Audio Deepfake Generation and Detection using Deep Learning for Digital Forensic Analysis

Mahra Alnaqbi and Richard Ikuesan
Zayed University, Abu Dhabi, UAE

Mahar.Alnaqbi@gmail.com
Richard.Ikuesan@zu.ac.ae

Abstract: Deepfake audio technology has developed rapidly and now poses serious risks in cybersecurity and digital forensic analysis. Synthetic speech can imitate authentic human voices, which makes manual and automated detection difficult. This study investigates deep learning-based methods for detecting audio deepfakes generated using modern speech synthesis and voice conversion techniques. The work focuses on both generation and detection in order to understand how different deepfake methods affect detection performance. The study employs multiple audio generation models, including Coqui TTS, GAN, RealNVP, VAE, and WaveNet, to generate realistic speech. Several deep learning detection models are evaluated, including CNN, LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM, conditional GAN, and hybrid architectures that combine convolutional and recurrent layers. Three datasets are used for training and evaluation. These include a self-generated deepfake dataset, a public Kaggle deepfake dataset, and the ASVspoof 2021 deepfake dataset. This combination allows evaluation under both controlled and real-world conditions. Experimental results indicate apparent performance differences among model types. Simple sequential models, such as LSTM and BiLSTM, perform poorly when deepfake audio exhibits strong naturalistic characteristics. This issue is most evident in Coqui TTS-generated audio, which is difficult to detect because of its natural tone and smooth articulation. In comparison, hybrid models that combine convolutional and recurrent learning consistently achieve higher accuracy and stronger generalization across datasets. The Hybrid cGAN with BiLSTM achieves near-perfect detection performance and exhibits stable performance across cross-validation folds and on independent test data. These results confirm that combining spatial and temporal feature learning improves robustness against advanced deepfake attacks. The study also introduces AudioForenX, an interactive forensic tool that integrates the best-performing models. AudioForenX enables real-time analysis, waveform visualization, and classification of audio samples as authentic or synthetic. The findings confirm that hybrid deep learning architectures provide reliable, balanced, and highly accurate detection of synthetic audio. This study contributes practical insights for digital forensics and supports the development of effective tools to counter audio deepfake threats.

Keywords: Audio deepfake detection, Bidirectional LSTM, Conditional GAN, Convolutional neural network, Deep learning, Speech synthesis

1. Introduction

Audio deepfakes now represent a critical security threat in modern communication due to their ability to imitate authentic human voices with high accuracy. They are already being used for fraud, impersonation, misinformation, and cyberattacks, and current manual or human judgment is no longer sufficiently reliable to detect them. This creates a significant risk in environments where voice is used for authentication, evidence, or identity verification. This study begins with this core challenge and investigates how deep learning (DL) can reliably support the forensic need to distinguish real human speech from synthetic speech.

Audio deepfake technology builds on DL models to create synthetic speech that is almost indistinguishable from real human speech. These models can manipulate or generate voice recordings that replicate the tone, pitch, rhythm, and emotion of a real person. Speech synthesis and voice conversion now allow the cloning of vocal identity with very high precision. The term deepfake joins deep learning and fake to show the technology that enables this transformation (McGettigan et al., 2024). Audio deepfakes, like many other deepfakes, were initially used for lawful purposes. They made a personalized speech for people with speech disorders as well as made the virtual assistants such as Siri and Alexa more natural. Unfortunately, the same technology is being used maliciously today. Such intelligence is exploited by criminals for identity theft, impersonation, and misinformation campaigns. It has raised concerns among security and media professionals, who are struggling to detect fabricated audio (Akhtar et al., 2024; Liu et al., 2024).

Audio deepfakes have become more and more convincing due to the fast evolution of DL models. Early robotic speech was robotic and unnatural-sounding. WaveNet and Tacotron generate speech so realistically that they almost perfectly imitate tone, pitch, and emotion (Li et al., 2025). The level of realism is so high that professionals trained to detect such audio manipulation face difficulty without advanced tools. The increasing use of audio deepfakes in cyberattacks underscores the need for robust detection mechanisms (Alali & Theodorakopoulos, 2024). In the previous decade, the development of audio deepfake technology has advanced significantly. Initially, speech synthesis used feedback based on basic signal-processing techniques to

produce synthetic speech. The early systems were successful only in a limited sense: they replicated natural human speech. DL changed many things. With the arrival of generative adversarial networks (GANs) and recurrent neural networks (RNNs), it became possible to generate highly realistic speech (Alshehri et al., 2024; Choi et al., 2024). Using publicly available software such as VoiceClone and Resemble AI, even non-experts can create realistic synthetic voices (Mirsky & Lee, 2021). As a result, audio-based threats have risen in number due to this ease of access.

