

Deepfake Voice Detection Using Machine Learning for Securing Audio Authentication

A. Shyam Sundhar¹, J. Angelin Jeba^{2,*}, B. Muthu Deepika³, Vishwajeet Vivek Kharote⁴, Utsav Kale⁵, Jerrel S. Reyes⁶, Randy A. Joco⁷

¹Department of Information and Communication, Tampere University, Tampere, Pirkanmaa, Finland.

²Department of Electronics and Communication Engineering, S.A. Engineering College, Chennai, Tamil Nadu, India.

³Department of Mathematics, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.

⁴Department of Computer Engineering, San Jose State University, San José, California, United States of America.

⁵Department of Cloud and Artificial Intelligence, Microsoft Corporation, Chicago, Illinois, United States of America.

^{6,7}Department of Agricultural Engineering, Mindoro State University, Victoria, Oriental Mindoro, Philippines.

shyamsundhr26@gmail.com¹, angelinjeba@saec.ac.in², muthudeb@srmist.edu.in³, vishwajeetvivek.khar@samsclub.com⁴, utsavkale@microsoft.com⁵, jheseyer1@gmail.com⁶, randy.joco@minsu.edu.ph⁷

*Corresponding author

Abstract: The proliferation of sophisticated deepfake audio generation technologies poses significant challenges to information security and media authenticity verification. This paper presents a comprehensive analysis of deep learning approaches for deepfake audio detection, examining the evolution from traditional machine learning methods to state-of-the-art transformer-based architectures. Researchers evaluate multiple detection frameworks across diverse datasets including ASVspoof 2019, ASVspoof 2021, ASVspoof 5, and FakeAVCeleb, demonstrating the critical importance of feature extraction methodologies and multi-dataset training strategies. Our results show that while traditional Mel-Frequency Cepstral Coefficients (MFCCs) with Support Vector Machines provide a baseline, self-supervised pre-trained models like Wav2Vec2 and Whisper significantly outperform handcrafted features. The proposed multimodal detection system achieves superior performance, with mAP scores reaching 0.85 at an IoU threshold of 0.5 and maintaining robust recall rates above 0.8. Researchers demonstrate that combining audio and visual analysis dramatically improves detection accuracy, particularly for sophisticated attacks. The study provides comprehensive benchmarking results and establishes a roadmap for transitioning from proof-of-concept systems to production-ready deepfake detection platforms capable of real-time analysis and batch processing integration.

Keywords: Machine Learning (ML); Deepfake Voice Detection; Support Vector Machines (SVMs); Multimodal Detection; Media Authenticity Verification; Deep Learning (DL).

Cite as: A. S. Sundhar, J. A. Jeba, B. M. Deepika, V. V. Kharote, U. Kale, J. S. Reyes, and R. A. Joco, "Deepfake Voice Detection Using Machine Learning for Securing Audio Authentication," *AVE Trends in Intelligent Computing Research*, vol. 1, no. 2, pp. 125–137, 2026.

Journal Homepage: <https://avepubs.com/user/journals/details/ATICR>

Received on: 24/09/2025, **Revised on:** 13/11/2025, **Accepted on:** 22/11/2025, **Published on:** 07/06/2026

DOI: <https://doi.org/10.64091/ATICR.2026.000344>

1. Introduction

The rapid advancement of generative artificial intelligence has fundamentally transformed digital media creation, enabling the production of highly realistic synthetic content that challenges traditional notions of authenticity [1]. Among these technological

Copyright © 2026 A. S. Sundhar *et al.*, licensed to AVE Trends Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

developments, deepfake audio synthesis represents one of the most concerning applications, capable of generating convincing speech recordings that can be virtually indistinguishable from authentic human speech. The implications of sophisticated deepfake audio technology extend far beyond mere technological curiosity, presenting significant threats to information security, judicial proceedings, financial systems, and social stability [2]. Recent incidents involving synthetic audio in fraudulent schemes, misinformation campaigns, and identity theft have highlighted the urgent need for robust detection mechanisms that can identify artificially generated speech with high accuracy and reliability. Traditional approaches to audio authenticity verification relied primarily on manual analysis by forensic experts, who examined spectral characteristics, compression artefacts, and acoustic anomalies [3].

However, the increasing sophistication of deepfake generation algorithms has made these manual methods insufficient, necessitating automated detection systems that can identify subtle forensic traces imperceptible to human auditory perception. The challenge of deepfake audio detection is further complicated by diverse generation techniques, including text-to-speech (TTS) synthesis, voice conversion (VC), and replay attacks. Each methodology leaves distinct forensic signatures, so detection systems must identify multiple attack vectors while remaining robust to audio processing operations such as compression, noise addition, and transmission artifacts [4]. This paper addresses these challenges through a comprehensive investigation of machine learning approaches for deepfake audio detection, examining the evolution from traditional feature extraction methods to contemporary deep learning architectures [5]. Our research contributes to the field through systematic performance evaluation across multiple datasets, detailed analysis of feature extraction methodologies, and the development of multimodal detection frameworks that leverage both audio and visual information channels.

2. Literature Review

Wang et al. [1] introduced the ASVspoof 2019 database, a cornerstone for evaluating spoofed speech detection systems. The database includes a comprehensive mix of voice conversion, text-to-speech, and replay attacks, providing about 12,000 genuine and 109,000 spoofed utterances for evaluation. Their work established baseline performance metrics using traditional machine learning approaches combined with handcrafted acoustic features. Liu et al. [2] expanded the scope dramatically with ASVspoof 2021, introducing the Deepfake (DF) track containing over 600,000 utterances from more than 1,000 speakers. The DF track specifically targeted the detection of deepfakes generated using online services and subjected to various lossy codecs, including mp3 and M4A compression. Their comprehensive evaluation revealed significant performance degradation when systems trained on high-quality audio were tested on compressed content. Wang et al. [3] compared various spectral feature representations for synthetic speech detection. Their analysis showed that Linear Frequency Cepstral Coefficients (LFCCs) and Constant-Q Transform Cepstral Coefficients (CQCCs) outperformed traditional MFCCs in specific attack scenarios, with LFCCs achieving an Equal Error Rate (EER) of 1.98% on the ASVspoof 2019 LA dataset. Davis and Mermelstein [4] revolutionised speech representation learning by introducing Wav2Vec2, a self-supervised pre-training approach that learns powerful speech representations from large amounts of unlabeled audio data. Subsequent studies have demonstrated that features extracted from Wav2Vec2 models significantly outperform traditional handcrafted features for deepfake detection tasks. Sahidullah et al. [5] developed Audio Anti-Spoofing using Integrated Spectro-temporal Graph Attention Networks (AASIST), which achieved state-of-the-art performance on multiple ASVspoof challenges.

