

RESEARCH ARTICLE

OPEN ACCESS

A MULTI-DOMAIN FEATURE FRAMEWORK FOR ROBUST DEEPFAKE AUDIO DETECTION

Akshat Chhatrivala^{1*}, Ishita Akolkar², Namrata Shroff³, Shankar Parmar⁴

^{1, 2, 3}Computer Engineering Department, Government Engineering College, Bharuch, India,

⁴Electronics and Communication Engineering Department, Government Engineering College, Godhra, Gujarat, India

¹<http://orcid.org/0009-0005-1256-3075>, ²<http://orcid.org/0009-0002-3237-2565>

³<http://orcid.org/0000-0001-6343-6205>, ⁴<http://orcid.org/0000-0002-3034-5144>

E-mail: 220140107009@gtu.edu.in, 220140107002@gtu.edu.in, namrata.gecb@gmail.com, shankar.parmar@gtu.edu.in

ARTICLE INFO

Article History

Received on: March 04, 2026

Reviewed on: April 02, 2026

Accepted on: July 01, 2026

Published on: July 31, 2026

Keywords:

Deepfake Audio Detection,
Synthetic Speech,
Multi-Domain Feature Fusion,
Audio Forensics

ABSTRACT

Audio deepfakes generated by modern text-to-speech and voice conversion systems pose serious threats to security, privacy, and trust in digital communication. This study proposes a multi-domain feature fusion framework for robust deepfake audio detection under realistic, in-the-wild conditions. A large-scale dataset comprising 31,780 audio samples, evenly split between genuine and synthetic speech and covering diverse speakers, languages, and recording environments, is utilized. Acoustic, compression-related, emotional, phase-based, prosodic, and statistical-spectral features are extracted and fused, and classification is performed using a lightweight fully connected neural network evaluated via stratified five-fold cross-validation. The proposed system achieves an average validation accuracy of 98.03% and an AUC of 0.998, demonstrating strong and stable discriminative performance. Ablation experiments and SHAP-based analysis highlight the critical role of compression-related features in enhancing robustness when combined with other feature domains. While the study focuses on audio-only detection and does not address adversarial or multimodal scenarios, the results indicate that multi-domain feature fusion offers a practical and generalizable solution for real-world deepfake audio detection, particularly in environments involving diverse codecs and synthesis techniques.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The evident growth of generative models shows significant transformation in the field of synthetic media, leading to the creation of hyper-realistic audio content commonly referred to as deepfake audio. These deepfake audios or artificially manipulated speech signals convincingly create a replica of the voice, tone, and prosody of the target speaker, even with minimal input data. Such advancement, even with certain legitimate applications in entertainment and personalized user interface, poses serious threats to privacy, security, and information integrity. In particular, the capacity of audio deepfakes to convincingly impersonate individuals has opened avenues for misinformation campaigns, identity theft, and high-stakes fraud [1]. Audio deepfakes are predominantly produced using neural architectures such as text-to-speech (TTS) synthesis, voice conversion (VC), and end-to-end speech generation using generative adversarial networks (GANs), variational autoencoders (VAEs), or transformers [2], [3], [4]. To capture the intricate patterns of human speech, large-scale datasets and complex acoustic modelling are being used. Near-human naturalness in synthesized speech and significant narrowing of the perceptual gap between real and fake audio have been demonstrated by recent TTS systems like VITS [5] and StyleTTS [6]. Consequently, the state-of-the-art automatic speaker verification (ASV) systems are vulnerable to adversarial exploitation. The threats associated with the audio deepfakes are not merely hypothetical but have posed real issues to society. These threats force the need to develop a reliable and explainable detection mechanism capable of performing under an adversarial environment, and real-world, noisy data [7]. Even with the evident increased attention towards audio deepfakes, this field remains under-explored compared to visual deepfakes.

The significant problems for the audio deepfake detection include the diversity of generative models and synthesis techniques, the lack of comprehensive and generalized datasets, and the difficulty of identifying consistent acoustic artifacts across variable conditions [8]. The state-of-the-art detection methods generally rely on spectral anomalies, prosodic irregularities, or embeddings learnt through deep neural networks. However, these models fail to be accurate beyond the specific conditions of the dataset under which they were developed, limiting their practical usage [9], [10]. This fast development of audio deepfake technology is a major challenge, which demands the advancement of robust detection methods that are capable of addressing both the technical and ethical implications of these manipulations [11], [12]. Through the application of a multi-domain feature fusion strategy utilizing acoustic, compression, emotional, phase, prosodic, and spectral features, this paper addresses these problems. The system aims at high accuracy and generalization in a wide range of conditions with a robust and interpretable deep neural network architecture. A large, in-the-wild dataset that has been carefully curated to reflect the conditions in the real world is employed for the study. The efficacy and interpretability of the proposed method are established through multiple experiments, setting a new benchmark for reliable audio deepfake detection [13]. In the subsequent sections, a comprehensive literature review is discussed in Section 2, the proposed methodology is described in detail in Section 3, experimental setup and findings are given in Section 4, key findings and future work are elaborated in Section 5, and a list of contributions is summarized in Section 6.

II. LITERATURE REVIEW:

Prior research in deepfake audio detection has evolved through distinct methodological paradigms, from spectral feature analysis to advanced temporal modeling and generalization strategies. This section synthesizes key advancements, highlighting their contributions, limitations, and persistent challenges in achieving robust, generalizable detection across diverse synthesis techniques and real-world conditions.

II.1 SPECTRAL FEATURE DOMINANCE:

A work on generalization in audio deepfake detection investigates CNN-based models trained on log-Mel spectrograms using the ASVspoof 2019 Logical Access corpus. The model is well-performing on familiar vocoders with an Equal Error Rate (EER) of around 1.4%, but there was a significant dip in performance around 18% EER when tested with unfamiliar vocoders. This is one of the biggest challenges of generalization since the models tend to overfit on spoof types [14]. A study employed MFCCs along with basic classifiers such as k-NN, SVM, and decision trees to identify deepfake TTS-synthesized speech. With 85–88% accuracy on the dataset it created, the paper demonstrated that although MFCCs offer a good baseline, they fall short in detecting more sophisticated or novel manipulations. This study thus benchmarks simple features and models for the detection of deepfakes [15]. A benchmark study on deep learning compared CNN, CRNN, LSTM, and Transformer architectures with spectrogram input. CRNN models showed optimal trade-off between accuracy and efficiency with EERs of around 3–4%, whereas Transformers demonstrated better generalization. Scalability and performance with real-time usage was the primary emphasis of this research with utilizing ASVspoof 2019 and WaveFake datasets [16]. SpecViT used a Vision Transformer on log-Mel spectrogram patches and included speaker embeddings. Outperforming CNNs with an EER of roughly 1.8%, SpecViT showed good generalization to out-of-dataset data. This method brought ViT-style token learning to time-frequency modeling for deepfake detection from datasets such as WaveFake and ASVspoof LA [2].