One of the most important milestones in this evolution is the development of neural text-to-speech (TTS) models. Unlike traditional TTS systems, neural TTS models learn complex patterns in human speech from data (Li et al., 2024). In fact, they do not just record what people say; they also record how that person expresses their emotions. With advances in this technology, synthetic audio is becoming more realistic, and detecting it is becoming more difficult (Wang et al., 2024). While deepfake generation models have improved rapidly, equally strong detection models have not developed as rapidly. Many existing methods fail when tested against novel deepfake types or voices generated by advanced speech synthesis techniques. This creates a clear research gap for detection models that are accurate, adaptable, and capable of operating under real forensic constraints.

Digital forensics plays an important role in the creation, storage, and alteration of digital evidence. It provides investigators with a defined process for examining data in a careful and controlled manner. It also ensures that the evidence is reliable when used for legal or security decisions. The rapid growth of audio deepfakes has made this forensic task more difficult. Modern deepfake models do not leave such visible clues. They replicate human tone, pitch, and time with great accuracy. Signal transformations, including the Wavelet Transform (WT), Constant-Q Transform (CQT), and Short-Time Fourier Transform (STFT), can be used in digital forensics to convert audio into visual patterns that reveal important details (Pham et al., 2024). Existing studies explain why the field of digital forensics is now under new pressure. They demonstrate that audio evidence is no longer readily credible. They also indicate that deepfake audio continues to improve rapidly.

This study follows this direction by examining the impact of deepfake audio on speech structure. It also develops detection methods that support precise and reliable forensic decisions. These forensic requirements directly motivate the study and guide the development of detection models and the final forensic tool. The study aims to develop and evaluate DL-based methods for generating and detecting audio deepfakes for use in digital forensic analysis. It also aims to build a practical tool for real-time verification of human speech authenticity. The study focuses on improving accuracy, generalization, and forensic reliability in realistic digital environments. The study sets the following objectives: to study and organize the state-of-the-art methods in audio deepfake creation and detection, to examine different DL models and identify the most suitable method for detecting audio deepfakes, to build a dataset of real and fake audio samples for forensic testing, and to develop a forensic analysis tool that supports the investigation and reporting of audio deepfakes.

2. Literature Review

Many studies focus on how audio deepfakes work and how to prevent them. Early work focused on speech synthesis and voice conversion. Later, DL models changed all this. Some studies are about the risks, and others focus on fake voice detection by ML. Detection is difficult because deepfake technology continues to improve. This section reviews key studies on audio deepfake generation and detection.

Li et al. (2025) developed SONAR, a benchmark framework for detecting AI-synthesized audio across various TTS and Voice Conversion (VC) models. As this Realistic AI Speech technology has advanced, distinguishing the authentic voice of a human from deepfakes has become challenging because current detection methods generalize poorly. SONAR fills this gap by evaluating detection systems on a novel dataset from nine varied audio synthesis platforms. They demonstrate that foundation models generalize better than traditional systems due to their larger scale and richer pretraining data. Xie et al. (2024) present FakeSound, a dataset designed explicitly for deepfake general audio detection. FakeSound is a strong benchmark for evaluating deepfake detection models. A general-purpose audio pre-trained framework is proposed, which outperforms state-of-the-art deepfake speech detectors and human testers. Alali and Theodorakopoulos (2024) present a detailed review of the methods for generating and detecting TTS, VC, and partially fake audio. TTS and VC deepfake detection models are very accurate and robust to unseen data. However, partially synthetic audio, which consists of a mixture of real and synthetic speech, remains a challenging problem. The authors discuss key characteristics of publicly available datasets and highlight the growing importance of detecting partially synthetic audio.

Zhang et al. (2024) contributed to the increasing threat of deepfake audio by their DeepFakeVox-HQ dataset, which comprises 1.3 million samples, including 270,000 high-quality deepfakes. They introduced Frequency-Selective Adversarial Training, which focuses on high-frequency components that attackers often use to hide. Results show improvement to model performance in difficult situations and validate the importance of using robust and realistic datasets for deepfake detection. Dagar and Vishwakarma (2022) studied the use of GANs and Variational Autoencoders in the generation of audio deepfakes. They focus on how rapidly these generative models have grown and discuss the modern methods, datasets, and architectures. The challenges and limitations of the technology are also discussed, as are future directions for improved security and detection.

Sharma et al. (2024) propose a multimodal deepfake detection method that leverages video and audio data. They use DL to extract important features from both modalities. In audio media, Convolutional Neural Networks (CNNs) are used to detect patterns of fake content, typically using mel-spectrograms. Their model improves the accuracy and precision by combining the audio and video data. Wani et al. (2024) used CNN and Bidirectional Long Short-Term Memory (BiLSTM) to analyze multiple audio features, including CQT, Constant-Q Cepstral Coefficients (CQCC), mel-spectrograms, and Mel Frequency Cepstral Coefficients (MFCCs). CNN layers created multi-dimensional feature representations, and BiLSTM captured temporal patterns. The model achieved high accuracy on ASVSpoof 2019 and FoR datasets. Alshehri et al. (2024) presented Sonic Sleuth, an Artificial Intelligence (AI) model for detecting audio deepfakes. Sonic Sleuth has an accuracy of 98.27%. The model performs well even under noisy conditions and supports multiple languages.