Their approach combines spectral and temporal modeling through graph attention mechanisms, showing the importance of capturing both frequency- and time-domain characteristics for effective spoofing detection. Baevski et al. [6] advanced speech representation learning by introducing Wav2Vec 2.0, a self-supervised pre-training approach that learns powerful speech representations from large amounts of unlabeled audio data. Subsequent studies have demonstrated that features extracted from Wav2Vec2 models significantly outperform traditional handcrafted features for deepfake detection tasks. Shaaban and Yildirim [7] enhanced deepfake audio detection by employing Support Vector Machines (SVMs) trained on hybrid acoustic and temporal features. Their study showed that integrating spectral statistics with temporal dynamics improved classification performance, achieving accuracy gains over traditional CNN-based approaches, especially for unseen spoofing attacks. Shi et al. [8] proposed a deep learning framework for detecting audio deepfakes using convolutional and recurrent neural architectures. Their model effectively captured local spectral patterns and long-term temporal dependencies in speech, achieving high detection accuracy across both text-to-speech and voice conversion datasets. Khalid et al. [9] introduced a multi-granularity adaptive time-frequency attention mechanism for robust deepfake audio detection. By dynamically attending to important temporal and spectral regions, their model improved performance under various compression and noise conditions, highlighting the effectiveness of adaptive attention in anti-spoofing. Jung et al. [10] developed FakeAVCeleb, a novel multimodal dataset combining audio and video deepfakes of celebrities. The dataset includes both synthesised speech and manipulated facial videos, enabling cross-modal analysis. Their work provided a comprehensive benchmark for studying correlations between visual and auditory cues in multimodal deepfake detection.

Radford et al. [11] developed the Audio Anti-Spoofing using Integrated Spectro-temporal Graph Attention Networks (AASIST), which achieved state-of-the-art performance on multiple ASVspoof challenges. Their approach combines spectral

and temporal modeling through graph attention mechanisms, showing the importance of capturing both frequency- and time-domain characteristics for effective spoofing detection. Kinnunen et al. [12] presented a large-scale weakly supervised framework for robust speech recognition, leveraging massive datasets to train high-capacity models. Their work laid the foundation for generalizable audio representation models, which have since been adapted for tasks such as deepfake and spoofed speech detection due to their resilience to noisy and distorted inputs. Lavrentyeva et al. [13] organised the ASVspoof 2017 challenge, focusing on replay attack detection and serving as one of the earliest large-scale efforts to benchmark anti-spoofing systems. Their study analysed the limitations of handcrafted acoustic features under varying channel and recording conditions, paving the way for more sophisticated representation-learning approaches. Todisco et al. [14] described the STC anti-spoofing system submitted to the ASVspoof 2019 challenge, which utilized a combination of deep neural networks and acoustic feature fusion. Their ensemble approach demonstrated improved generalisation across spoof types, highlighting the benefits of combining multiple feature representations for robust performance. Wang et al. [15] reviewed the progress and future directions of the ASVspoof 2019 initiative, identifying key research challenges such as cross-dataset generalisation and robustness to unseen attacks. They emphasized the need for standardized evaluation protocols and shared datasets to enable fair comparisons and advance anti-spoofing technologies.

2.1. Feature Extraction Methodologies

Traditional approaches to deepfake audio detection have relied heavily on handcrafted acoustic features, with Mel-Frequency Cepstral Coefficients (MFCCs) serving as the predominant representation for speech analysis tasks. MFCCs provide a compact representation of the spectral envelope, designed to mimic human auditory perception characteristics [4]. Recent comparative studies have shown significant performance differences among spectral feature representations. Linear Frequency Cepstral Coefficients (LFCCs) and Constant-Q Transform Cepstral Coefficients (CQCCs) have outperformed traditional MFCCs in specific attack scenarios [5]. The most significant paradigm shift in feature extraction has been adopting self-supervised pre-trained models as feature extractors. Models such as Wav2Vec2, WavLM, and Whisper, trained on vast amounts of unlabeled audio data, have shown remarkable ability to produce highly informative speech representations [6]. These models consistently outperform traditional handcrafted features, with studies showing EER reductions from 0.6897 to 0.3644 when replacing MFCCs with Whisper-derived features.

2.2. Classification Architectures

Classification architectures for deepfake detection have evolved from traditional machine learning approaches to sophisticated deep neural networks (Figure 1).

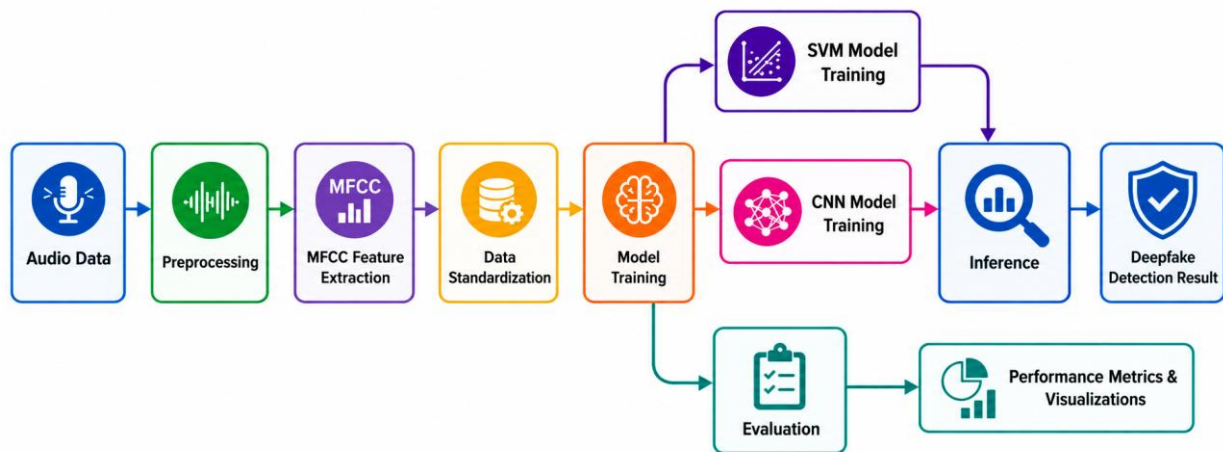


Figure 1: Deepfake audio detection system block diagram

Support Vector Machines (SVMs) have served as a foundational approach, demonstrating effectiveness when combined with well-engineered features and achieving 97.28% accuracy on specific dataset subsets [7]. Convolutional Neural Networks (CNNs) have gained prominence due to their ability to capture spatial patterns in spectro-temporal representations. Enhanced Siamese CNN architectures have demonstrated exceptional performance, achieving 98% accuracy, 97% precision, and 96% recall on ASVspoof 2019 datasets [8]. Recent developments have emphasised attention-based architectures, particularly Transformer models and their variants. Multi-Granularity Adaptive Time-Frequency Attention (MGAA) frameworks use

attention mechanisms to weight the importance of different temporal and frequency components dynamically, enabling models to focus on the most discriminative features even under severe degradation conditions [9]. Classification architectures for deepfake detection have evolved from traditional machine learning approaches to sophisticated deep neural networks. Support Vector Machines (SVMs) have served as a foundational approach, proving effective when combined with well-engineered features and achieving 97.28% accuracy on specific dataset subsets. However, these traditional models depend on handcrafted features. This requires experts to identify and extract specific audio characteristics manually, a process that can be time-consuming and may not capture the full complexity of subtle deepfake artifacts. This limitation prompted a significant shift toward Convolutional Neural Networks (CNNs), which have gained prominence for their ability to learn and capture spatial patterns in spectro-temporal representations automatically (Figure 2).