II.2 TEMPORAL AND MULTI-SCALE DYNAMICS:

A two-branch CNN-BiGRU model combined spectral and temporal evidence through cross-attention. With an EER below 2% on ASVspoof LA and DFDC-Audio, this technique emulated human perception processes by combining features dynamically in a multiscale manner. It offered an effective architecture for identifying varied audio deepfakes [17]. Another research emphasized voiced and unvoiced region analysis, demonstrating that fake cues are more prominent in voiced parts of speech. Through the evaluation of energy, zero-crossing rate, and variation of pitch in these parts, the research attained 91% accuracy with straightforward models. It presented a new outlook for localized temporal analysis for better detection efficiency [18]. A work utilized direct and reverberant forms of audio signals to enhance robustness in deepfake detection. By training a dual-branch model on both channels jointly with contrastive loss, it learned subtle cues from the reverberation and improved generalizability. Tested on LA and DF-TIMIT datasets, the model achieved an EER of 5.4%, better than single-channel methods and demonstrating the contribution of acoustic realism to fake audio detection [19]. Wave-spectrogram fusion was introduced in an attempt to better handle audio deepfake detection with improved generalizability. The WaveSpec model simultaneously processes raw waveform and spectrogram features with the help of multi-scale fusion and cross-modal alignment, which allows it to better identify out-of-domain forgeries. It obtained state-of-the-art results on ASVspoof2021 and In-The-Wild datasets, demonstrating improved resistance against diverse manipulation methods [20].

II.3 GENERALIZATION AND ROBUSTNESS STRATEGIES:

A study introduced a GAN-based anomaly detection technique trained on only real audio. Assuming counterfeited audio as an anomaly, the model had about 92% detection accuracy with generator loss as an anomaly score. The unsupervised technique prevents overfitting to a certain type of spoof and promises better generalization. The system was tested on the ASVspoof LA dataset [21]. A refinement strategy for the latent space coupled with manifold mixup augmentation was introduced to enhance generalizability in audio deepfake detection. Aligning the latent features and increasing training diversity, the technique minimized overfitting to individual deepfake generators. Evaluated on ASVspoof and FakeAVCeleb, the technique registered a 10–15% decrease in EER on novel attack types, highlighting the strength of feature-space regularization [22]. An incremental learning system for unsupervised domain adaptation was introduced in a paper to prevent performance degradation in dynamic audio scenarios. By employing a memory-light domain discriminator and contrastive learning, the model learns to adapt incrementally without access to target labels.

Cross-domain performance improved by up to 30% on ASVspoof and In-the-Wild datasets during evaluation, making its real-world deployment without retraining from scratch prominent [23]. A study proposed a two-branch adversarial training paradigm that localizes partial deepfake segments in audio through feature gradient refinement. The method improves contrast between real and artificial speech segments with a gradient-guided loss and sharpening in feature space. Evaluated on the Fake-or-Real and Partial-Fake datasets, it recorded more than 80% EER gain over baselines, which indicates its robustness in localizing partial manipulations in speech [24].

II.4 SPECIALIZED AND SURVEY PERSPECTIVES:

A study resolved the specific problem of identifying deepfakes in singing voices. Based on MFCC, pitch, jitter, and shimmer features, a traditional combination of random forest and MLP models attained 89% accuracy on a proprietary singing dataset. The study emphasized how voice distortions affect prosody differently in singing, thus prosodic features being more effective [25]. FakeSound study presented a deepfake general audio detection task and a highly realistic fake audio dataset. The dataset uses inpainting with AudioLDM and AudioSR, featuring spliced, regenerated, and partially manipulated audio. A benchmark model trained on the dataset outperformed human detection accuracy (<60%) and previous SOTA on ADD 2023, particularly on challenging half-truth samples and zero-shot situations [26]. A survey presented an in-depth overview of audio deepfake detection approaches, categorizing them as feature-based, model-based, and hybrid. The survey did not present experiments, but noted important gaps in the research. Certain gaps highlighted in the paper included cross-language generalization, low-resource manipulation, and interpretability. The ASVspoof and FakeAVCeleb datasets were noted as benchmarks of the field [16]. Although targeting visual deepfakes, this review offers cross-domain insights that can be applied to audio. It grouped detection techniques into frequency-based, biological signal-based, and transformer models, and highlighted challenges such as generalizability and real-time deployment.

The review offered hybrid multimodal architectures merging audio-visual cues for next-generation detection, hinting at cross-modal prospects for audio deepfake research [27]. A research centered on algorithm attribution for synthetic audio detection. It proposed a two-stage network incorporating attention-guided attribution maps and contrastive loss to distinguish synthesis methods. Based on datasets such as ASVspoof and In-the-Wild, the model yielded above 90% accuracy in recognizing the source algorithm, highlighting the significance of interpretability and fine-grained attribution in generalized detection environments [28]. Confidence calibration in deepfake audio detection was investigated when a calibration layer was added to CNN outputs. This had better reliability with domain shifts and reached an AUC of approximately 94%. The authors highlighted the need for calibrated confidence scores for deployment in real-world applications, utilizing the FakeAVCeleb dataset [29]. A new multi-stage approach based on contrastive learning and GAN-based data augmentation was proposed to make deepfake detection resilient for audio. Through the integration of pre-trained models (such as Wav2Vec2), HiFi-GAN for generating spoofed samples, and contrastive loss, this technique enhanced generalization against fine manipulations and novel spoof types. The proposed framework was compared against conventional PTM-based systems over ASVspoof 2019, In-The-Wild, and DECRO corpora and achieved better performance [30].

