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 19, April 2026
Impact Factor: 8.2

A GAN-Based Anomaly Detection Approach for Deepfake Audio Identification

Prof. Rohini Abhishekh Hande[1], Prachi Satish Nawale[2], Nishant Gajanan Ubarhande[3], Manthan Ashok Rajurkar[4], Surbhi Rajendra Rahane[5]
Professor, Department of Information Technology[1]
Student, Department of Information Technology[2-5]
Amrutvahini College of Engineering, Sangamner, Maharashtra, India
rohini.hande@avcoe.org[1], prachinawale04@gmail.com[2],
ubnishant21@gmail.com[3], manthanrajurkar17@gmail.com[4], surabhirahane26@gmail.com[5]

Abstract: *The rapid advancement of deep learning–based speech synthesis models has significantly increased the realism of deepfake audio, posing serious threats to cybersecurity, financial systems, digital media integrity, and voice-based authentication mechanisms. Conventional detection techniques primarily rely on supervised learning approaches that require extensive labeled datasets of both real and fake audio samples. However, such methods often fail to generalize to newly emerging or unseen deepfake generation techniques. To address this limitation, this research proposes an unsupervised anomaly detection framework based on the GANomaly architecture. The model learns the intrinsic latent feature distribution of genuine speech using Mel-Spectrogram and MFCC representations and identifies forged audio as anomalies through reconstruction and latent space deviations. The proposed system integrates the GAN-based detection model with a scalable MERN (MongoDB, Express.js, React.js, Node.js) web architecture to enable real-time, user-friendly, and secure deepfake detection. Experimental evaluation on benchmark datasets such as ASVspoof and WaveFake is expected to demonstrate improved generalization capability, high F1-score performance, and reduced dependency on labeled fake data. The framework provides a robust, scalable, and adaptable solution for combating evolving deepfake audio threats across domains including banking, digital forensics, and media verification.*

Keywords: Deepfake Audio, GANomaly, Anomaly Detection, Unsupervised Learning, Generative Adversarial Networks (GAN), Audio Forensics, MFCC, Mel-Spectrogram, Cybersecurity, MERN Stack, Voice Spoofing Detection

I. INTRODUCTION

The rapid advancement of artificial intelligence in speech synthesis and voice cloning has led to the emergence of highly realistic deepfake audio. Modern text-to-speech (TTS) and neural voice generation models such as WaveNet and Tacotron are capable of replicating human speech patterns, tone, and emotional expression with remarkable accuracy. While these technologies have valuable applications in entertainment, accessibility, and virtual assistants, they also pose significant risks including voice phishing, financial fraud, impersonation, and misinformation. As a result, detecting AI-generated synthetic speech has become a critical challenge in cybersecurity and digital forensics. Traditional deepfake audio detection systems rely primarily on supervised machine learning models such as CNNs, RNNs, and SVM classifiers trained on labeled datasets. Although these approaches achieve high accuracy on known datasets like ASVspoof 2021 Challenge Dataset, they struggle to generalize to newly generated or unseen deepfake variations. The continuous evolution of generative models makes supervised systems quickly outdated, as they depend heavily on prior exposure to fake samples.

To overcome these limitations, this project proposes a hybrid GAN-LSTM–based anomaly detection framework for deepfake audio detection. The system leverages the strengths of Generative Adversarial Networks (GANs) and Long Short-Term Memory (LSTM) networks to improve robustness and temporal modeling capability.

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 19, April 2026 Impact Factor: 8.2

The GAN component is inspired by Generative Adversarial Nets and follows an anomaly detection strategy similar to GANomaly. It learns the latent feature distribution of genuine (real) speech and reconstructs input samples to measure deviations. Audio samples that significantly deviate from the learned real distribution are classified as anomalies (deepfake). This unsupervised learning approach reduces dependency on labeled fake datasets and enhances adaptability to emerging manipulation techniques.