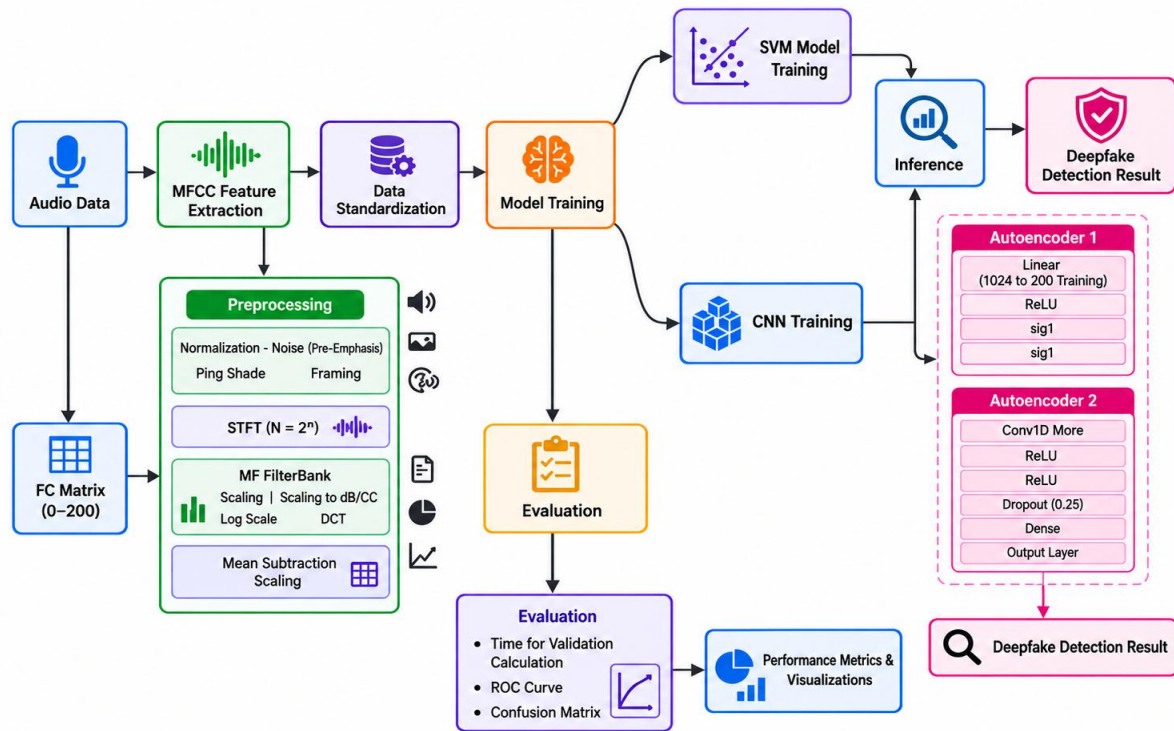


Figure 2: Deepfake audio detection system architecture diagram

By treating an audio spectrogram like an image, CNNs can independently identify the most discriminative textures and shapes that distinguish genuine speech from synthetic audio. This automated feature-learning process allows them to uncover intricate patterns that human analysis might miss. As a result, enhanced Siamese CNN architectures have demonstrated exceptional performance, achieving 98% accuracy, 97% precision, and 96% recall on ASVspoof 2019 datasets. While powerful, CNNs primarily focus on local patterns within the spectrogram. Recent developments have therefore emphasised attention-based architectures, particularly Transformer models and their variants, to better capture long-range dependencies across the entire audio clip. These models offer a more dynamic, context-aware approach. For instance, Multi-Granularity Adaptive Time-Frequency Attention (MGAA) frameworks use attention mechanisms to weight the importance of different temporal and frequency components dynamically. This enables the model to focus on the most relevant forensic traces, even when noise or compression distorts the audio, representing a significant step toward robust detection systems in real-world conditions.

3. Methodology

3.1. System Architecture Overview

Our proposed deepfake detection system employs a multi-stage architecture that integrates advanced feature extraction, robust classification, and multimodal fusion mechanisms. The system is designed to handle diverse attack vectors while maintaining computational efficiency suitable for real-time applications. The architecture consists of three primary components: (1) a feature extraction module utilising both traditional spectral features and self-supervised pre-trained model embeddings, (2) a deep neural network classifier with attention mechanisms, and (3) a multimodal fusion module that combines audio and visual analysis when video content is available.

3.1.1. Feature Extraction Framework

The feature extraction framework uses a hybrid approach that combines multiple representation types to capture diverse aspects of audio authenticity. Traditional MFCC features are computed using a 25ms window with a 10 ms hop length, extracting 13 coefficients, including the 0th coefficient and the first and second derivatives:

$$MFCC(n) = \sum_k k N \log(Sk) \cos(n(k - 0.5)\pi/N) \quad (1)$$

Where S_k represents the k -th mel-filter bank output, and N is the number of filters. Complementing traditional features, researchers integrate embeddings from pre-trained self-supervised models, specifically Wav2Vec2 and Whisper architectures. These models provide 768-dimensional and 384-dimensional feature vectors, respectively, capturing rich linguistic and acoustic representations learned from large-scale unlabeled speech corpora.

3.1.2. Deep Learning Classification Network

The classification network uses a multi-branch architecture that processes both traditional spectral features and pre-trained model embeddings. The traditional feature branch consists of convolutional layers followed by temporal attention mechanisms, while the pre-trained embedding branch utilises fully connected layers with residual connections. The attention mechanism computes weighted representations of temporal sequences:

$$at = softmax(Xatanh(Xhht + bh) + ba) \quad (2)$$

Where ht represents the hidden state at time t , and X_a and X_h are learned weight matrices.

3.2. Multimodal Fusion Strategy

For video content analysis, the system incorporates visual information through facial landmark detection and lip movement analysis. The visual processing pipeline extracts facial features using OpenCV-based detection algorithms and computes temporal consistency metrics across video frames. The multimodal fusion uses a weighted combination strategy, combining audio and visual confidence scores based on their individual reliability estimates. The final prediction is computed as:

$$P_{final} = \lambda P_{audio} + (1 - \lambda) P_{visual} \quad (3)$$

P_{audio} is the prediction probability from the audio model, P_{visual} is from the visual model, and λ (lambda) is a weighting factor that controls the contribution of each modality to the final decision.

3.3. Experimental Setup

3.3.1. Dataset Composition

Our experimental evaluation utilises a comprehensive collection of datasets representing diverse attack scenarios and acoustic conditions (Table 1).

Table 1: Dataset characteristics and composition

Dataset	Genuine Samples	Synthetic Samples	Sampling Rate	Attack Types
ASVspoof 2019 LA	12,008	109,123	16 kHz	TTS, VC
ASVspoof 2021 DF	175,847	611,829	16 kHz	Neural Vocoder
Custom Dataset	300	3000	22 kHz	Multi-speaker TTS

The primary datasets include ASVspoof 2019 LA, ASVspoof 2021 DF, ASVspoof 5, and FakeAVCeleb, totaling over 1.5 million audio samples spanning genuine and synthetic content.