II.5 RESEARCH GAPS:

One of the most persistent and unsolved issues in audio deepfake detection is low generalizability to unknown spoofing techniques, particularly vocoders and synthesis pipelines not included in the training set. Most models work flawlessly on datasets such as ASVspoof 2019 if tested in-domain but suffer hugely when tested on out-of-distribution data or novel generation methods. This exposes a central flaw in constructing strong, universal detectors that can generalize over a wide range of manipulations. The area is missing techniques that abstract away from overfitting to dataset-specific signals and instead aim to capture deep, modality-invariant properties of natural speech. With the advent of strong speech representation models such as BERT, Wav2Vec2.0, and ECAPA, their incorporation into deepfake detection pipelines is still under-researched or under-optimized [31], [32], [33]. A limited number of papers work on semantic, emotional, or prosodic inconsistencies, which tend to be subtle but indicative of synthesis. Yet few models actually combine linguistic comprehension with acoustic modeling, leaving it to holistic, context-sensitive detection mechanisms. Furthermore, emotional regularization and pragmatic content (tone, irony, stress) are areas where deepfakes often fail, but these remain underutilized in detection systems. Most recent research draws upon ASVspoof, WaveFake, or other prominent though comparatively narrow datasets. There is an urgent need for more multilingual, cross-domain, and variable-spoken language, accent, codec, emotional, and realistic audio noise environments datasets. Very few research efforts are devoted to language transferability or speaker accent variation, which are essential for global usage.

Besides, experiment frameworks are mostly static and fail to mimic adaptive attackers or live generation conditions. Therefore, there is an increasing demand for dynamic, changing benchmark suites and ongoing learning evaluation tests. Interpretability is another indispensable but insufficiently addressed gap. Most deep models act as black boxes, providing outputs without explaining why a sample is counterfeit. This is undesirable in forensics, legal testimony, and high-stakes settings. A few works start to leverage tools such as Grad-CAM or saliency maps, but systematic human-understandable explanations are mostly missing. Moreover, uncertainty modeling and calibration of confidence are usually overlooked despite being essential for identifying whether a sample should be accepted, checked, or rejected. The research community does not have effective tools to quantify, visualize, and describe model behavior on difficult inputs. Overall, the surveyed literature reveals a clear methodological evolution from early spectral and MFCC-based baselines toward spectro-temporal fusion, anomaly-based learning, confidence calibration, and domain-generalizable training pipelines. While existing systems achieve strong in-domain performance on benchmarks such as ASVspoof, WaveFake, Fake-or-Real, FakeAVCeleb, and the recently introduced FakeSound, consistent gaps remain in cross-corpus robustness, low-resource and cross-lingual settings, and interpretable, deployment-ready detectors. To contextualize the above discussion, Table 1 provides a summary of prior works in audio deepfake detection.

Table 1: Summary of Prior Research Done.

Sr. No.	Paper Title	Methodology	Features Used	Dataset	Findings
1	[28]	CNN-based classifier on log-Mel spectrograms	Spectral (contrast, roll-off)	ASVspoof 2019	1.4% EER (in-domain) but 18% EER on unseen vocoders due to spoof-type overfitting
2	[15]	Classical ML (SVM, k-NN)	MFCC (spectral), ZCR	Custom voice dataset	85–88% accuracy; MFCCs good baseline but limited on sophisticated TTS
3	[34]	Taxonomy of fake gen. + detection	Broad coverage (not experimental)	Multiple (overview)	Identified gaps: cross-language, low-resource, interpretability
4	[23]	Benchmark: CNN, RNN, CRNN, Transformer	Spectrograms (statistical)	ASVspoof, WaveFake	CRNN optimal at 3–4% EER; Transformers better generalization
5	[26]	Feature extraction + SVM	ZCR, Spectral (entropy), Prosody (pitch)	FakeSound (proposed)	Outperformed human accuracy (<60%) and ADD 2023 SOTA
6	[21]	GAN trained on real audio only	Time-domain (RMS), Environment (artifacts)	ASVspoof 2019 (LA)	92% accuracy using generator loss; prevents spoof-specific overfitting
7	[25]	Feature extraction + ensemble (RF, MLP)	MFCC, Jitter, Shimmer, Pitch	Custom singing dataset	89% accuracy; prosodic features effective for singing deepfakes
8	[18]	Voiced vs. unvoiced region splitting + stats	ZCR, Pitch, Energy, Jitter	LA ASVspoof	91% accuracy; fake cues prominent in voiced regions
9	[2]	ViT on spectrogram patches	Spectral (token-level) + Speaker awareness	WaveFake, ASVspoof LA	1.8% EER; superior out-of-dataset generalization
10	[17]	Dual-branch CNN-RNN + fusion	Spectrogram (multi-scale), Temporal variation	LA, DFDC-Audio	<2% EER; emulates multiscale human perception

Source: Authors, (2026).

III.2 MULTI-DOMAIN FEATURE REPRESENTATION:

The authenticity of speech manifests across multiple interdependent acoustic and temporal dimensions. Deepfake generation techniques, despite remarkable progress, still fail to reproduce the complete hierarchy of correlations present in human speech — from micro-level spectral variations to macro-level prosodic flow. To capture these complementary cues, the proposed system constructs a multi-domain feature representation that integrates six distinct yet interconnected families of descriptors: acoustic, compression-related, emotional, phase-based, prosodic, and statistical/spectral features. Each family encapsulates a specific physical or perceptual aspect of speech production, and their fusion provides a discriminative basis for differentiating genuine and synthetic audio. Figure 1 presents an overview of this multi-branch feature-extraction pipeline.