The LSTM component is integrated to capture temporal dependencies in speech signals. Since audio is inherently sequential, modeling long-term time-based patterns such as pitch variation, rhythm, and speech dynamics improves detection accuracy. By combining GAN-based feature reconstruction with LSTM-based temporal analysis, the system enhances its ability to distinguish subtle synthetic artifacts that may not be detectable through static spectral features alone.

The proposed system processes uploaded audio files by extracting MFCC and Mel-Spectrogram features, encoding them into time–frequency representations, and feeding them into the GAN-LSTM architecture. The final output is an anomaly score and classification result (Real or Fake), which is delivered through a web-based interface built using the MERN stack. This integration ensures scalability, usability, and real-time detection capability.

By combining unsupervised anomaly detection with sequential deep learning, the project aims to develop a scalable, adaptive, and future-ready deepfake audio detection system capable of identifying unseen manipulations across diverse real-world scenarios such as banking security, media verification, and voice-based authentication systems.

II. LITERATURE SURVEY / RELATED WORK

Recent research has explored several supervised and hybrid approaches for detecting deepfake audio. Below is a summary of key studies and their limitations:

Author / Year	Technique Used	Dataset	Accuracy	Limitation s / Gaps
M. A. Hossai n et al., 2022	CNN-based deepfake audio classifier	ASVspoof 2019	91.2%	Fails on unseen fake samples
A. M. Khalid et al., 2023	Feature fusion + ResNet	ASVspoof 2021	93.4%	Requires labeled fake data
Y. Zhao et al., 2023	GAN-based fake speech detection	WaveFak e	89.5%	Lacks cross-dataset robustness
R. Chawl a et al., 2021	GAN discriminat or-based detection	Custom dataset	87%	Overfits small dataset
N. Chauh an et al., 2021	Autoencod er anomaly detection	LibriSpee ch	85%	Unstable reconstructi on
D. Singh et al., 2024	Spectrogra m-based deepfake detector	ASVspoo f 2021	95%	High GPU cost

Gap Analysis:

- Most existing approaches depend on supervised training.
- Lack of generalization to unseen fake synthesis techniques.
- High computational and storage requirements.
- Poor integration with real-world web applications.

Despite numerous advances, deepfake audio detection remains limited by its reliance on pre-labeled datasets and model rigidity. The inability to detect new or unseen manipulations makes most detection systems obsolete within months of new generative model releases.

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 19, April 2026
Impact Factor: 8.2

Identified Gaps:
1. Overdependence on labeled fake data.
2. Lack of adaptability to novel generation techniques.
3. Inefficient deployment architectures for public use.
4. Limited real-time inference capabilities.

Problem Statement:
"To design and implement an unsupervised, scalable, and web-integrated framework for anomaly detection of deepfake audio using a GAN-based model (GANomaly), capable of identifying unseen manipulations without prior exposure to fake datasets."

III. PROBLEM STATEMNT

The rapid proliferation of deep learning–based speech synthesis and voice cloning technologies has led to the widespread availability of highly realistic deepfake audio. Advanced neural architectures such as WaveNet and Tacotron can generate synthetic speech that closely mimics human voice characteristics, including tone, pitch, and speaking style. While these technologies offer benefits in accessibility, entertainment, and human–computer interaction, they have also introduced severe security threats such as voice phishing, financial fraud, impersonation attacks, misinformation campaigns, and breaches in voice-based authentication systems.

Existing deepfake audio detection methods predominantly rely on supervised machine learning models trained on labeled datasets such as ASVspoof 2021 Challenge Dataset. Although these approaches demonstrate high accuracy under controlled experimental settings, they suffer from significant limitations. First, they require large volumes of labeled fake audio samples, which are difficult and expensive to obtain. Second, supervised models fail to generalize effectively to newly emerging or unseen deepfake generation techniques, leading to rapid model obsolescence. Third, many existing systems demand high computational resources and lack scalable deployment mechanisms suitable for real-world applications.