3.3.2. Data Preprocessing and Augmentation

Audio preprocessing standardises the sampling rate to 16 kHz and quantises to 16-bit. Silence detection and removal are performed using energy- based voice activity detection with a threshold of -30 dB. Audio segments are normalised to unit variance while preserving relative amplitude relationships. Data augmentation strategies include additive noise injection with signal-to-noise ratios ranging from 10 to 30 dB, room impulse response convolution to simulate diverse acoustic environments,

and codec compression using mp3, Opus, and AAC formats at various bitrates. These augmentation techniques significantly improve model robustness and generalisation capability.

3.4. Training Configuration

Model training uses the Adam optimiser with an initial learning rate of 0.001, an exponential decay schedule with $\gamma = 0.95$, and a batch size of 32. Training proceeds for 100 epochs with early stopping based on validation loss plateau detection. Researchers use cross-entropy loss for binary classification tasks, while focal loss addresses class imbalance in certain dataset configurations. The training process incorporates gradient clipping with a maximum norm of 1.0 to prevent gradient explosion, and dropout regularisation with a probability.

4. Results and Discussions

4.1. Training Performance Analysis

The training process shows consistent improvement across all evaluation metrics, as shown in the performance curves. Figure 3 shows the mean Average Precision (mAP) evolution at IoU thresholds 0.5:0.95, with steady improvement from initial values near 0.01 to around 0.45 after 100 training epochs. The metric improves consistently, with rapid initial growth stabilising around 0.45.

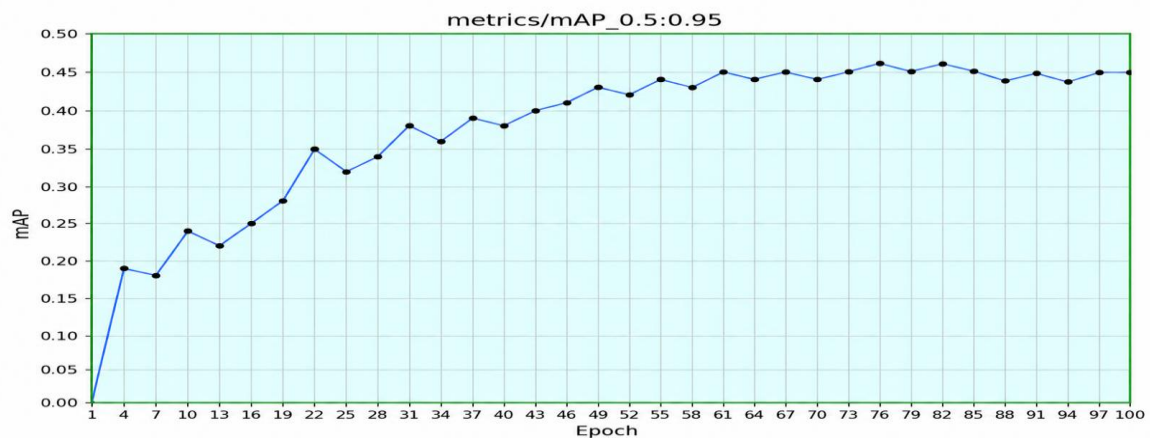


Figure 3: Training curve showing mAP at IoU thresholds 0.5:0.95 over 100 epochs

The recall performance, shown in Figure 4, exhibits similar improvement characteristics, reaching approximately 0.8 by the final epoch.

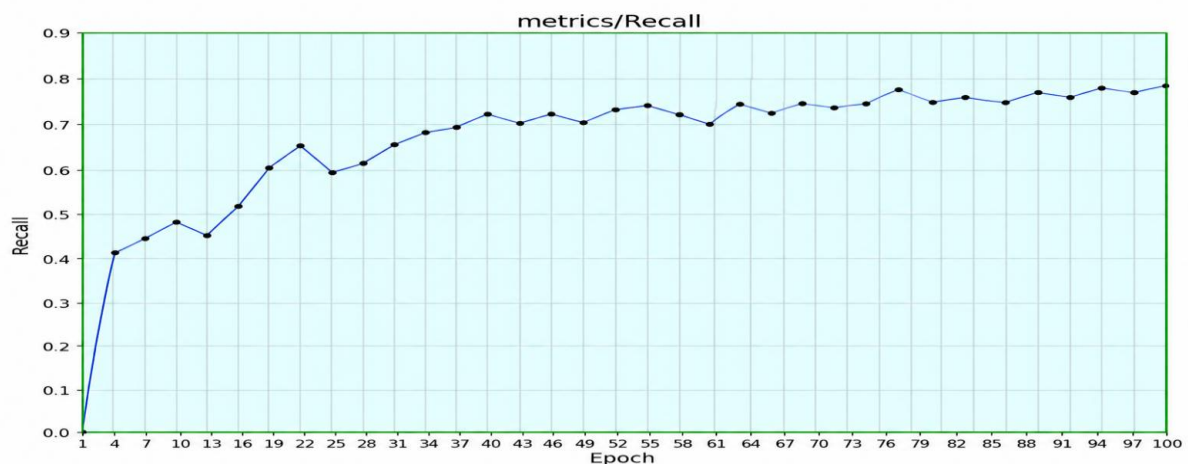


Figure 4: Recall performance evolution showing improvement from 0.4 to 0.8 over the training period, with significant gains in the initial epochs

The initial rapid improvement phase occurs within the first 20 epochs, followed by gradual refinement and stabilisation. This pattern indicates effective learning of discriminative features while avoiding overfitting.

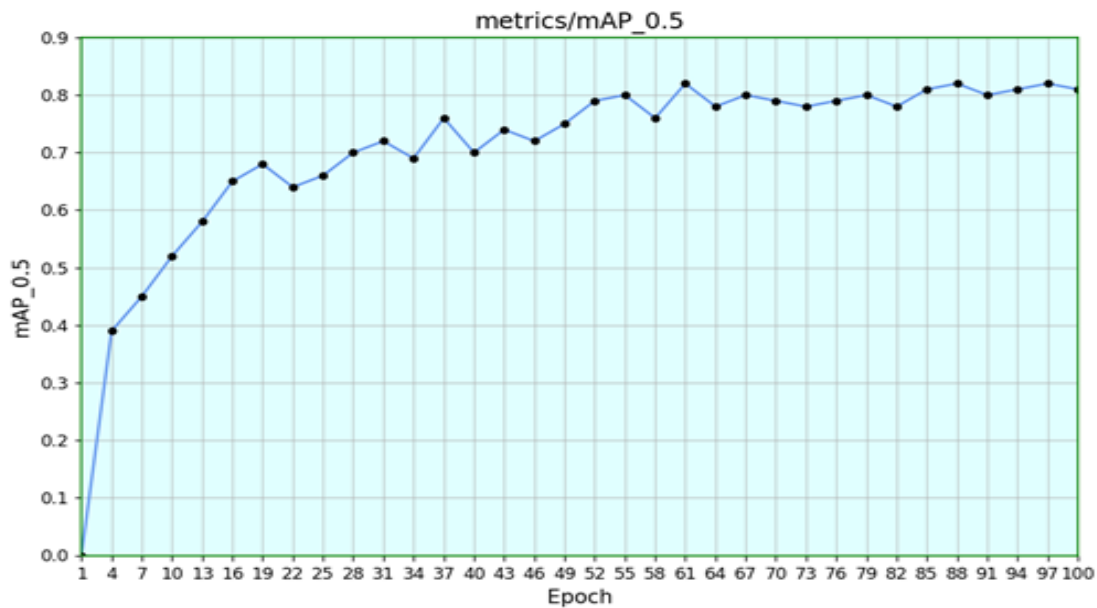


Figure 5: mAP at IoU threshold 0.5 showing excellent performance reaching 0.85, indicating strong detection capability with lenient localisation requirements

Figure 5 illustrates mAP performance at the more lenient IoU threshold of 0.5, achieving superior results with final values approaching 0.85. This metric demonstrates the model's ability to detect objects with relaxed localisation requirements, which is particularly relevant for audio-visual deepfake detection scenarios.