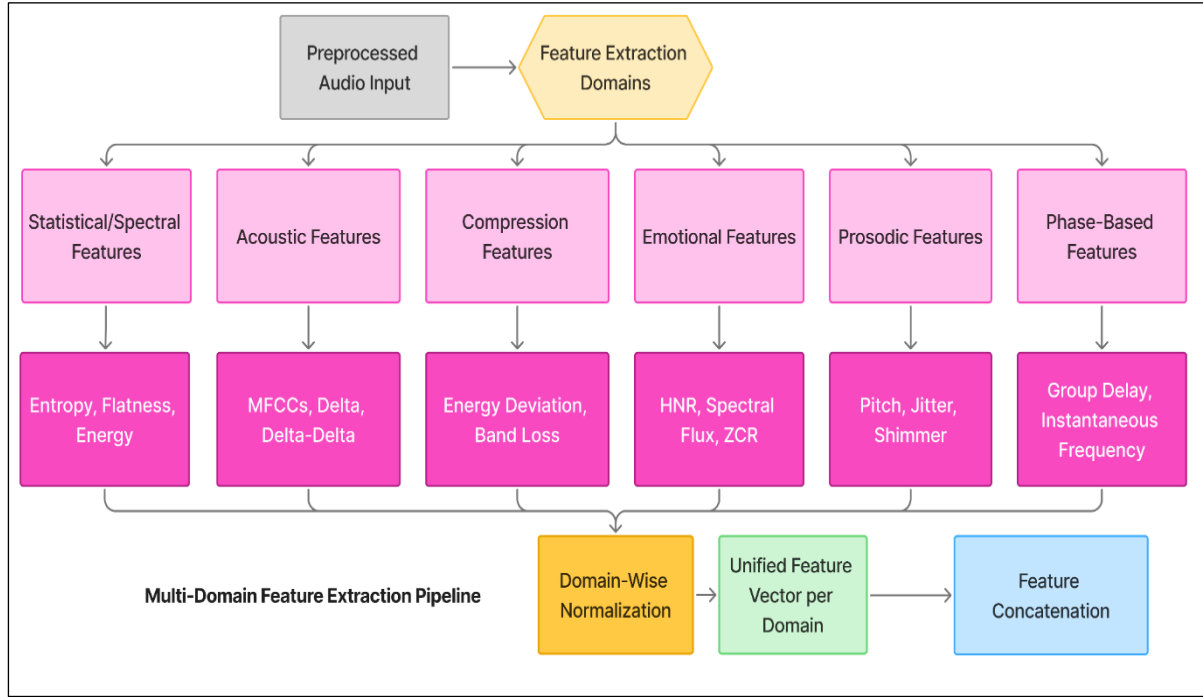


Figure 1: Multi-Domain Feature Extraction Pipeline.

Source: Authors, (2026).

III.2.1 Materials and Methods

Acoustic descriptors capture the spectral envelope and short-term resonance structure of the human vocal tract. Genuine speech exhibits smooth transitions in its spectral shape as the articulators move continuously; conversely, deepfake audio often suffers from vocoder-induced over-smoothing or discontinuous formant trajectories. To model these differences, we extract Mel-Frequency Cepstral Coefficients (MFCCs) using the Librosa library [35]. The waveform is divided into 25 ms Hamming-windowed frames with 10 ms overlap, and a 512-point Fast Fourier Transform (FFT) is applied to obtain the magnitude spectrum. The power spectrum is then filtered through 40 Mel-spaced triangular filters to approximate the human auditory frequency response. The logarithm of the filter-bank energies is decorrelated using a discrete cosine transform to produce 13 cepstral coefficients per frame.

Equation 1.

$$c_n(t) = \sum_{k=1}^K \log E_k(t) \cos \left[\frac{n\pi}{K} \left(k - \frac{1}{2} \right) \right] \quad (1)$$

Here, $E_k(t)$ denotes the Mel filter-bank energy in the k -th band at frame t , K is the total number of Mel filters, $c_n(t)$ is the n -th cepstral coefficient at frame t , and N_c is the number of retained MFCCs. To capture spectral dynamics, first- and second-order temporal derivatives (Δ and Δ^2 MFCCs) are computed, resulting in a 39-dimensional acoustic feature vector per frame. Frame-wise mean and variance statistics are finally pooled to obtain an utterance-level representation. This process emphasizes subtle spectral variations suppressed in synthetic speech. In natural recordings, the MFCC trajectory demonstrates smooth formant motion, while deepfake samples reveal reduced variance and abrupt spectral transitions due to neural vocoder quantization.

III.2.2 Compression-Related Features:

Deepfake audio often undergoes digital post-processing distinct from that of real recordings. Genuine recordings, captured through physical microphones, naturally preserve micro-level noise and non-linearities that interact predictably with audio codecs. Synthetic audio, generated numerically, exhibits different quantization behavior when compressed. To exploit this difference, compression-related features are computed by measuring how perceptual and spectral statistics vary before and after standard compression-decompression. Each audio signal is encoded using MP3 at 128 kbps, decoded back to waveform, and both versions are analyzed for spectral distortion, log-energy loss, and band-energy ratio differences. For every Mel band k , the energy deviation is defined as

Equation 2.

$$\Delta_k = \frac{E_k^{orig} - E_k^{comp}}{E_k^{orig}} \quad (2)$$

Where E_k^{orig} and E_k^{comp} denote Mel-band energies before and after compression. The mean, variance, and skewness of Δ_k across bands constitute the compression feature vector. In real audio, this deviation follows a smooth decay pattern; synthetic samples show irregular distortions because vocoder-generated signals lack the micro-redundancy that codecs exploit for efficient encoding.

III.2.3 Emotional Features:

Human speech carries rich emotional prosody resulting from complex coordination between vocal-fold vibration, subglottal pressure, and articulatory movement. Deepfake systems typically generate emotion using averaged style embeddings or prosody templates, which fail to reproduce the spontaneous micro-variations in tone and energy that characterize genuine affect. The openSMILE toolkit, configured with the emobase parameter set, is employed to extract emotional descriptors comprising a series of low-level descriptors (LLDs), including the harmonic-to-noise ratio (HNR), zero-crossing rate (ZCR), spectral flux, spectral centroid, and pitch-related functionals. For each LLD, a set of statistical functionals—namely mean, standard deviation, skewness, and kurtosis—is computed over the entire utterance, resulting in a consolidated emotional feature vector for subsequent analysis. HNR quantifies the ratio of periodic to aperiodic energy, reflecting voice quality and stability [36]. Synthetic voices often exhibit unnaturally high HNR due to vocoder smoothing, while real speech displays subtle irregularities associated with breathiness or emotional tension. Spectral flux and centroid, representing the rate of spectral change and the “brightness” of the signal respectively, further differentiate dynamic natural expressions from their flattened synthetic counterparts.