Deepfake forensics focuses on how investigators study, verify, and explain audio evidence reliably. The field assists in legal work, where the truth of the recording must be determined before it is used in a case. Several studies aim to address this gap by presenting models, tools, and workflows for investigating both speech and synthetic speech. The reviewed studies show that the DL models have been the most effective tools in identifying audio deepfakes. The CNNs, BiLSTMs, and hybrid models achieve good performance by learning spectral and temporal features. Models that incorporate augmentation, privacy protection, or multimodal inputs also yield more accurate results by processing more complex patterns (Sharma et al., 2024). Extensive datasets, such as DeepFakeVox HQ and FakeSound, support these models by providing a wide range of high-quality samples (Zhang et al., 2024; Xie et al., 2024).

However, many gaps remain. Most detection models struggle when tested on types of deepfakes they have never encountered. New TTS and VC systems generate voices that are more natural than those in older datasets, which limits generalization. These gaps indicate the need for further research. There is a clear need for models that can detect deepfakes across different synthesis methods, operate on noisy data, explain their decisions, and support forensic reporting.

3. Methodology

This section describes how audio deepfakes are generated and detected. It explains the dataset preparation, feature extraction, deepfake generation techniques, and DL-based detection models used in the study. Figure 1 demonstrates the operational framework employed in the study. The study selected deepfake generation and detection algorithms following a detailed exploratory process. This explains that the selection was purposeful and guided by investigation.