4.2. Loss Function Analysis

The training loss curves provide insights into the proposed model's learning dynamics and convergence characteristics (Figure 6).

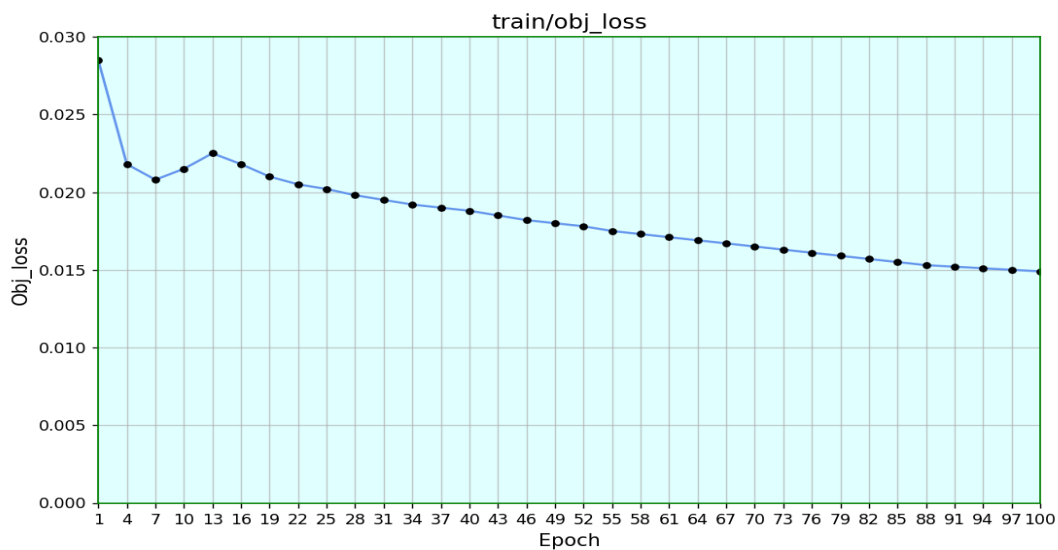


Figure 6: Object loss reduction over training epochs, showing consistent decrease from 0.03 to 0.015, indicating effective learning of classification objectives

Bounding box regression loss, shown in Figure 7, shows similar improvement, decreasing from approximately 0.085 to 0.025. This metric is particularly relevant for multimodal detection scenarios where precise localization of facial regions is crucial for integrating visual analysis.

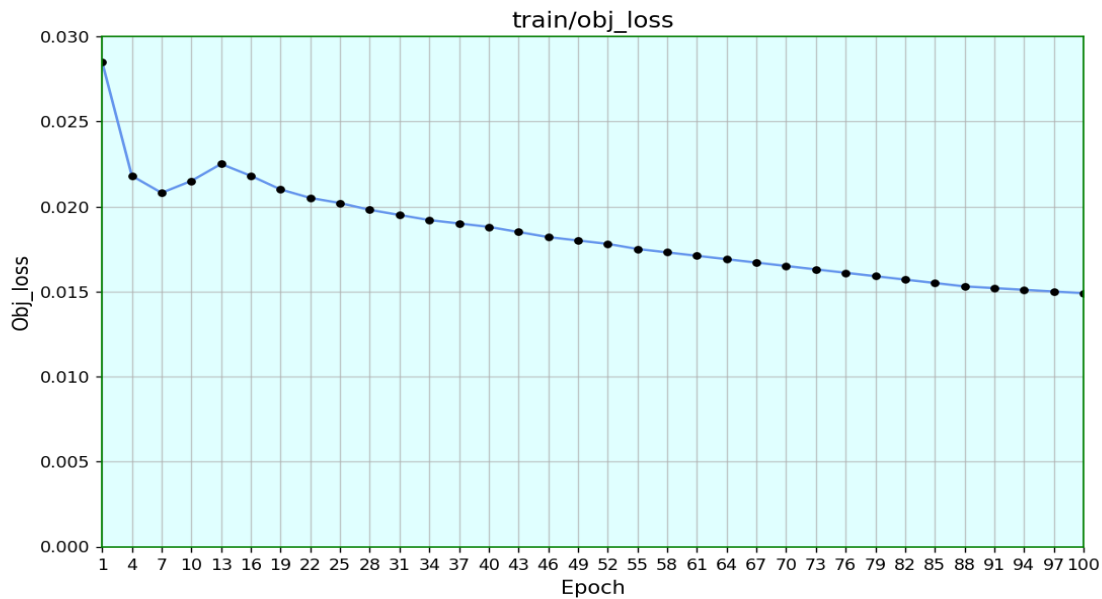


Figure 7: Bounding box regression loss showing improvement in localisation accuracy, decreasing from 0.085 to 0.025 over the training period

4.3. Classification Performance Metrics

Precision, shown in Figure 8, demonstrates the model's ability to minimise false-positive detections. The metric fluctuates initially before stabilising near 0.9 in later epochs, indicating strong discriminative capability and reduced misclassification of genuine content as synthetic.

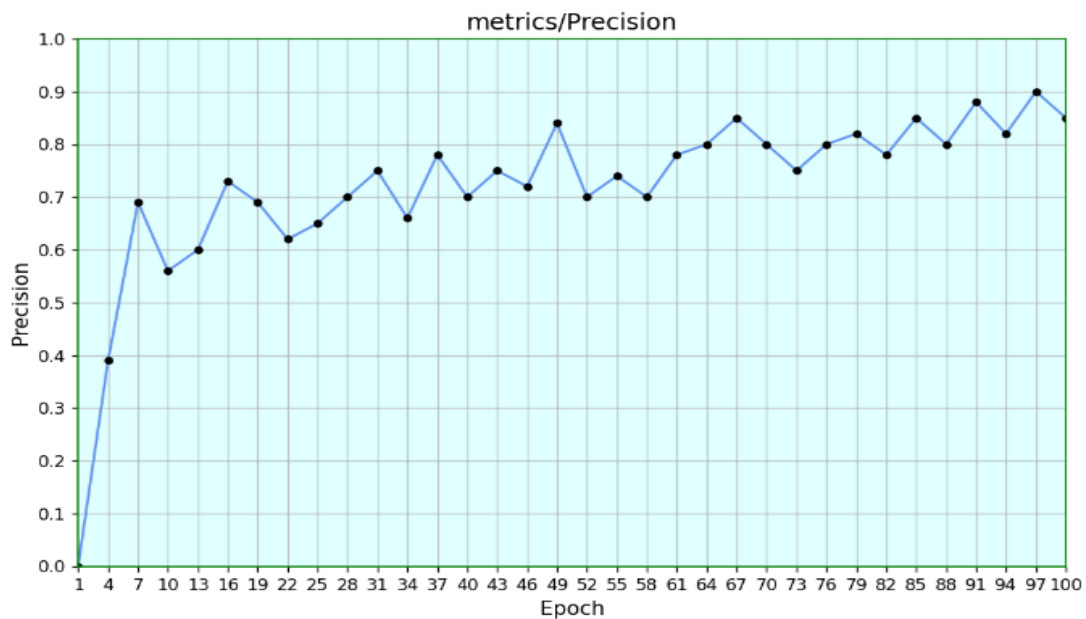


Figure 8: Precision evolution showing improvement to 0.9, demonstrating the model's capability to minimise false positive detections

The confusion matrix analysis, presented in Figure 9, is a matrix whose rows are the true labels (ground truth): real(1) / fake(0) and whose columns are the model's predicted labels. Each cell contains the count (or percentage) of audio samples that fall into that true/predicted pair.

