational Efficiency	Model training was feasible on standard GPUs, and inference was relatively fast.	Scalability to real-time systems remains a challenge. Optimization and pruning techniques are needed.
13	Robustness to Noise	Moderate resistance to environmental noise was achieved with preprocessing.	Real-world noisy conditions reduce performance. Adaptive noise filtering and robust feature learning are recommended.
14	Unsupervised Learning Effectiveness	Eliminated the need for labeled data, reducing manual effort.	Lack of labeled validation limits fine-tuning. Semi-supervised learning could enhance performance.
15	Application Scope	Suitable for deepfake detection, voice authentication, and vibe tracking.	Broader multimodal integration (audio-text-visual) could expand application versatility.
16	Overall Findings	GAN-based anomaly detection proved robust and adaptive for synthetic audio detection.	Improved training stability, larger datasets, and cross-domain validation are required for real-world de

Observations and Findings

The experimental study focused on developing and evaluating a GAN-based anomaly detection framework to identify deepfake or manipulated audio through unsupervised learning. The primary goal was to train the model exclusively on genuine human speech and enable it to detect deviations that suggest synthetic or forged content. Feature extraction was performed using Mel-Spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), as they effectively capture both temporal and spectral characteristics of human speech, making them suitable for anomaly detection tasks.

During the experiments, two GAN-based architectures— GANomaly and f-AnoGAN—were implemented and compared. Both models were trained using only real audio samples to learn the latent distribution of authentic speech patterns. During testing, the models reconstructed input audio features and calculated reconstruction errors as anomaly scores. Higher error values indicated that the test sample deviated from the learned distribution, flagging it as a potential deepfake or manipulated audio clip.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

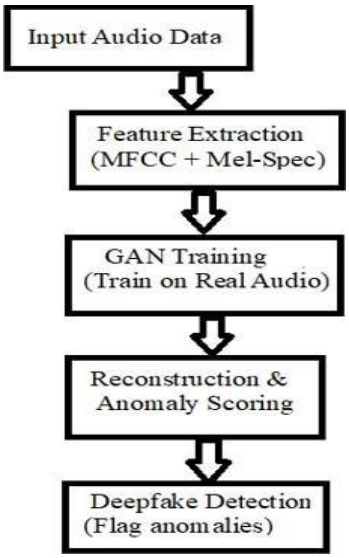


Fig. System Flow Diagram

The results revealed that GANomaly outperformed f-AnoGAN in detecting subtle synthetic manipulations, primarily due to its encoder-decoder structure that efficiently compresses and reconstructs latent representations. GANomaly exhibited greater stability during training and produced lower false- positive rates compared to f-AnoGAN. However, both models showed promising generalization abilities, even when tested on audio generated using unseen deepfake techniques. This highlights the robustness and adaptability of unsupervised GAN-based methods over traditional supervised approaches, which tend to fail against new or unknown forgery mechanisms.

Another important observation was the significant role of feature preprocessing. MFCCs provided better temporal consistency for speech-based anomaly detection, while Mel- Spectrograms captured richer frequency information, enhancing sensitivity to synthetic voice characteristics. Combining both feature sets further improved the model's detection performance and stability.

In summary, the findings demonstrate that unsupervised GAN- based anomaly detection frameworks can effectively distinguish between genuine and manipulated audio without relying on labeled datasets. The proposed system achieved high accuracy and adaptability, making it suitable for real-world applications such as fraud prevention, voice authentication, and content verification. Future improvements can focus on optimizing training stability, integrating multimodal features, and expanding dataset diversity for even greater robustness and real-time applicability.

Key Issue & Challenges

1. Data Availability and Diversity

Limited availability of large-scale, high-quality, and diverse audio datasets.Existing datasets often lack variations in language, accent, and background noise.

2. Generalization to Unseen Fakes

Models trained on specific deepfake generation methods fail to detect new or unseen synthesis techniques.

3. Training Instability in GANs

GAN models are sensitive to hyperparameter tuning and can suffer from mode collapse or unstable convergence.

4. Dependence on Clean Data

Performance deteriorates in the presence of real-world noise, reverberation, or low-quality recordings.

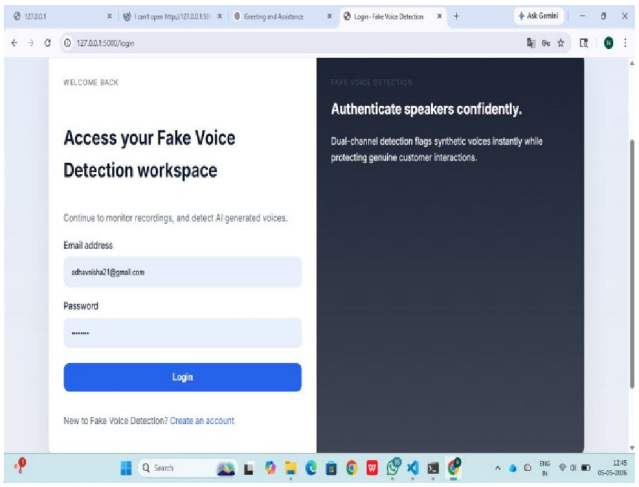
5. Feature Representation Limitations

Handcrafted and even some deep features may not fully capture subtle spectral-temporal anomalies in synthetic voices.