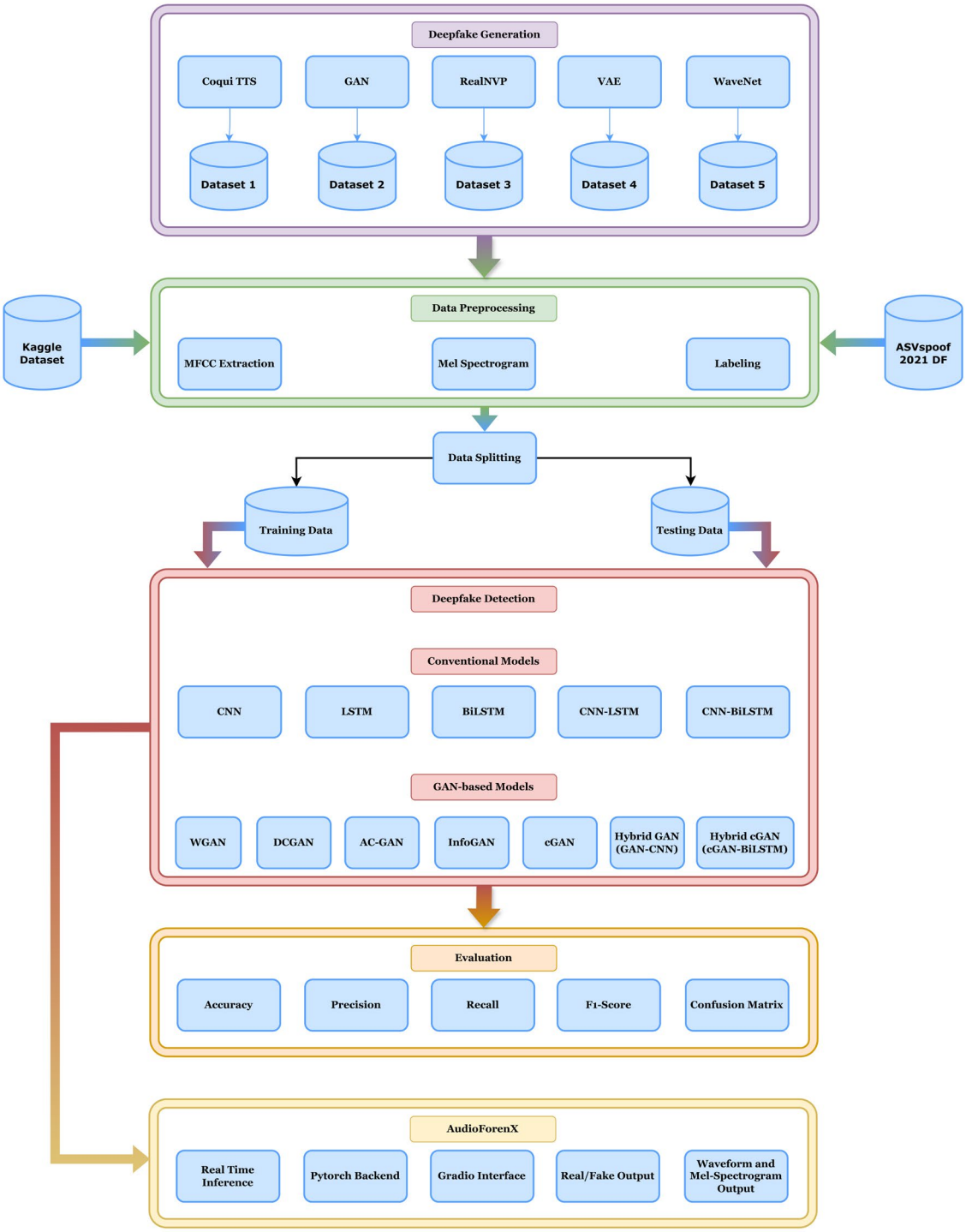

Figure 1: Methodology Overview

3.1 Dataset Generation and Preparation

Real audio samples were generated based on the Coqui TTS model and other methods. The Coqui TTS model was used to synthesize human-like speech from predefined text prompts. A variety of sentences were used to produce a variety of speech recordings. These clips retained their natural tone, pitch, and rhythm, which are pretty similar to human speech. All generated real audio samples were saved in a separate dataset folder. Each file was saved in .wav format with a standardized sampling rate. The dataset was designed to be consistent to enable high-quality speech synthesis.

Fake audio samples were produced by altering the real speech using DL models. Multiple techniques were employed to make the synthesized audio natural while retaining some of the variables that distinguish it from

actual speech. The modification techniques included changes to the speech waveform's frequency, slowing or speeding the audio without altering pitch, and the introduction of background noise. In addition, deep generative models were also used to generate fake audio datasets. Each technique introduced different variations. These techniques generated diverse datasets, resulting in robust detection. The total number of audio samples per technique was 10,000, comprising 5,000 real and 5,000 synthetic samples.

3.2 Public Dataset Collection

Two public datasets were used. The Kaggle "Deepfake Detection in Audio" dataset contains labelled real and fake audio samples and supports supervised classification (Sizakele, 2024). The ASVspoof 2021 Speech Deepfake dataset contains real and synthetic speech generated under realistic conditions and was used to test robustness and generalization (Skenospeniel, 2021).

3.3 Feature Extraction and Preprocessing

In order to effectively train DL models, it is essential to transform raw audio data into meaningful representations. The feature extraction technique has been used to extract essential characteristics of speech. MFCCs were extracted from each audio sample. MFCCs model how humans perceive sound and have been widely used in speech recognition and deepfake detection. Each audio file is converted into a 40-dimensional MFCC vector. This conversion preserves key features while reducing noise and redundancy. Mel spectrograms are visual representations of sound that show the evolution of frequency components over time. CNNs use spatial structures to detect deepfake patterns through spectrograms. To improve model robustness, the spectrograms were augmented with noise. To prevent overfitting and improve generalization, factual distortions were simulated by adding Gaussian noise.

All audio files from Kaggle and ASVspoof 2021 datasets were resampled to 16 kHz and converted to single-channel waveforms. Each audio clip was padded or trimmed to four seconds. Every waveform was converted to a mel-spectrogram using the short-time Fourier transform. A total of 64 mel-frequency bins were extracted. The mel-spectrograms were normalized to the range -1 to 1. Time masking and frequency masking were applied to enhance generalization.

3.4 Deepfake Audio Generation Techniques

Powerful DL techniques are used in the generation of audio deepfakes. This study employed five generative models: Coqui TTS, GAN, Real-Valued Non-Volume-Preserving (RealNVP), Variational Autoencoder (VAE), and WaveNet. Coqui TTS generated speech from written text. It produced audio that sounded human-like, capturing tone, pitch, and rhythm. GANs used a generator and a discriminator in adversarial training to produce synthetic audio. The generator learned speech patterns from real data and refined its output through feedback. RealNVP generated deepfake audio by learning an invertible transformation between real and fake distributions. VAE learned latent representations of real speech and reconstructed synthetic samples. WaveNet generated speech sample by sample and captured fine temporal and spectral details of human voices. Each generation method produced datasets containing 5,000 real and 5,000 fake samples. These balanced datasets enabled reliable training and evaluation.

3.5 Deepfake Detection Models

Detecting deepfake audio requires powerful DL models. The study used CNN, Long Short-Term Memory (LSTM), BiLSTM, CNN-LSTM, CNN-BiLSTM,, Wasserstein GAN (WGAN), Deep Convolutional GAN (DCGAN), Auxiliary Classifier GAN (AC-GAN), Information Maximizing GAN (InfoGAN), Conditional GAN (cGAN), Hybrid GAN, and Hybrid cGAN with Bi-LSTM. CNNs extracted spatial features from spectrograms. LSTMs captured temporal dependencies. BiLSTMs analyzed sequences in both forward and backward directions. Hybrid CNN-LSTM and CNN-BiLSTM models combined spatial and temporal learning. GAN-based models learned discriminative patterns through adversarial training. Five-fold cross-validation was applied to ensure reliable evaluation.