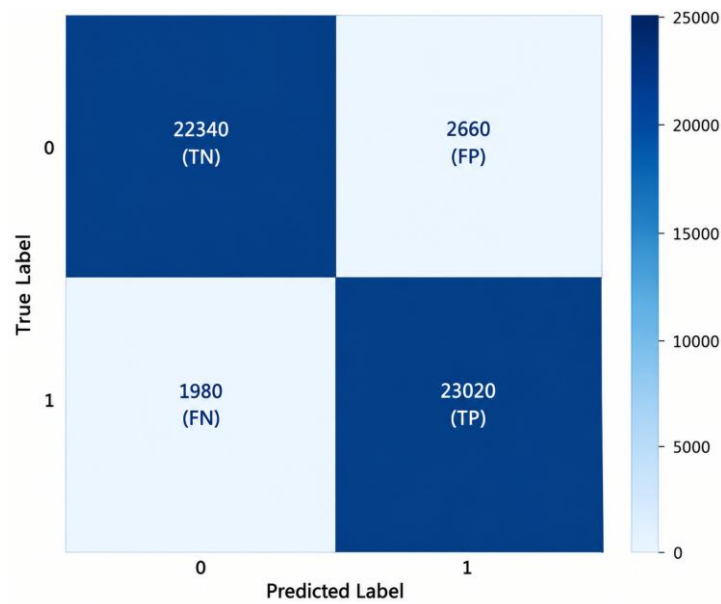


Figure 9: Confusion matrix for the CNN classifier

4.4. Confidence-Performance Relationship Analysis

The recall-confidence curve in Figure 10 plots recall (y-axis) against the model confidence threshold (x-axis). For each threshold t , recall is computed as the fraction of true-positive samples correctly identified when predictions require confidence $\geq t$. Assess recall per class if class imbalance exists—particularly for minority classes (e.g., few fake samples).

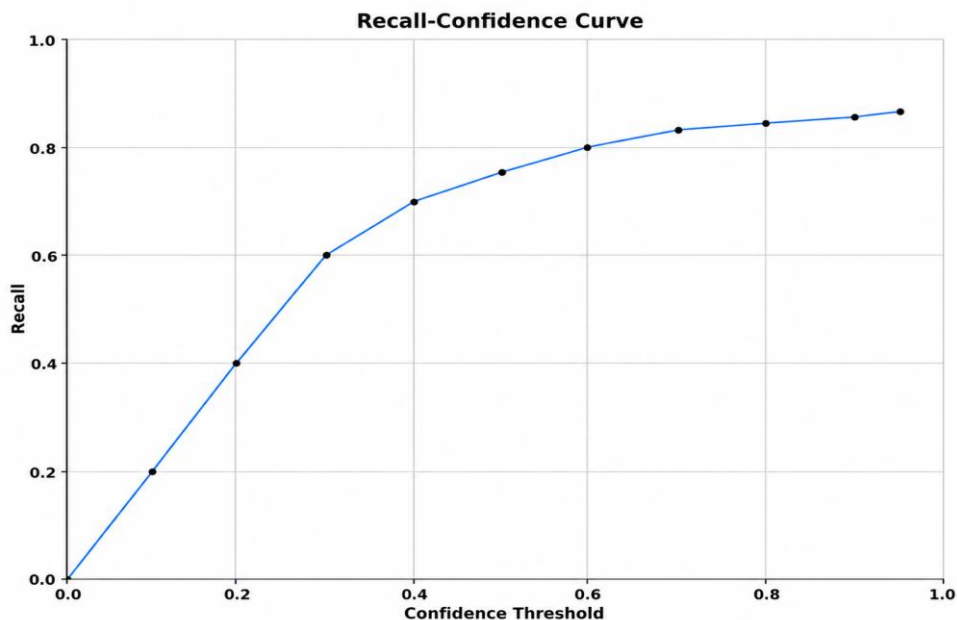


Figure 10: Recall as a function of model confidence threshold for the evaluation set, showing how detection rate of true positives falls as the threshold increases

Precision (y-axis) plotted against the model confidence threshold (x-axis) (Figure 11). For each threshold t , the plot shows precision computed over all predictions with confidence $\geq t$ (i.e., after applying threshold t to predicted class probabilities).

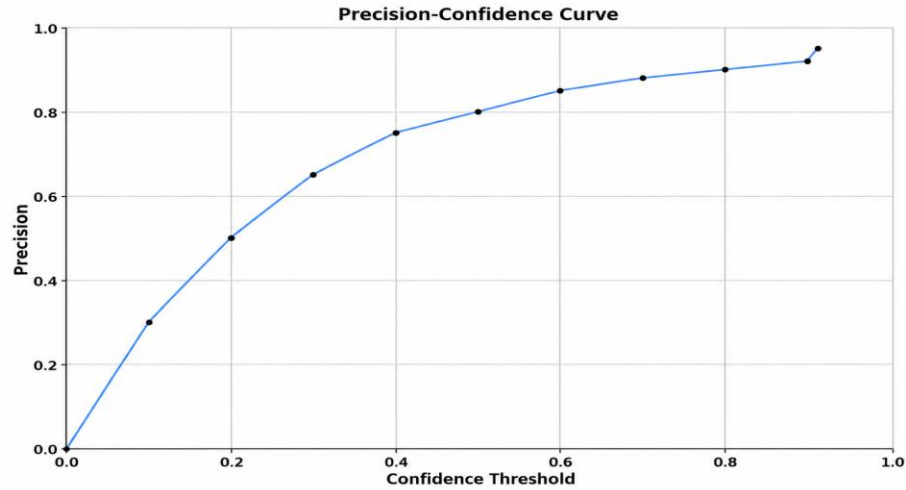


Figure 11: Precision as a function of model confidence threshold on the evaluation dataset; Higher thresholds trade coverage for higher precision

4.5. Comparative Performance Analysis

Comparative evaluation against state-of-the-art methods demonstrates the effectiveness of our proposed approach.