III.2.4 Phase-Based Features:

The phase spectrum encodes the fine temporal alignment of harmonics and reveals source–filter coherence in human phonation. While most speech synthesis methods accurately reconstruct magnitude spectra, they frequently neglect phase continuity, leading to temporal incoherence and phase jumps across frames. Phase information is extracted from the short-time Fourier transform (STFT) of the waveform. For each frequency bin f and time frame t , the instantaneous phase $\phi(f, t)$ is computed as the arctangent of the complex STFT components. From this, we derive the group delay and instantaneous frequency statistics:

Equation 3.

$$\tau_g(f, t) = -\frac{\partial \phi(f, t)}{\partial f}, \quad \omega_{inst}(f, t) = \frac{\partial \phi(f, t)}{\partial t} \quad (3)$$

Frame-level averages, standard deviations, and entropy of these quantities are then calculated over the spectral domain. Real speech shows consistent, smoothly varying group-delay trajectories governed by physiological glottal excitation, whereas deepfake audio demonstrates discontinuities or flattened phase patterns due to imperfect reconstruction. Capturing these discrepancies provides strong discriminative evidence of synthesis artifacts.

III.2.5 Prosodic Features:

Prosody encompasses intonation, loudness, rhythm, and temporal structure — the higher-level organization of speech beyond individual phonemes. It is a crucial indicator of naturalness, reflecting cognitive and emotional states. Synthetic voices often display uniform rhythm, quantized pitch patterns, and limited expressive range. Prosodic analysis is performed by computing fundamental frequency (F_0) contours, intensity, and temporal-rhythm descriptors. The F_0 contour is extracted via autocorrelation using Librosa’s `pitchtrack` function. Local perturbations in pitch and amplitude are summarized using jitter and shimmer metrics:

Equation 4.

$$Jitter = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|T_i - T_{i+1}|}{T} \quad (4)$$

Equation 5.

$$Shimmer = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|A_i - A_{i+1}|}{A} \quad (5)$$

Where T_i and A_i represent the period and amplitude of successive cycles. Aggregate descriptors such as mean F_0 , standard deviation, range, energy variance, and voiced/unvoiced ratio are also computed. Genuine recordings exhibit natural fluctuations in F_0 and intensity driven by respiratory control, while deepfake speech tends to produce smoother, less variable contours, revealing synthetic regularity in rhythm and stress.

III.2.6 Statistical and Spectral Features:

In addition to domain-specific attributes, global statistical and spectral summaries offer compact yet powerful characterization of the signal’s overall distribution. These metrics assess the entropy, randomness, and harmonic structure of the waveform, complementing the localized descriptors described above [37], [38]. We compute short-time energy (STE), spectral entropy (H_s), and spectral flatness (SF) from the STFT magnitude spectrum.

Equation 6.

$$H_s(t) = -\sum_f p(f, t) \log p(f, t) \quad (6)$$

Equation 7.

$$p(f, t) = \frac{|S(f, t)|^2}{\sum_f |S(f, t)|^2} \quad (7)$$

Equation 8.

$$SF(t) = \frac{(\pi_f |S(f, t)|^2)^{1/F}}{\frac{1}{F} \sum_f |S(f, t)|^2} \quad (8)$$

STE measures temporal energy variation; spectral entropy quantifies the uniformity of energy distribution; and spectral flatness captures the balance between tonal and noise-like components. Real speech, characterized by quasi-periodicity, exhibits low SF and moderate H_s values, while synthetic speech often yields higher flatness and entropy due to excessive smoothing and lack of micro-harmonic detail [39]. The multi-domain design provides a comprehensive view of both physiological and algorithmic aspects of speech production. Acoustic features capture vocal-tract behavior, prosodic and emotional cues reflect cognitive and expressive intent, phase-based features encode fine-grained temporal coherence, compression features exploit codec-sensitivity differentials, and statistical measures summarize global spectral organization [40]. By combining these complementary cues, the proposed representation detects inconsistencies that remain imperceptible in any single domain. This holistic integration is essential for countering the rapid evolution of generative speech models, many of which may correct one type of artifact while still revealing another. Consequently, the multi-domain representation forms a robust, interpretable foundation for the deepfake detection framework that follows.

III.2 MODEL ARCHITECTURE:

The fused multi-domain representation is used as input to a fully connected neural network optimized for binary deepfake classification. The network architecture is deliberately lightweight to emphasize interpretability and generalization while avoiding the over-parameterization that commonly affects high-dimensional feature sets. Figure 1 illustrates the overall network design.

III.3.1 Architectural Design

The classifier consists of three layers: an input layer, a single hidden layer of 128 neurons, and an output layer that produces posterior probabilities for the two target classes—genuine and synthetic audio. Each hidden unit employs the Rectified Linear Unit (ReLU) activation function to introduce non-linearity and enhance feature discrimination. To mitigate overfitting and encourage generalization, a dropout layer with a probability of 0.3 is applied after the hidden layer, randomly deactivating a subset of neurons during training. The final layer uses a softmax activation, transforming the linear output into normalized probabilities:

Equation 9.

$$\hat{y} = \text{softmax}(W_2 F_{\text{fused}} + b_1) \quad (9)$$

Where $h_1 = \text{ReLU}(W_2 F_{\text{fused}} + b_1)$ denotes the hidden representation of the fused feature vector F_{fused} , and W_i, b_i are the trainable parameters of the network. This shallow structure is intentionally chosen: it provides sufficient representational capacity to model non-linear relationships among features while remaining interpretable and computationally efficient.

III.3.2 Training Objective and Optimization:

Training is performed using the cross-entropy loss function, defined for a binary problem as

Equation 10.

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (10)$$

Where y_i represents the ground-truth label and $\hat{y}_i$ the predicted probability of the sample being genuine. This loss encourages the model to maximize the likelihood of correct classification while penalizing confident mispredictions. Optimization is conducted with the Adam optimizer (learning rate = 1×10^{-3} , $\beta_1 = 0.9$, $\beta_2 = 0.999$). To facilitate smooth convergence and adaptive exploration of the parameter space, a cosine-annealing learning-rate scheduler with warm restarts is applied, gradually reducing the learning rate within each epoch cycle. This scheduling approach prevents premature convergence and helps the model escape shallow local minima. The network is trained for up to 15 epochs with early stopping activated if the validation loss does not improve for five consecutive epochs. The early-stopping criterion ensures optimal trade-off between bias and variance, effectively halting training before overfitting begins.