3.6 AudioForenX Tool Development

The study developed an end-to-end deepfake detection tool called AudioForenX. The system uses a Python backend implemented in PyTorch and an interactive Gradio interface. It loads trained cGAN, Hybrid GAN-CNN, and Hybrid cGAN-BiLSTM models using original files. Uploaded audio is normalized, resampled, padded, and converted to mel-spectrograms using the same pipeline as the training process. Predictions are displayed

along with waveform and spectrogram visualizations. The tool supports CPU- and GPU-based execution and does not store user audio, thereby ensuring privacy.

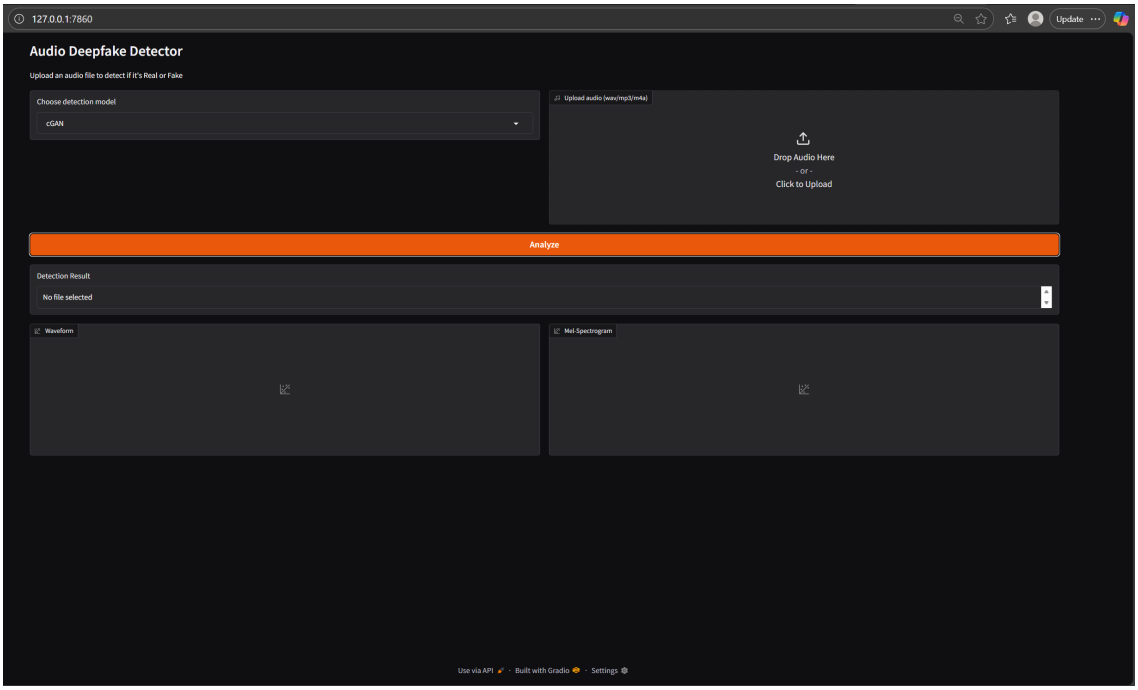

Figure 2: AudioForenX Interface

4. Results and Analysis

This section presents the results of all deepfake detection models on the generated and collected audio. These results explain the generation methods' ability to produce high-quality deepfakes and demonstrate how well the detection models identify them. The developed tool is also evaluated for accuracy and interpretability to clarify its forensic value.

4.1 Results on Generated Deepfake Data

Coqui TTS-generated data posed a challenge for all detection models. This dataset included real audio and high-quality synthetic audio that were difficult to separate. CNN achieved the best performance across all metrics, achieving 93% accuracy. However, LSTM and BiLSTM models struggled to detect deepfakes, achieving less than 90% accuracy. This indicates that simple sequential models cannot accurately detect deepfakes in Coqui-generated data. Coqui TTS-generated data was the most difficult to detect. Some models misclassified fake samples. This makes the Coqui TTS technique more efficient than others. The dataset generated with the GAN model contained high-quality synthetic samples. All detection models achieved perfect classification on GAN-generated data. The perfect accuracy observed on GAN-generated data does not indicate that the models have unlimited generalization ability; instead, the GAN audio contained clear spectral artifacts that made the samples easier to detect. The RealNVP model produced high-quality audio, and all detection models achieved almost perfect classification. The VAE model also produced high-quality audio, and most models achieved perfect classification, although CNN had minor misclassifications. The WaveNet model produced deepfake audio that all models perfectly detected. Hybrid models, CNN-LSTM and CNN-BiLSTM, performed best on all datasets.

Table 1: Results on Generated Deepfake Data

Dataset	Best Model(s)	Best Accuracy	Key Notes
Coqui TTS	CNN	0.93	High-quality deepfakes. Sequential models struggled.
GAN	All models	1.00	Clear patterns. Perfect detection.
RealNVP	All models	1.00	All models detected samples perfectly.
VAE	Most models	1.00	CNN had minor errors. Others perfect.
WaveNet	All models	1.00	Perfect detection.