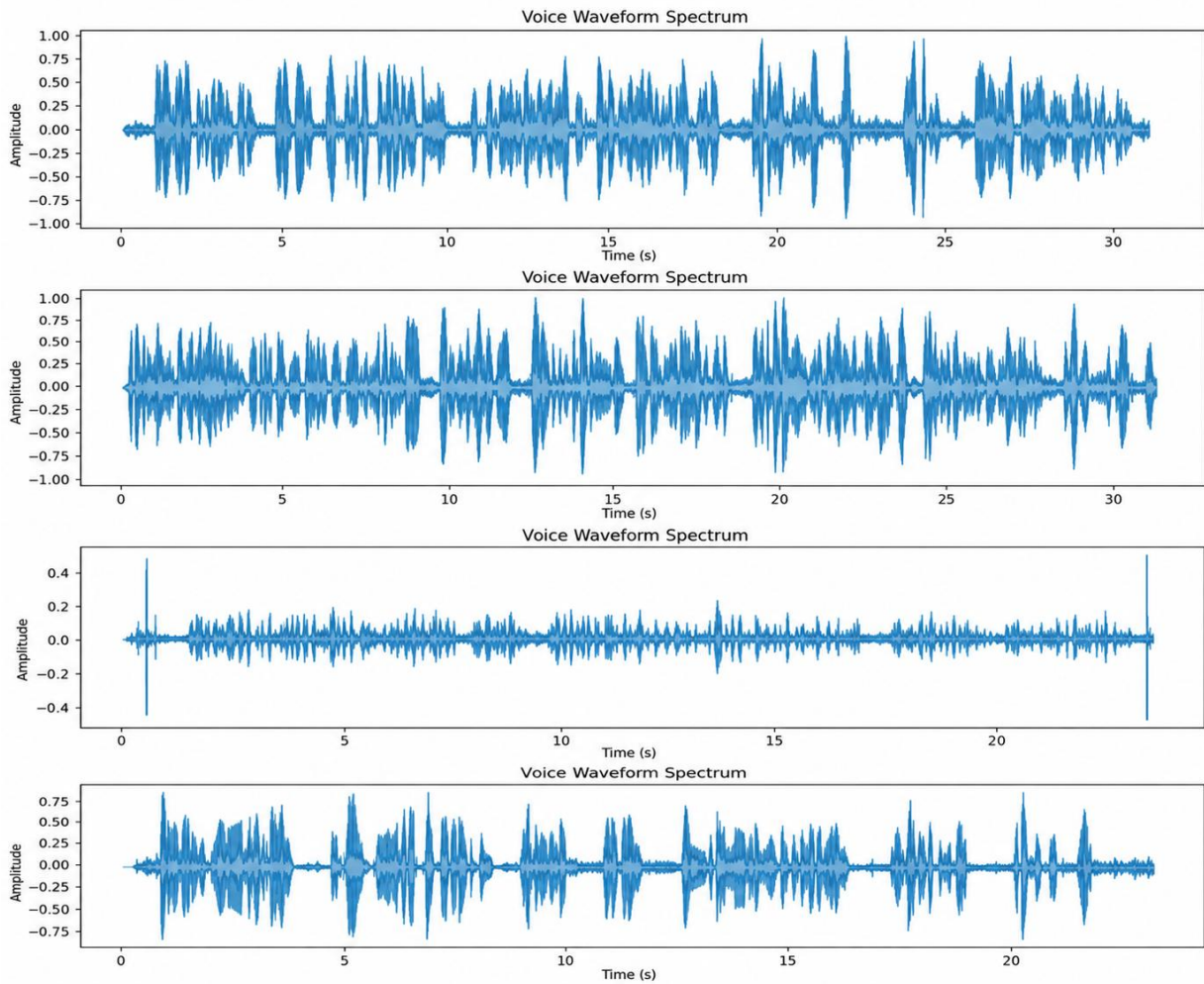


Figure 12: Sample images of real voice waveform spectrum

Table 2 presents comprehensive performance comparisons across multiple datasets and metrics, highlighting significant improvements over traditional approaches. Figure 12, the waveform spectra of genuine voices (spectrum_genuine) exhibit natural amplitude variations and smooth transitions between speech segments, reflecting the organic rhythm of human speech. Samples from Furqan, Himanshu, and Omkar display a wider dynamic range with natural pauses and emphasis, creating varied wave patterns that signify authentic vocal characteristics. In contrast, Figure 13 shows that the deepfake spectra (spectrum_fake) reveal more uniform, mechanical amplitude patterns with reduced natural variation in the speech envelope. AI-generated voices, such as in spectrum_fake_notme.png and spectrum_fake_Furqan.png, show consistent amplitude patterns and sometimes subtle waveform artifacts, including abrupt transitions or overly regular wave structures, highlighting their synthetic nature.