III.3.3 Batching and Data Handling:

Training is conducted with a batch size of 128, using stratified batching to maintain class balance within every mini-batch. This prevents gradient bias arising from unequal class representation and ensures stable optimization. All inputs are standardized before training and shuffled at the start of each epoch to improve randomness and reduce dependency on sample ordering. To ensure deterministic reproducibility, all random seeds across Python, NumPy, and PyTorch environments are fixed. This guarantees consistent results across multiple runs—a crucial requirement for forensic and reproducible AI research. Figure 2 illustrates the end-to-end workflow. The normalized and fused features are loaded into mini-batches, propagated through the dense network, and optimized via back-propagation. Validation accuracy and loss are monitored at each epoch; the model achieving the lowest validation loss is retained for testing. This disciplined training regimen—combining early stopping, stratified data handling, and adaptive learning rates—ensures a stable and reproducible optimization process. The trained model subsequently undergoes evaluation.

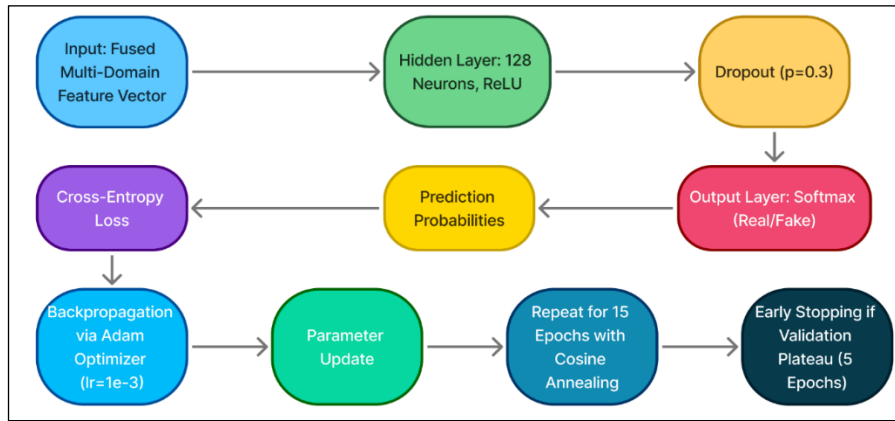


Figure 2: Neural Network Architecture and Training Workflow.

Source: Authors, (2026).

IV. EXPERIMENTAL SETUP, IMPLEMENTATION, AND RESULTS

This section presents a detailed description of the experimental setup, implementation workflow, and evaluation results of the proposed multi-domain deepfake audio detection framework. The discussion is structured to ensure methodological clarity and reproducibility, beginning with dataset handling and preprocessing, followed by model configuration and training strategy, and finally presenting quantitative results and analytical observations. All experiments were conducted using a stratified five-fold cross-validation protocol to ensure robustness and statistical reliability of the reported performance.

IV.1 DATASET LOADING AND EXPERIMENTAL PARTITIONING:

The experiments were conducted on a large-scale in-the-wild audio dataset comprising a total of 31,780 audio samples, evenly divided between genuine (real) and synthetic (fake) speech. The dataset was curated to reflect realistic deployment conditions, incorporating diverse speakers, languages, recording environments, and multiple deepfake generation techniques. The balanced class distribution was intentionally maintained to prevent classification bias and to enable fair evaluation of both false acceptance and false rejection behavior. For experimental evaluation, the dataset was organized into two primary directories corresponding to real and fake audio classes. A stratified five-fold cross-validation strategy was employed, ensuring that each fold preserved the original class distribution. In each iteration, four folds were used for training while the remaining fold served as the validation set. This procedure was repeated until all folds were used once for validation, and the final performance was reported as the average across all folds. Such a validation strategy provides a robust estimate of generalization performance while reducing the dependency on any single data split.

IV.2 PREPROCESSING PIPELINE:

Prior to feature extraction and model training, all audio samples underwent a standardized preprocessing pipeline designed to minimize irrelevant variability while preserving speech-specific characteristics critical for deepfake detection. Voice Activity Detection (VAD) was applied using the Silero VAD model to remove silent and non-informative regions from the recordings. This step ensured that subsequent feature extraction focused solely on speech-bearing segments. All processed audio signals were resampled to a uniform sampling rate of 16 kHz to standardize temporal resolution across samples. Given the natural variability in recording durations, each audio clip was either zero-padded or truncated to a fixed duration of five seconds. This uniform length constraint enabled consistent batch processing and ensured compatibility with the feature extraction and classification pipeline. The preprocessing stage was optimized for large-scale execution using GPU acceleration and mixed-precision computation, allowing efficient handling of the dataset without altering signal fidelity.

IV.3 FEATURE EXTRACTION AND REPRESENTATION

Following preprocessing, each audio sample was transformed into a comprehensive multi-domain feature representation as described in Section 3. Feature extraction was performed independently for each domain, and utterance-level representations were constructed using statistical pooling. The extracted features included acoustic descriptors, compression-related features, emotional features obtained via openSMILE, phase-based features derived from the short-time Fourier transform, prosodic attributes such as pitch and perturbation measures, and global statistical and spectral descriptors. All extracted features were concatenated to form a unified fused feature vector representing each audio sample. Before classification, the feature vectors were standardized to zero mean and unit variance to ensure stable optimization and prevent dominance of any single feature domain. This fused representation served as the sole input to the neural classifier, allowing the model to jointly exploit complementary cues arising from multiple aspects of speech production and synthesis artifacts.

IV.4 MODEL ARCHITECTURE AND TRAINING CONFIGURATION

The classification model employed in this study is a lightweight fully connected neural network specifically designed to balance discriminative capacity with interpretability and computational efficiency. The network accepts the fused multi-domain feature vector as input, followed by a single hidden layer consisting of 128 neurons activated using the Rectified Linear Unit (ReLU) function.