4.2 Results on Kaggle Deepfake Data

The Kaggle deepfake dataset contains real and synthetic audio, which serves as the basis for deepfake classification. The WGAN achieved an overall accuracy of 60.23%. The results show that the model was biased towards the real class. The DCGAN achieved moderate performance but misclassified many fake samples. The AC-GAN achieved lower accuracy and performed better at detecting fakes than at detecting real data. The InfoGAN achieved moderate accuracy but exhibited a bias toward real audio. The cGAN achieved perfect performance across all validation folds. The model had an accuracy, precision, recall, and f1-score of 1.00. The Hybrid GAN with CNN achieved high stability and near-perfect performance. The Hybrid cGAN with a Bi-LSTM achieved near-perfect accuracy and strong generalization. These results show that hybrid models are more reliable for deepfake detection.

Table 2: Results on Kaggle Deepfake Data

Model	Accuracy	Strength	Weakness
WGAN	0.60	Good recall for real class	Poor fake detection
DCGAN	0.61	Moderate real detection	Many fake
AC-GAN	0.48	Better at fake detection	Poor real detection
InfoGAN	0.65	Good real detection	Poor fake detection
cGAN	1.00	Perfect detection	None observed
Hybrid GAN (GAN-CNN)	0.99	High stability	Small misclassifications
Hybrid cGAN (cGAN-BiLSTM)	1.00	Perfect performance	None observed

Although some models achieved very high accuracy, this does not imply universal performance against all types of deepfake. The results show that hybrid models learned stable temporal and spectral features, but unseen generators may still affect performance.

4.3 Results on ASVSpoof 2021 Deepfake Data

The ASVspoof 2021 deepfake dataset includes real and deepfake audio generated under realistic conditions. The WGAN achieved an accuracy of 50.33% and showed bias towards the fake class. The DCGAN achieved very low accuracy and failed to detect fake samples correctly. The AC-GAN showed strong fake detection but weak real detection. The InfoGAN achieved very low overall accuracy and exhibited strong bias toward real samples. The cGAN achieved perfect classification on both the validation and test data. The Hybrid GAN with CNN achieved strong fake-detection performance but low recall on real samples. The Hybrid cGAN with Bi-LSTM achieved perfect detection with no misclassifications. These results show that hybrid conditional models are more effective on large-scale and imbalanced datasets.

Table 3: Results on ASVspoof 2021 Deepfake Data

Model	Accuracy	Strength	Weakness
WGAN	0.50	Moderate recall	Bias toward fake class
DCGAN	0.12	High recall for real	Poor fake detection
AC-GAN	0.90	Strong fake detection	Weak real detection
InfoGAN	0.04	Good real detection	Almost no fake detection
cGAN	1.00	Perfect detection	None observed
Hybrid GAN (GAN-CNN)	0.98	Strong fake detection	Low recall for real
Hybrid cGAN (cGAN-BiLSTM)	1.00	Perfect performance	None observed

The high detection performance on the ASVspoof 2021 dataset is attributable to the dataset's structure and strong class separability. These results should not be interpreted as universal detection performance.

4.4 AudioForenX: Forensic Analysis Tool

AudioForenX loads and runs three trained DL architectures, which include cGAN, Hybrid GAN with CNN, and Hybrid cGAN with Bi-LSTM. Figures 38 and 39 show the detection results of real and fake samples using cGAN. The waveform and mel-spectrogram of both samples are displayed, which explain the acoustic structure of the

audio samples. AudioForenX also accurately distinguishes real and synthetic samples using a Hybrid GAN with a CNN and a Hybrid cGAN with a Bi-LSTM.