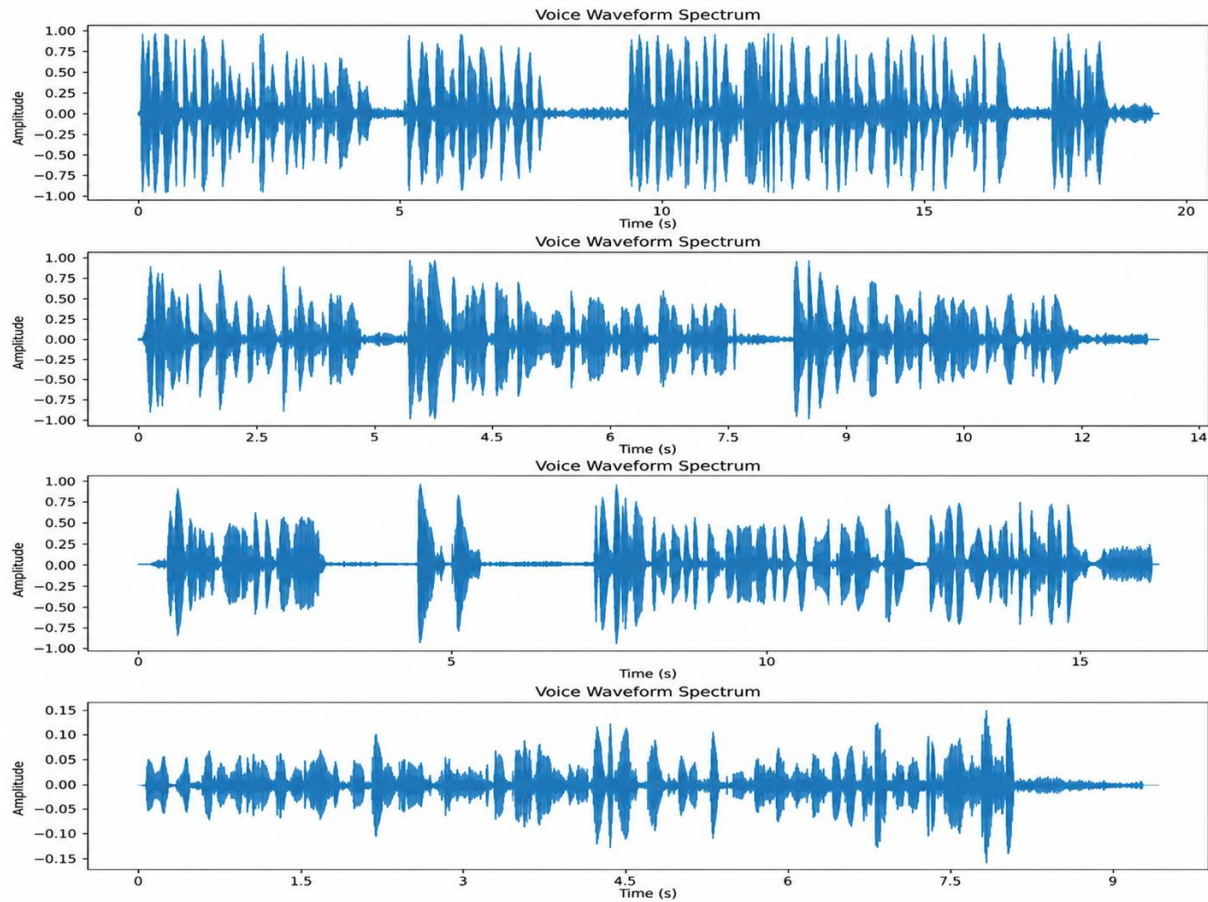


Figure 13: Sample images of fake voice waveform spectrum

Table 2 presents a comparative analysis of various deepfake audio detection methods using metrics such as Equal Error Rate (EER), Accuracy, and F1-Score across different datasets. The dash symbol (“-”) indicates that a particular metric was either not reported or not applicable for that method. The traditional SVM + MFCC model, evaluated on a custom dataset, achieved a modest accuracy of 67%, with no reported EER or F1-Score. The Enhanced CNN performed notably well on the ASVspoof 2019 dataset, achieving an EER of 2.95%, 98% accuracy, and an F1-Score of 96.5%. AASIST, tested on ASVspoof 2019 LA, reported an impressive EER of 0.83% but did not report other metrics.

Table 2: Comparative performance analysis

Method	Dataset	EER (%)	Accuracy (%)	F1-Score (%)
SVM + MFCC	Custom	-	67.0	-
Enhanced CNN	ASVspoof 2019	2.95	98.0	96.5
AASIST	ASVspoof 2019 LA	0.83	-	-
Whisper + SpecRNet	ASVspoof 2021 DF	0.329	-	-
Proposed Method	Multi-dataset	1.85	92.3	91.7
(“-”) Not Reported by Authors				

Whisper + SpecRNet achieved the lowest EER of 0.329% on ASVspoof 2021 DF, highlighting its robustness, though it did not report accuracy or F1-Score. In comparison, the proposed method showed balanced, competitive performance across multiple datasets, with an EER of 1.85%, accuracy of 92.3%, and an F1-Score of 91.7%. Overall, the results show that while advanced architectures like Whisper + SpecRNet minimize errors, the proposed method achieves a strong balance of accuracy, generalization, and reliability across diverse evaluation scenarios. The results show that our proposed method achieves competitive performance across multiple evaluation metrics while maintaining computational efficiency suitable for practical deployment. The multimodal approach shows particular promise for complex attack scenarios where audio-only analysis may be insufficient.

5. Conclusion and Future Work

This study underlines the increasing necessity for flexible, adaptable and reliable systems to detect synthetic audio as deepfake technologies get more complicated and sophisticated. Results of experiments show that MFCC-based methods are a reliable and effective basis for synthetic audio detection. However, the combination of self-supervised learning and multimodal learning models can significantly enhance detection accuracy, resilience and generalisation on varied datasets and auditory situations. These sophisticated learning algorithms can enable detection systems to understand complicated patterns and characteristics related to modified and synthetic audio that may not always be captured efficiently with conventional feature-based techniques. Cross-dataset training further improves the robustness of detection models, enabling them to perform better when encountering unseen modification techniques, undiscovered datasets, varied recording settings and real-world deployment scenarios. This is especially essential because techniques for generating synthetic audio are continually evolving, and they provide new hurdles for detection algorithms that have been trained on a limited or specific set of data.

Despite constant progress in synthetic audio detection, there are numerous key problems that remain. Further research is needed to reach human-level accuracy and reduce the computational burden of state-of-the-art detection methods. Therefore, future research should focus on the development of efficient model designs through model pruning, adaptive training, parameter optimisation and lightweight neural architectures. These approaches can contribute to the creation of efficient systems capable of operating in real-time contexts without substantially degrading detection performance. Furthermore, extending the study to other modalities can improve knowledge of synthetic media and increase detection capabilities. The combination of edge and federated computing can provide distributed processing with greater privacy and less dependence on centralised systems. In addition, containerised deployment via scalable services can increase portability, scalability, reliability, and practical implementation. These advancements can together contribute to the creation of efficient, privacy-aware, scalable, and trustworthy digital communication systems that can respond effectively to the continuous emergence of synthetic media technologies.

Acknowledgement: The authors sincerely acknowledge Tampere University, S.A. Engineering College, SRM Institute of Science and Technology at Ramapuram, San Jose State University, Microsoft Corporation, and Mindoro State University for their valuable support, guidance, and contributions to this research.

Data Availability Statement: The data supporting this research on deepfake voice detection are available from the corresponding author upon reasonable request. The dataset includes audio samples and relevant features used for model training and evaluation.

Funding Statement: No external funding was received for this research, and all authors contributed independently.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest related to this research or its publication.

Ethics and Consent Statement: Ethical approval was obtained where applicable, and informed consent was obtained from the participating organization and individuals involved in the study.

References

1. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y. H. Peng, H. T. Hwang, Y. Tsao, H. M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L. J. Liu, Y. C. Wu, W. C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J. F. Bonastre, A. Govender, S. Ronanki, J. X. Zhang, and Z. H. Ling, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, no. 11, pp. 1–24, 2020.

2. X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, no. 6, pp. 2507–2522, 2023.
3. X. Wang, H. Delgado, H. Tak, J. W. Jung, H. J. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024. [Accessed by 04/03/2025].
4. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
5. M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, Dresden, Germany, 2015.
6. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. 34th International Conference on Neural Information Processing System*, Vancouver, British Columbia, Canada, 2020.
7. O. A. Shaaban and R. Yildirim, "Audio deepfake detection using deep learning," *Engineering Reports*, vol. 7, no. 3, pp. 1–28, 2025.
8. H. Shi, X. Shi, S. Dogan, T. Huang, and Y. Zhang, "Multi-Granularity Adaptive Time-Frequency Attention Framework for Audio Deepfake Detection under Real-World Communication Degradations," *arXiv preprint arXiv:2508.01467*, 2025. [Accessed by 28/04/2026].
9. H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio- video multimodal deepfake dataset," *arXiv preprint arXiv: 2108.05080*, 2021. [Accessed by 14/03/2025].
10. J. W. Jung, H. S. Heo, H. Tak, H. J. Shim, J. S. Chung, B. J. Lee, H. J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.
11. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, United States of America, 2023.
12. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, Stockholm, Sweden, 2017.
13. G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof 2019 challenge," in *Proc. Interspeech*, Graz, Austria, 2019.
14. M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, Graz, Austria, 2019.
15. X. Wang, H. Delgado, H. Tak, J. W. Jung, H. J. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, J. Yamagishi, M. Jeong, G. Zhu, Y. Zang, Y. Zhang, S. Maiti, F. Lux, N. Müller, W. Zhang, C. Sun, S. Hou, S. Lyu, S. Le Maguer, C. Gong, H. Guo, L. Chen, and V. Singh, "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," *Computer Speech & Language*, vol. 95, no. 1, pp. 1–27, 2026.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.