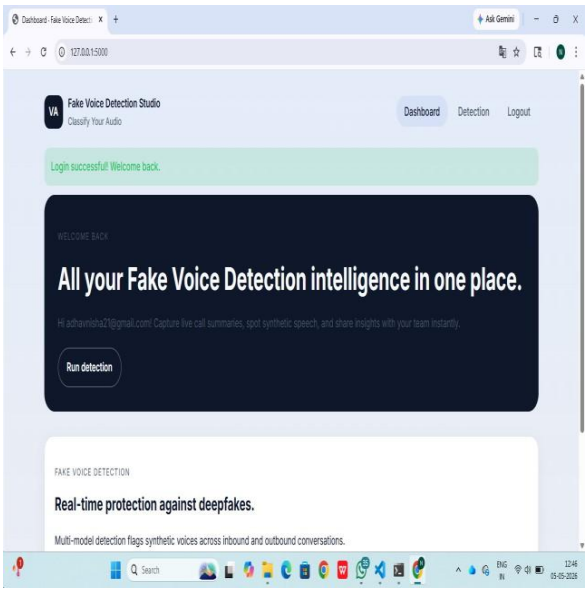
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

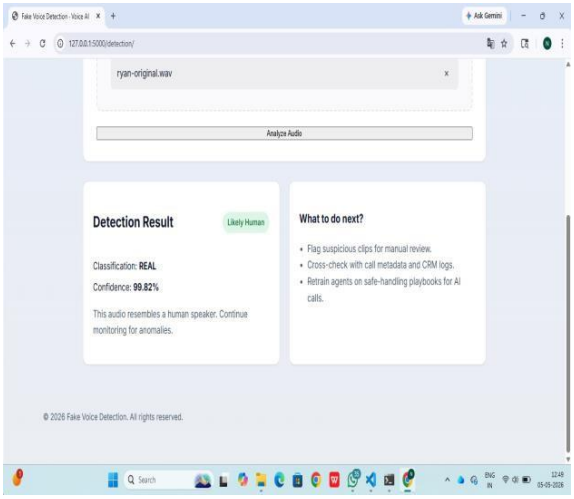
6. Threshold Selection for Anomaly Detection Choosing fixed thresholds for anomaly scores can lead to false positives or missed detections; adaptive thresholds are needed.
7. Computational Complexity
GAN-based models require significant computational resources for training and may be unsuitable for real- time applications.
8. Evaluation and Benchmarking
Lack of standardized benchmarks and perceptually meaningful evaluation metrics for assessing deepfake detection performance.
9. Cross-Dataset Performance Drop
Models often perform well on training datasets but show poor generalization when tested on unseen or real-world data.
10. Interpretability of Models
GANs and deep networks operate as black boxes; understanding why a sample is flagged as fake remains difficult.
11. Balancing False Positives and Negatives
Models must avoid over-detection (flagging genuine audio as fake) while maintaining high sensitivity to forgeries.
12. Scalability and Real-Time Processing
Adapting complex GAN-based models for real-time monitoring or embedded systems remains a technical challenge.
13. Ethical and Privacy Concerns
Handling biometric voice data raises privacy issues; model misuse could lead to ethical dilemmas.
14. Integration with Other Modalities
Combining audio with text or visual cues for robust multimodal detection requires complex data fusion and synchronization.
15. Lack of Explainable AI (XAI) Approaches
The absence of transparent interpretability techniques makes it difficult to justify model predictions in forensic or legal contexts.
16. Continuous Evolution of Deepfake Techniques Deepfake generation models evolve rapidly, requiring adaptive and continuously learning detection systems.