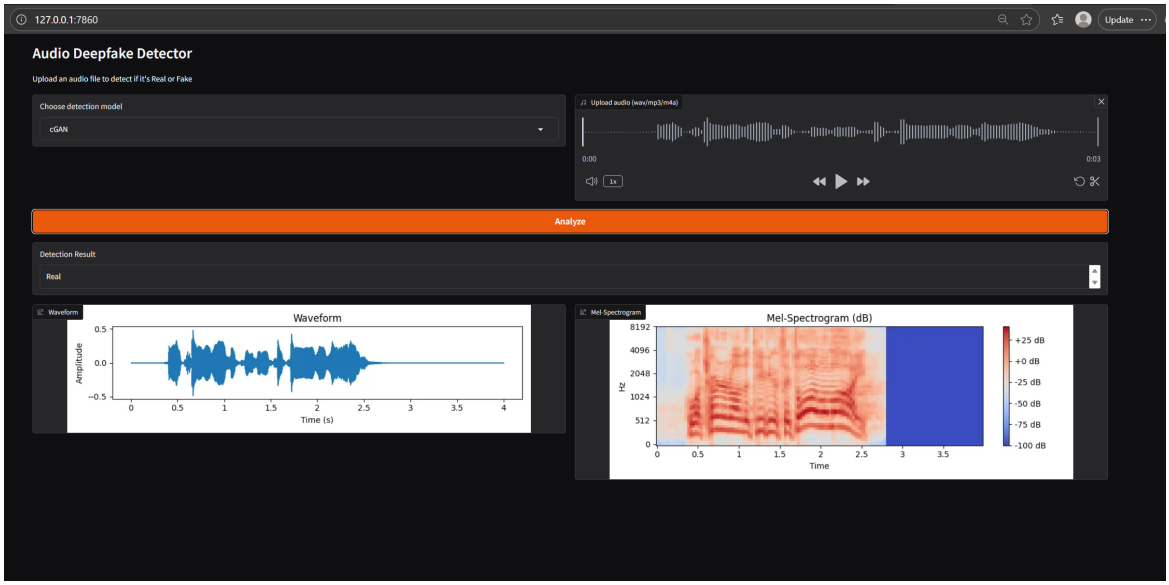

Figure 3: Detection using cGAN on a Real Audio Sample

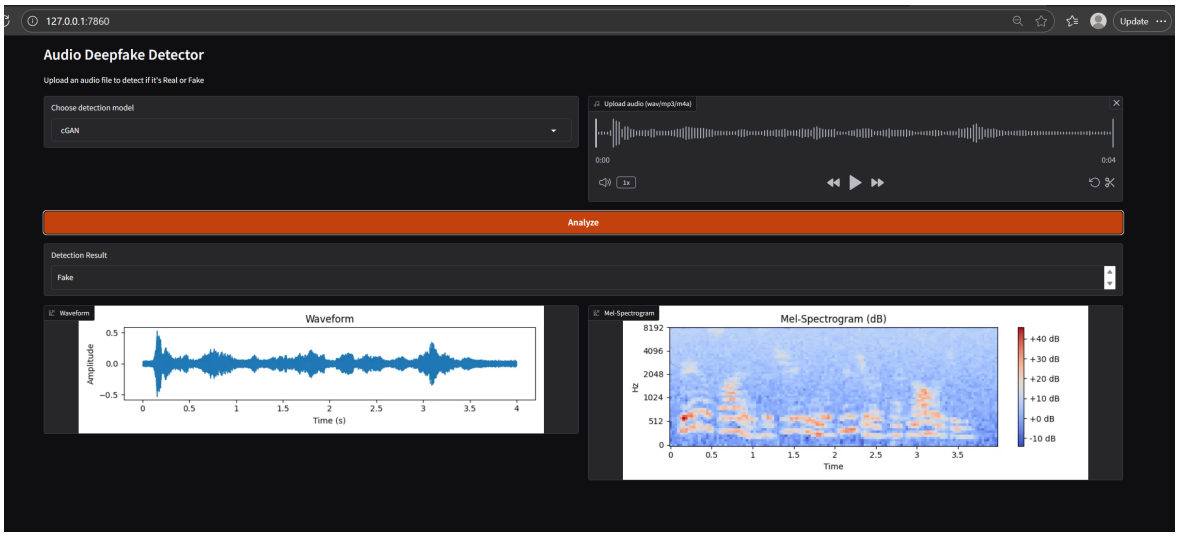

Figure 4: Detection using cGAN on a Fake Audio Sample

These visual results corroborate the quantitative findings and demonstrate how the models distinguish between real and synthetic audio. The tool demonstrates strong accuracy and clear forensic value.

5. Discussion

This research began with the problem that audio deepfakes pose an increasing security threat because they can synthesize human voices to facilitate fraud, impersonation, and misinformation (Akhtar et al., 2024). The findings validate that human judgment is no longer reliable for identifying synthetic audio and that DL is required to support forensic decision-making. Experiments on generated data show that CNN-based and hybrid temporal-spectral models analyze acoustic structures that are not readily perceived by humans, supporting earlier findings that machine-based analysis is more effective for forensic voice detection (Alshehri et al., 2024).

One important contribution is an explanation of why specific deepfake generation strategies are more complex to detect. Coqui TTS employs advanced neural text-to-speech synthesis and produces highly realistic human speech, which led to higher misclassification rates in LSTM and BiLSTM models. This supports earlier observations that modern neural TTS models generate speech with authentic tone, pitch, and emotional

patterns (Sharma et al., 2024). The results provide empirical evidence that higher synthesis quality increases detection difficulty and that realistic deepfakes alter temporal patterns in ways similar to real speech.

The study also shows near-perfect detection performance on GAN, RealNVP, VAE, and WaveNet-generated data. These results indicate that earlier generation models leave clear acoustic artifacts, which supports existing observations that constrained generative methods lack sufficient variability (Dagar & Vishwakarma, 2022). However, these results do not indicate universal detector robustness.

On Kaggle and ASVspoof datasets, basic GAN variants such as WGAN, DCGAN, InfoGAN, and AC-GAN showed bias and poor class balance. These failures highlight the limitations of single-architecture models and confirm that hybrid conditional and temporal models exhibit stronger generalization. The strong performance of hybrid cGAN with BiLSTM supports the need for combined spectral and temporal learning in realistic forensic environments (Zhang et al., 2024).