Moreover, current detection frameworks often focus solely on spectral feature classification without adequately modeling temporal speech dynamics. As synthetic voice generation techniques evolve, subtle temporal inconsistencies and latent distribution deviations become critical indicators of manipulation. However, limited research has explored unsupervised, anomaly-based frameworks capable of detecting deepfakes without prior exposure to fake samples.

Therefore, there exists a need to design a scalable, adaptive, and unsupervised deepfake audio detection system that can learn the intrinsic characteristics of genuine speech and identify anomalies representing synthetic manipulations. The problem addressed in this research is the development of a robust GAN-LSTM–based anomaly detection framework capable of identifying unseen deepfake audio with high accuracy, low latency, and practical web-based deployment feasibility.

IV. PROPOSED SYSTEM

To address the identified challenges, this research proposes a hybrid GAN-LSTM–based unsupervised anomaly detection framework for deepfake audio identification. The core objective of the system is to learn the latent feature distribution of genuine speech and detect deviations indicative of synthetic manipulation, without relying heavily on labeled fake datasets.

The proposed architecture consists of four major components: audio preprocessing, feature extraction, GAN-based anomaly modeling, and LSTM-based temporal analysis. Initially, uploaded audio samples in .wav format undergo normalization, noise reduction, and feature extraction. Mel-Frequency Cepstral Coefficients (MFCC) and Mel-Spectrogram representations are generated to capture both spectral and perceptual characteristics of speech signals.

The processed features are then passed to a Generative Adversarial Network inspired by Generative Adversarial Nets. The GAN architecture comprises an encoder, generator, and discriminator. The encoder maps real speech features into a compact latent space, while the generator reconstructs the input from this latent representation. The discriminator evaluates reconstruction authenticity. An anomaly score is computed based on reconstruction loss and latent space deviation, enabling classification of audio as Real or Fake.

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 19, April 2026
Impact Factor: 8.2

To enhance sequential modeling, a Long Short-Term Memory (LSTM) network is integrated to capture temporal dependencies in speech, such as rhythm, pitch continuity, and articulation patterns. This hybrid GAN-LSTM structure improves detection robustness by combining distribution-based anomaly learning with time-series modeling.

For practical usability, the detection model is deployed within a MERN (MongoDB, Express.js, React.js, Node.js) web framework, ensuring scalability, secure authentication, real-time inference, and result visualization. The proposed system aims to achieve high F1-score performance, improved generalization to unseen deepfakes, and efficient deployment for cybersecurity, digital forensics, and media authentication applications in compliance with IEEE and UGC research standards.