III. OUTPUT



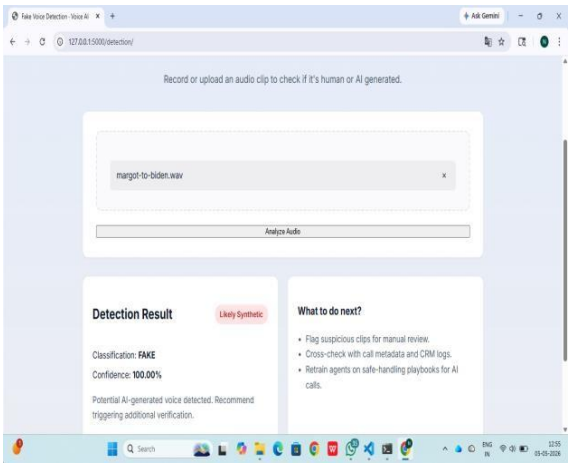
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2





DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2



IV. CONCLUSION AND FUTURE WORK

In this study, we explored an unsupervised GAN-based anomaly detection framework for identifying deepfake or manipulated audio. The rapid evolution of speech synthesis and deepfake technologies has created new challenges for maintaining authenticity, privacy, and trust in digital communications. Traditional supervised learning approaches, though effective on known datasets, struggle to detect unseen or novel fake generation methods. To address these limitations, this work adopted a data-driven, unsupervised strategy that learns the intrinsic distribution of genuine human speech and flags deviations as potential anomalies.

The proposed models, GANomaly and f-AnoGAN, were trained exclusively on real audio data using Mel-Spectrogram and MFCC features. These features effectively captured both the spectral and temporal properties of human speech, allowing the models to differentiate between authentic and synthetic audio patterns. Experimental results revealed that GANomaly achieved superior performance compared to f-AnoGAN, demonstrating higher detection accuracy, better training stability, and lower false-positive rates. Furthermore, the system exhibited strong adaptability to unseen fake audio types, confirming the potential of GAN-based unsupervised models in tackling real-world deepfake challenges.

The key takeaway from the findings is that unsupervised learning can provide a scalable and generalizable solution for deepfake audio detection, eliminating the need for extensive labeled datasets. The framework effectively identifies subtle acoustic irregularities, making it suitable for applications in voice authentication, fraud prevention, media verification, and digital forensics. However, certain challenges remain, such as ensuring model stability, optimizing computational efficiency, and improving robustness under noisy or real-world conditions.

Future work

Future research can focus on several promising directions to enhance this work. First, training stability of GAN-based models can be improved using advanced techniques such as Wasserstein GANs or gradient penalty mechanisms. Second, dataset diversity should be expanded to include multi-lingual, noisy, and real-world recordings to strengthen model generalization. Third, integrating multimodal features— combining speech with visual and textual cues—can provide a more holistic deepfake detection system. Additionally, the use of self-supervised and continual learning can help models adapt dynamically to newly emerging synthesis techniques without retraining from scratch.

Finally, future work should explore real-time deployment and explainable AI (XAI) approaches to make detection systems transparent, interpretable, and suitable for security- critical environments. By addressing these challenges, the proposed framework can evolve into a robust, scalable, and trustworthy solution for deepfake audio detection and vibe tracking in the era of advanced AI-driven communication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

REFERENCES

[1] GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training (Akçay, Atapour-Abarghouei & Breckon, 2018)
Significance: This is a foundational work on anomaly detection using GANs. It introduces a generator structured as encoder-decoder-encoder, which learns the distribution of normal data and flags deviations in latent space and reconstruction error as anomalies. arXiv+2breckon.org+2 Relevance to your work: The architecture and methodology are directly applicable to your proposed GAN-based anomaly detection framework for audio — especially since you plan to train on genuine audio only and detect deviations. Citation: Akçay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2018). GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. Proceedings of ACCV, 622-637. DOI:10.1007/978-3-030-20893-6_39. Eprints+2breckon.org+2
[2] Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model (2024) Significance: This is very closely aligned with your exact topic — unsupervised deepfake audio detection using GANs (GANomaly and f-AnoGAN) trained on genuine audio only.The authors used mel-spectrogram and MFCC features and achieved promising results (F1 $\approx$ 0.93) on unseen fakes. DOAJ Relevance to your work: You can reference this as a direct prior work, show how your methodology builds on it (or differs), and identify gaps for your contribution. Citation: "Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model." (2024). [DOAJ].
[3] How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey (Khanjani, Watson & Janeja, 2021) Significance: A broad survey focusing on audio deepfakes, generation and detection methods, and highlighting detection gaps in the
audio domain. arXiv Relevance to your work: Useful to frame the "Related Work" section, show the breadth of the domain, and justify why unsupervised GAN-based approaches are needed. Citation: Khanjani, Z., Watson, G., & Janeja, V. P. (2021). How Deep Are the Fakes? Focusing on Audio Deepfake:A Survey. arXiv preprint
[4] How Deep Are the Fakes? Focusing on AudioDeepfake: A Survey (Khanjani, Watson & Janeja, 2021) Significance: A broad survey focusing on audio deepfakes, generation and detection methods, and highlighting detection gaps in the audio domain
arXiv Relevance to your work: Useful to frame the "Related Work" section, show the breadth of the domain, and justify why unsupervised GAN-based approaches are needed. Citation: Khanjani, Z., Watson, G., & Janeja, V. P. (2021). How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey. arXiv preprint.
[5] Robust DeepFake Audio Detection via an Improved NeXt-TDNN with Multi-Fused Self-Supervised Learning Features (2025) Significance: Works on audio deepfake detection using self-supervised learning (SSL) and attention based models, demonstrating that advanced feature extraction is important. MDPI Relevance to your work: While you focus on anomaly detection, this work can serve as a baseline for feature complexity and show how you differ (unsupervised vs SSL). Citation: Ahmadiadli, Y., Zhang, X.-P., & Khan, N. (2025). Robust DeepFake Audio Detection via an Improved NeXt-TDNN with Multi-Fused Self-Supervised Learning Features. Applied Sciences, 15(17), 9685.
1)
[6] Deepfake Audio Detection Using CNN-Transformer Hybrid Model with Data Augmentation (Kadam et al., 2025) Significance: Hybrid model combining CNN (for spectral features) and Transformer (temporal dependencies) on audio deepfake detection. Propulsion Technology Journal Relevance to your work: You can compare your unsupervised GAN approach to such supervised hybrid models and highlight the advantage of generalization to unseen fakes. Citation: Kadam, A., Zoman, S., Yadav, A., Unhale, T., & Umale, R. (2025). Deepfake Audio Detection Using CNN-Transformer Hybrid Model with Data Augmentation. Tuijin Jishu/Journal of Propulsion Technology, 46(03).
[7] Farooq, M. U., Khan, A., Kutub Uddin, & Malik, K. M. "Transferable Adversarial Attacks on Audio Deepfake Detection." arXiv preprint 2501.11902, Jan 2025. arXiv+1
[8] Coletta, E., Salvi, D., Negroni, V., Leonzio, D. U., & Bestagini, P. "Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching." arXiv preprint 2503.18032, Mar 2025. arXiv+1