The findings show that DL can detect fake speech with strong accuracy when trained on structured and diverse data. Hybrid models perform best because they analyze both frequency and time patterns. Coqui TTS deepfakes are the hardest to detect, and traditional LSTM models struggle with advanced speech. This confirms the need for hybrid models in modern forensic environments. The study supports the need for automated systems in digital forensics. Future work should employ transformer-based and attention-driven models, apply explainable AI techniques, expand datasets with diverse accents and languages, integrate acoustic and linguistic properties, and develop multimodal systems that combine audio and video signals.

6. Conclusion

The study meets all four objectives stated at the beginning of the work. It first addresses the objective of studying and organizing state-of-the-art methods for audio deepfake creation and detection. The study reviews the main generative models and the existing detection methods. It explains how current systems function and highlights the limits of existing solutions. This provides a clear view of the field and supports the design of subsequent experiments. The study meets the second objective by testing a wide set of DL models for deepfake detection. The results show that CNN, LSTM, BiLSTM, and their hybrid forms are effective at learning spectral and temporal features that distinguish real from synthetic speech. The experiments show that the Hybrid cGAN with a BiLSTM achieves the highest accuracy and yields consistent results across all datasets. The study also meets the third objective by developing a comprehensive dataset of real and synthetic audio. It generates balanced sets of 5,000 real and 5,000 synthetic samples for each generation technique. These datasets support fair testing, reproducibility, and future research. The study meets the fourth objective by developing a forensic tool named AudioForenX. The tool provides real-time testing, visualization of model decisions, and supports forensic investigation and reporting. This tool represents the study's practical contribution.

References

Akhtar, Z., Pendyala, T. L., & Athmakuri, V. S. (2024). Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve. *Forensic Sciences*, *4*(3), 289–377.

Alali, A., & Theodorakopoulos, G. (2024, June). Review of existing methods for generating and detecting fake and partially fake audio. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics* (pp. 35–36).

Alshehri, A., Almalki, D., Alharbi, E., & Albaradei, S. (2024). Audio Deep Fake Detection with Sonic Sleuth Model. *Computers*, *13*(10), 256.

Choi, J. E., Schäfer, K., & Zmudzinski, S. (2024, June). Introduction to Audio Deepfake Generation: Academic Insights for Non-Experts. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation* (pp. 3–12).

Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International journal of multimedia information retrieval*, *11*(3), 219–289.

Li, X., Bu, F., Mehrish, A., Li, Y., Han, J., Cheng, B., & Poria, S. (2024). CM-TTS: Enhancing Real Time Text-to-Speech Synthesis Efficiency through Weighted Samplers and Consistency Models. *arXiv preprint arXiv:2404.00569*.

Li, X., Chen, P. Y., & Wei, W. (2025). Where are we in audio deepfake detection? A systematic analysis over generative and detection models. *ACM Transactions on Internet Technology*.

Liu, H., Chen, Y., Narayanan, A., Balachandran, A., Moreno, P. J., & Wang, L. (2024). Can DeepFake Speech be Reliably Detected?. *arXiv preprint arXiv:2410.06572*.

McGettigan, C., Bloch, S., Bowles, C., Dinkar, T., Lavan, N., Reus, J., & Rosi, V. (2024). Voice cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, *54*(1), 1–41.

Pham, L., Lam, P., Nguyen, T., Nguyen, H., & Schindler, A. (2024, September). Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)* (pp. 1–5). IEEE.

Sharma, V. K., Singh, S. N., & Caudron, Q. (2024). Combating Deepfakes Using an Integrated Framework forAudio and Video Deepfake Detection.

Sizakele, N. (2024). *Deepfake Detection in Audio* [Data set]. Kaggle. https://www.kaggle.com/datasets/noxolosizakele/deepfake-detection-in-audio

Skenospeniel. (2021). *ASVspoof 21 DF* [Data set]. Kaggle. https://www.kaggle.com/datasets/skenospeniel/asvspoof-21-df

Wang, K., Zhang, J., Ren, Y., Yao, M., Shang, D., Xu, B., & Li, G. (2024). SpikeVoice: High-Quality Text-to-Speech Via Efficient Spiking Neural Network. *arXiv preprint arXiv:2408.00788*.

Wani, T. M., Qadri, S. A. A., Comminiello, D., & Amerini, I. (2024, June). Detecting audio deepfakes: Integrating CNN and BiLSTM with multi-feature concatenation. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security* (pp. 271–276).

Xie, Z., Li, B., Xu, X., Liang, Z., Yu, K., & Wu, M. (2024). FakeSound: Deepfake General Audio Detection. *arXiv preprint arXiv:2406.08052*.

Zhang, Z., Hao, W., Sankoh, A., Lin, W., Mendiola-Ortiz, E., Yang, J., & Mao, C. (2024). I Can Hear You: Selective Robust Training for Deepfake Audio Detection. *arXiv preprint arXiv:2411.00121*.