A dropout layer with a probability of 0.3 is applied to the hidden layer during training to mitigate overfitting and enhance generalization. The output layer consists of two neurons corresponding to the genuine and synthetic classes, with softmax activation producing normalized posterior probabilities. The input layer receives a fused multi-domain feature vector of fixed dimensionality, formed by concatenating all extracted feature streams. Model training was performed using the cross-entropy loss function optimized via the Adam optimizer with a learning rate of 1×10^{-3} . A cosine annealing learning-rate scheduler with warm restarts was employed to promote stable convergence and avoid premature stagnation. Training was conducted for a maximum of 15 epochs, with early stopping activated if the validation loss failed to improve for five consecutive epochs. Mini-batches of size 128 were used, and stratified batching ensured balanced class representation within each batch. All experiments were implemented in PyTorch and executed on Google Colab Pro using an NVIDIA Tesla T4 GPU.

IV.5 OVERALL QUANTITATIVE PERFORMANCE

The model achieved consistently high performance across all five validation folds, with accuracies ranging from 97.79% to 98.28%, yielding an average validation accuracy of 98.03% (± 0.18). The mean Area Under the ROC Curve (AUC) was 0.998, indicating near-perfect separability between genuine and synthetic audio samples. Table 2 summarizes the fold-wise results for Accuracy and Loss. The narrow range of variation across folds demonstrates the model’s stability and reproducibility—a result of stratified partitioning, balanced batch training, and early stopping strategies. The high AUC value confirms that the model maintains discriminative strength across multiple threshold values, suggesting that it does not rely on a narrow confidence range to identify synthetic speech. The F1-score averaged 0.981, reflecting a balanced trade-off between detection rate and false-alarm control. Together, these results confirm that the proposed architecture—despite being shallow and computationally efficient—achieves state-of-the-art performance on a complex, real-world dataset.

Table 2: Validation Accuracy and Loss during each fold.

Folds	Validation Accuracy	Validation Loss
1.	0.9807	0.006125
2.	0.9828	0.005840
3.	0.9812	0.005390
4.	0.9779	0.005600
5.	0.9807	0.006200
FINAL	0.9806	0.005831

Source: Authors, (2026).

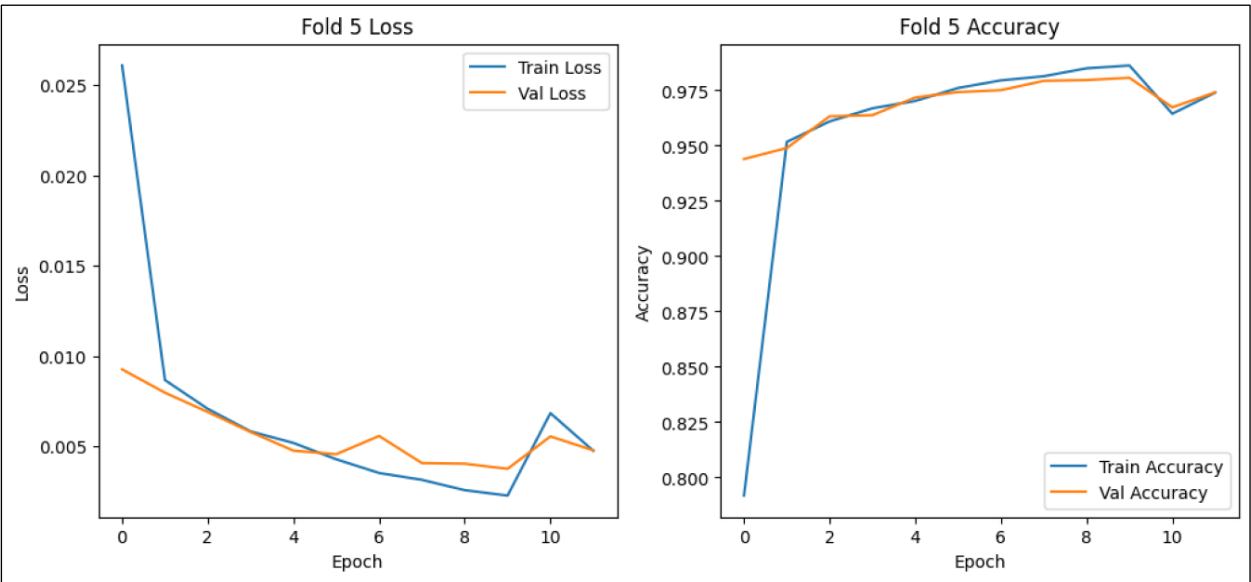


Figure 3: Training vs Validation Accuracy and Loss for Last Fold

Source: Authors, (2026).

IV.6 EVALUATION METRICS AND PERFORMANCE CRITERIA

To thoroughly assess the performance of the suggested deepfake audio detection system, a wide range of complementary metrics was utilized. These metrics offer subtle aspects of the system's discriminative ability, calibration, and class-specificity that are critical to evaluating robustness and reliability in real-life forensic and security applications. Accuracy is a baseline metric, measuring the proportion of correctly classified samples out of all predictions. Informative as this metric is, it can hide class-wise imbalances or trade-offs in errors, especially in light of the high-stakes application of deepfake detection. To better handle this, precision and recall metrics were calculated independently for every class.

Precision describes the ratio of correctly recognized positive instances to all positives predicted—quantifying the false positive control—whereas recall (sensitivity) takes note of the model's capacity to identify all actual positive instances, quantifying false negative rates. The F1-score, a harmonic mean between precision and recall, offers a balanced individual score of classification performance for each class. Given the approximately balanced dataset, macro-averaged (treating each class equally) and weighted-averaged (accounting for class sample sizes) versions of these statistics were also calculated to help summarize the class-level performance comprehensively. In addition to these point statistics, Receiver Operating Characteristic (ROC) curve and summary statistic Area Under the Curve (AUC-ROC) were used to analyze model separability at different classification thresholds. The model performed an average AUC close to 0.998, which presents almost perfect discrimination between original and manipulated audio samples. Complementarily, the Precision-Recall curve was also examined, which is especially relevant in security scenarios where false positive (false alarm) and false negative (forgery missed) costs are asymmetric. The PR curve again verified that high precision holds even at nearly complete levels of recall, highlighting practical application.