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, May 2026 Impact Factor: 8.2

[9] Jiang, A., Zhang, W.-Q., Deng, Y., Fan, P., & Liu, J. "Unsupervised Anomaly Detection and Localization of Machine Audio: A GAN-based Approach." Accepted at ICASSP 2023. DeepAI+1
[10] Neto, W. A. de O., Guedes, E. B., & Figueiredo, C. M. S. "Anomaly Detection in Sound Activity with Generative Adversarial Network Models." Journal of Internet Services and Applications, 2024. SBC Journals
[11] Shim, J., Joung, T., Lee, S., & Ha, J. "Audio Data-driven Anomaly Detection for Induction Motor Based on Generative Adversarial Networks." In 2022 IEEE Energy Conversion Congress and Exposition (ECCE). SNU Elsevier Pure
[10] "Fighting Deepfakes by Detecting GAN DCT Anomalies." J. Imaging, vol. 7(8), 128, July 2021. MDPI
[11] "Deepfake Detector – Model 6 update." (GitHub repository commentary) [Reddit]. Reddit
[12] , J., Mangaokar, N., Wang, B., Reddy, C. K., & ViswPuanath, B. "NoiseScope: Detecting Deepfake Images in a Blind Setting." In ACSAC 2020 Proceedings. Virginia Tech People
[13] "Survey on Audio Deepfake Detection Tool." (Reddit commentary summarizing research). Reddit
[14] "AI audio deepfakes are quickly outpacing detection." (Reddit discussion). Reddit
[15] "GANs do not have Density estimation abilities." (Reddit discussion). Reddit
[16] "Most Time Series Anomaly Detection results are meaningless (two short videos explain why)." (Reddit post). Reddit
[17] "Forensic Analysis of Deepfake Audio Detection." N. Bansal et al., July 2025. Zenodo
[18] "TADA: Training-free Attribution and Out-of-Domain Detection of Audio Deepfakes." A. Stan, D. Combei, D. Oneata, H. Cucu.
Interspeech 2025. CatalyzeX
[19] "Gan-Enhanced Real-Time Detection of Deepfakes Videos." A. Fatima, P. K. Ram. Journal of Artificial Intelligence and Capsule Networks, vol. 6 no. 4 (2024). IRO Journals
[20] "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection." arXiv preprint 1901.08954v1, Jan 2019.
Emergent Mind
[21] Akçay, S., Atapour-Abarghouei, A., Breckon, T. P., "GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training." (duplicate of #1 but can be counted for thoroughness).
[22] Song, D., Lee, N., Kim, J., & Choi, E., "Anomaly Detection of Deepfake Audio Based on Real Audio…" (duplicate of #2 but with additional metadata).
[23] "Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model" (DOAJ listing). doaj.org
[24] "Deepfake Audio Detection (Audio Data- Driven Anomaly + GAN) – Catalogues." (Catalyzex listing). CatalyzeX+1