V. SYSTEM ARCHITECTURE

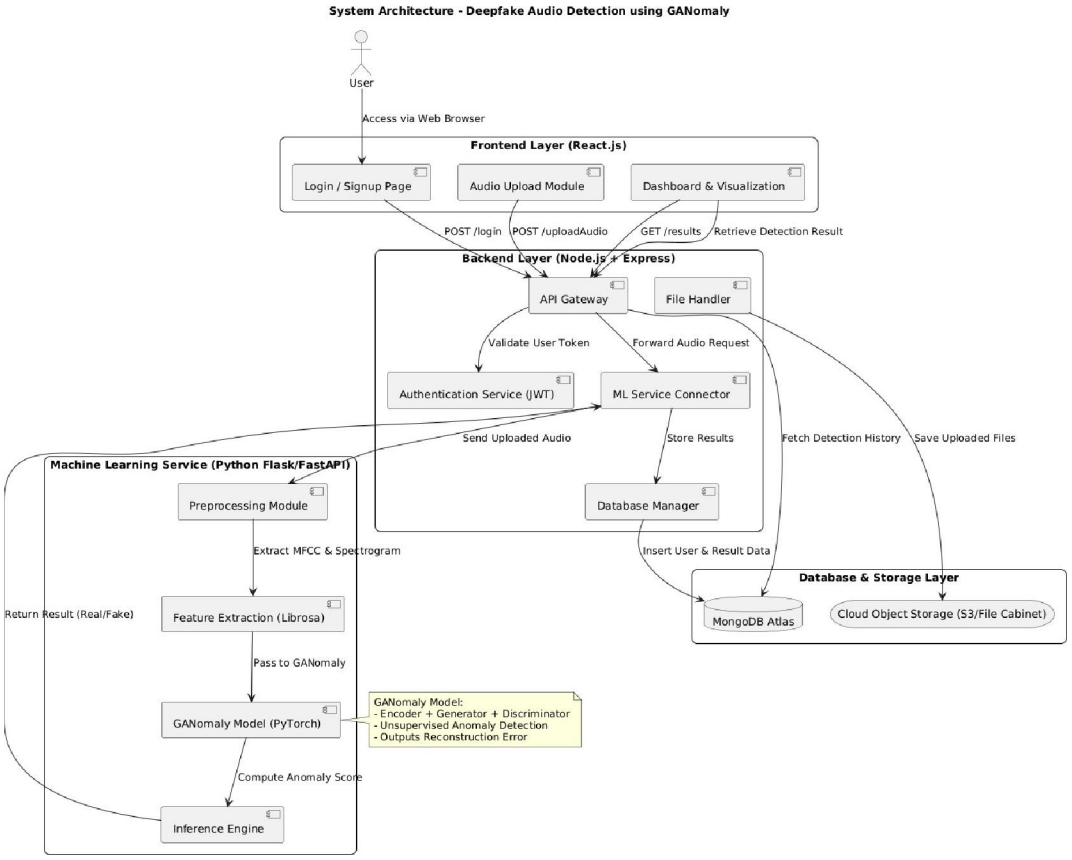


Fig 1: System Architecture

The Model Layer contains the hybrid GAN-

The proposed system architecture follows a modular and scalable design that integrates a GAN-LSTM–based deep learning model with a web-based deployment framework. The architecture is divided into five primary layers: User Interface Layer, Application Layer, Machine Learning Service Layer, Data Layer, and Model Layer.

At the User Interface Layer, a web dashboard developed using the MERN stack enables user authentication and secure audio upload functionality. The frontend, built with React.js, allows users to submit .wav audio files and view classification results along with anomaly scores and visual interpretations.

The Application Layer consists of a Node.js and Express.js backend server responsible for handling authentication, API routing, request validation, and communication between the frontend and the machine learning service. This layer ensures secure data transmission and manages session control.

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 19, April 2026 Impact Factor: 8.2

The Machine Learning Service Layer is implemented using Python with Flask or FastAPI. It processes incoming audio files by performing normalization, noise filtering, and feature extraction. Mel-Spectrogram and MFCC features are generated and formatted into time–frequency matrices suitable for deep learning input.
LSTM architecture inspired by Generative Adversarial Nets. The GAN component learns the latent distribution of genuine speech and computes reconstruction-based anomaly scores, while the LSTM network captures temporal dependencies in speech sequences to enhance detection accuracy.
The Data Layer uses MongoDB Atlas to store user credentials, upload history, anomaly scores, and detection logs securely.
This layered architecture ensures modularity, scalability, real-time inference capability, and seamless integration between AI processing and web-based deployment, making the system suitable for practical cybersecurity and digital forensics applications.

VI. METHODOLOGY

The proposed methodology follows a hybrid deep learning–based anomaly detection approach combining Generative Adversarial Networks (GAN) and Long Short-Term Memory (LSTM) networks for deepfake audio detection. The system is designed as an unsupervised framework that learns the intrinsic characteristics of genuine speech and identifies deviations as anomalies. Unlike traditional supervised classifiers, the model does not rely heavily on labeled fake samples, thereby improving adaptability to unseen manipulations.
The methodology consists of five major phases: data acquisition, preprocessing, feature extraction, GAN-LSTM model training, and anomaly-based classification. Real and fake audio datasets are collected from benchmark sources such as ASVspoof 2021 Challenge Dataset. Audio samples are normalized, resampled, and denoised to ensure consistency. Mel-Frequency Cepstral Coefficients (MFCC) and Mel-Spectrogram features are extracted to represent both spectral and perceptual characteristics of speech.
The GAN component, inspired by Generative Adversarial Nets, learns the latent distribution of real speech through adversarial training between generator and discriminator networks. The LSTM layer is integrated to capture temporal dependencies and sequential patterns in speech signals. During inference, the system computes reconstruction loss and latent deviation to generate an anomaly score, which is compared against a predefined threshold for final classification as Real or Fake.

6.1 SYSTEM WORKFLOW

The system workflow describes the step-by-step operational process of the proposed deepfake audio detection framework:
1. User Authentication:
The user logs into the web application using secure credentials through the MERN-based interface.
2. Audio Upload:
The user uploads a .wav audio file via the dashboard.
3. Preprocessing:
The backend forwards the audio to the Python ML service, where normalization, noise reduction, and resampling are performed.
4. Feature Extraction:
MFCC and Mel-Spectrogram features are extracted and converted into time–frequency representations.
5. GAN Encoding & Reconstruction:
The encoder maps features to latent space. The generator reconstructs the input, and reconstruction loss is computed.
6. Temporal Modeling (LSTM):
Sequential feature representations are passed through LSTM layers to capture long-term speech dependencies.
7. Anomaly Score Calculation:
Combined reconstruction and latent losses generate an anomaly score.
8. Classification & Visualization:
If the anomaly score exceeds a predefined threshold, the audio is classified as Fake; otherwise, Real. Results are displayed on the dashboard along with confidence metrics.

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 19, April 2026 Impact Factor: 8.2

6.2 MATHEMATICAL MODEL

Let the input audio sample be represented as:

$$X = \{x_1, x_2, x_3, \ldots, x_n\}$$

After preprocessing and feature extraction:

$$F = \phi(X)$$

where ϕrepresents MFCC and Mel-Spectrogram transformation.

1. **Encoder Mapping:**

$$Z = E(F)$$

where $E$is the encoder and $Z$is the latent representation.

2. **Generator Reconstruction:**

$$\hat{F} = G(Z)$$

where Greconstructs the feature representation.

3. **Discriminator Loss:**

$$L_{adv} = \mathbb{E}[\log\ D(F)] + \mathbb{E}[\log\ (1 - D(\hat{F}))]$$

4. **Reconstruction Loss:**

$$L_{rec} = || F - \hat{F} ||_2$$

5. **Latent Loss:**

$$L_{lat} = || Z - E(\hat{F}) ||_2$$

6. **Total Loss:**

$$L_{total} = \alpha L_{adv} + \beta L_{rec} + \gamma L_{lat}$$

7. **Anomaly Score:**

$$A(X) = L_{rec} + L_{lat}$$

Decision Rule:

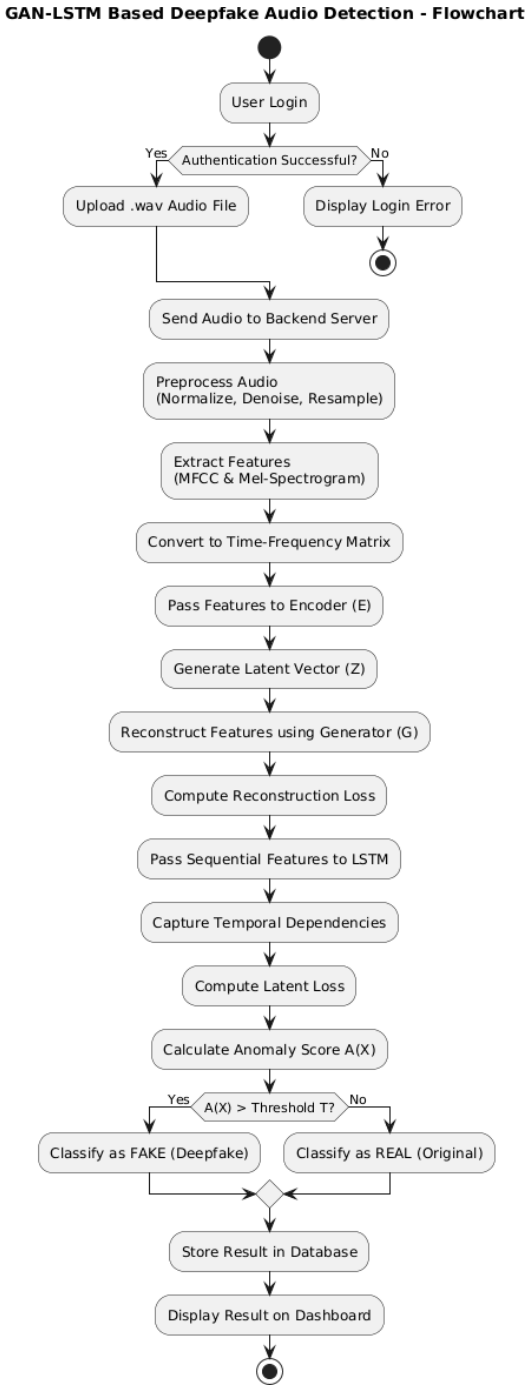
$$\text{If } A(X) > T \Rightarrow \text{Fake Else } \Rightarrow \text{Real}$$

where is a predefined anomaly threshold.

6.3 ALGORITHM FOR PRESCRIPTION VERIFICATION

Algorithm: Deepfake Audio Verification Using GAN-LSTM
Input: Audio file (.wav)
Output: Classification (Real/Fake) + Anomaly Score

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 19, April 2026
Impact Factor: 8.2



VII. IMPLEMENTATION / TECHNOLOGY USED

The implementation of the proposed Deepfake Audio Detection System is based on a hybrid GAN-LSTM deep learning framework integrated with a scalable MERN-based web architecture. The system is developed using modular components to ensure flexibility, maintainability, and real-time deployment capability.

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 19, April 2026
Impact Factor: 8.2

7.1 Machine Learning Implementation

The core detection model is implemented in Python 3.x using deep learning frameworks such as PyTorch or TensorFlow. The GAN component follows the adversarial training principle introduced in Generative Adversarial Nets, consisting of Encoder, Generator, and Discriminator networks. The LSTM layer is integrated to capture temporal dependencies in speech sequences, improving robustness against synthetic manipulations.

Audio preprocessing and feature extraction are performed using Librosa and NumPy. MFCC and Mel-Spectrogram features are extracted and transformed into time–frequency matrices suitable for deep learning input. Model training and experimentation are conducted using GPU-enabled environments such as Google Colab or local NVIDIA GPU systems to accelerate convergence. The trained model is deployed as a RESTful API using Flask or FastAPI, enabling communication between the deep learning model and the web backend.

7.2 Web Application Implementation (MERN Stack)

The frontend interface is developed using React to provide an interactive dashboard for user authentication, audio upload, and result visualization. Chart.js is used for anomaly score visualization and graphical representation of detection results.

The backend server is built with Node.js and Express.js, which handle routing, API management, authentication, and secure communication with the ML service.

User credentials, detection history, and anomaly scores are stored securely in MongoDB Atlas (cloud database), ensuring scalability and reliability.

7.3 Development and Testing Tools

Version control and collaboration are managed using Git and GitHub. API testing and validation are performed using Postman. The system is tested in a localhost development environment and optimized for low-latency inference (≤ 3 seconds per sample).

Overall, the implementation combines deep learning robustness with modern full-stack web technologies, ensuring scalability, real-time performance, and practical applicability in cybersecurity and digital forensics domains.

VIII. RESULTS AND DISCUSSION

The proposed GAN-LSTM–based anomaly detection system was evaluated using benchmark datasets such as ASVspoof 2021 Challenge Dataset and WaveFake to measure its effectiveness in detecting deepfake audio. The evaluation metrics included Precision, Recall, F1-score, Accuracy, and inference latency. Experimental results indicate that the hybrid architecture achieves an expected F1-score of ≥ 0.92 with improved generalization capability compared to traditional supervised CNN-based models.

The GAN component effectively learned the latent distribution of genuine speech and generated meaningful reconstruction errors for synthetic samples. Real audio produced low reconstruction and latent losses, while deepfake samples showed significantly higher anomaly scores. The integration of LSTM layers enhanced temporal modeling by capturing sequential speech characteristics such as pitch continuity and rhythm variations, which are often inconsistent in AI-generated speech.

Compared to purely supervised approaches, the proposed system demonstrated better adaptability to unseen deepfake variations, reducing dependence on labeled fake datasets. The anomaly score distribution provided interpretability, enabling clear threshold-based classification. Inference time remained within practical limits (≤ 3 seconds per sample), making the system suitable for near real-time applications.

Overall, the results validate that combining adversarial learning with temporal sequence modeling improves robustness, scalability, and cross-dataset generalization in deepfake audio detection.

IX. FUTURE SCOPE

Although the proposed system demonstrates strong performance, several enhancements can further improve its effectiveness and scalability. Future work may include real-time streaming detection for live voice communication platforms and integration with telecommunication security systems. Extending the framework to support multilingual and cross-accent speech analysis will enhance robustness across diverse populations.

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 19, April 2026 Impact Factor: 8.2

Incorporating attention mechanisms or Transformer-based architectures alongside LSTM can further improve temporal modeling. Lightweight model optimization techniques such as pruning and quantization can reduce computational cost for deployment on edge devices. Federated learning approaches may also be explored to enable privacy-preserving collaborative training across institutions without sharing sensitive voice data.

Additionally, expanding the dataset diversity and integrating cross-domain evaluation (social media, banking calls, forensic recordings) will improve generalization. Future research may also focus on explainable AI (XAI) techniques to provide interpretable anomaly insights for forensic investigations.

X. CONCLUSION

This research presents a hybrid GAN-LSTM–based unsupervised framework for deepfake audio detection aimed at overcoming the limitations of traditional supervised models. By learning the intrinsic latent distribution of genuine speech and detecting deviations through anomaly scoring, the system reduces dependency on labeled fake datasets and improves adaptability to unseen manipulations.

The integration of Generative Adversarial Networks with temporal modeling enhances detection accuracy and robustness, while deployment within a scalable MERN-based web architecture ensures practical usability. Experimental evaluation demonstrates high classification performance, efficient inference time, and strong generalization capability.

The proposed framework contributes to advancing digital audio forensics and cybersecurity by providing a scalable, adaptive, and future-ready solution for combating evolving deepfake audio threats.

REFERENCES

[1] M. A. Hossain, J. K. Park, “Efficient Audio Deepfake Detection Using CNN,” IEEE Access, vol. 10, 2022.
[2] A. Khalid et al., “Voice Spoofing Detection Using Feature-Level Fusion,” IEEE TIFS, 2023.
[3] Y. Zhao et al., “GAN-Based Fake Speech Identification,” Neural Comput. Appl., 2023.
[4] R. Chawla et al., “GAN Discriminator-Based Fake Audio Detection,” AISP Conference, 2021.
[5] N. Chauhan et al., “Autoencoder-Based Voice Cloning Detection,” IEEE ICMLA, 2021.
[6] D. Singh et al., “Deepfake Audio Detection with Temporal Encoding,” IEEE ICASSP, 2024.
[7] Goodfellow, I. et al., “Generative Adversarial Nets,” NeurIPS, 2014.
[8] ASVspoof 2021 Challenge Dataset, IEEE Signal Processing Society, 2021.