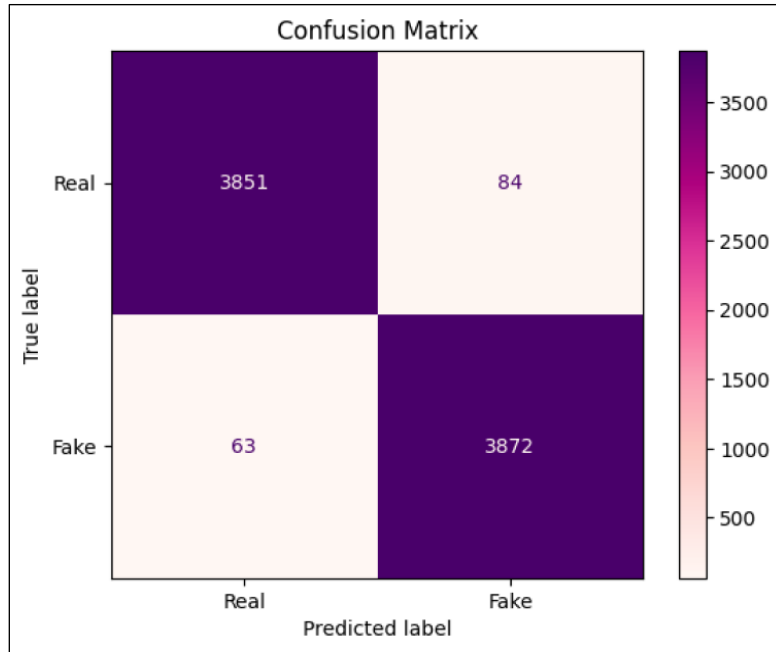


Figure 4: Confusion Matrix for the Validation Dataset.
Source: Authors, (2026).

Confusion matrices were constructed to provide an intuitive representation of the distribution of correct and incorrect predictions between classes, emphasizing the low but non-zero misclassification rates, which were mostly localized on difficult or borderline samples. Finally, predicted probability distributions and confidence margins were analyzed, illustrating excellent calibration of output probabilities and allowing reliable interpretations necessary for deployment environments calling for graded risk judgments. In combination, this set of performance metrics stringently confirms the accuracy, balance, and reliability of the detection framework. This complementary nature of the metrics, from baseline accuracy to advanced ROC-AUC and calibration, guarantees a rigorous and transparent performance estimation.

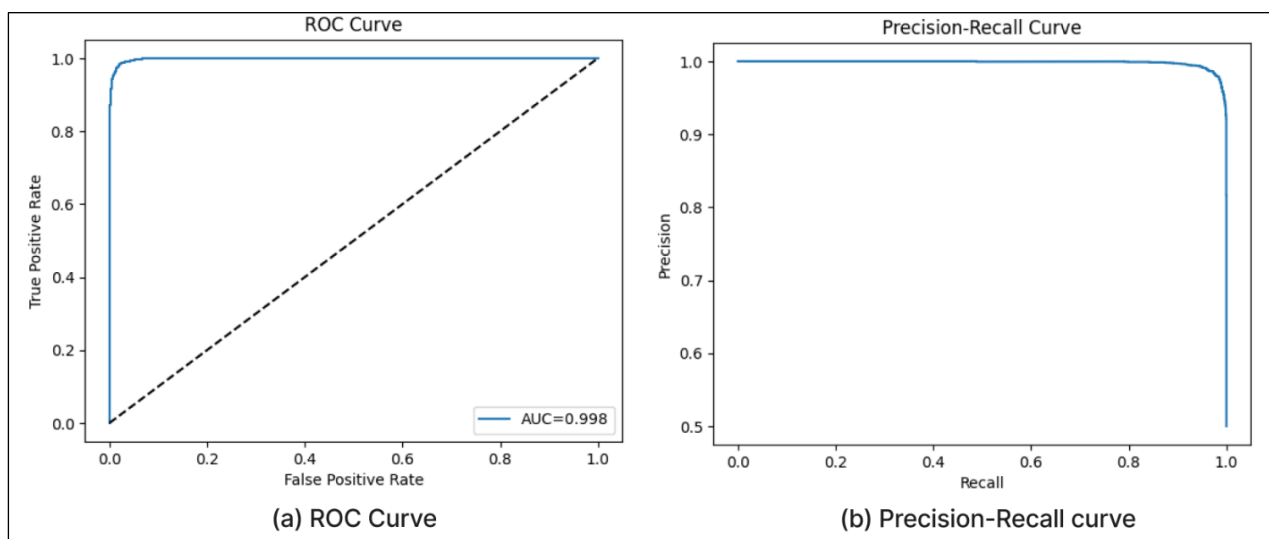


Figure 5: (a) Receiver Operating Characteristic (ROC) curve b) Precision-Recall.
Source: Authors, (2026).

ROC Curve of the suggested deepfake audio detection model averaged across five stratified folds. The excellent AUC value of 0.998 displays superb class separability. Precision-Recall curve indicating strong detection precision for varying recall levels, essential in forensic accuracy. To further examine the discriminative structure of the proposed multi-domain feature representation, a t-distributed Stochastic Neighbor Embedding (t-SNE) visualization was employed. While quantitative metrics such as accuracy and AUC provide aggregate performance evaluation, t-SNE offers an intuitive low-dimensional view of how genuine and synthetic audio samples are distributed in the learned feature space prior to classification. The t-SNE projection was generated using the fused multi-domain feature vectors extracted from the validation data. All samples were reduced from the original high-dimensional feature space into two dimensions using a fixed random seed to ensure reproducibility. No class information was provided to the dimensionality reduction algorithm, allowing the visualization to reflect the intrinsic structure of the learned representations rather than supervised separation. As shown in Figure 6, the t-SNE plot reveals a clear tendency toward class-wise clustering, with a substantial proportion of genuine and synthetic samples forming distinct regions in the projected space. However, partial overlap between the two classes is also observed, particularly in boundary regions. This overlap is expected given the use of real-world, in-the-wild audio data, where recording conditions, speaker variability, and high-quality synthesis techniques can reduce the prominence of individual artifact cues. Importantly, the presence of overlap does not contradict the strong classification performance reported earlier. Instead, it highlights the role of the proposed classifier in learning non-linear decision boundaries that effectively separate samples even when low-dimensional projections appear ambiguous. The high AUC and consistent fold-wise accuracy indicate that the classifier successfully exploits subtle multi-domain feature interactions that are not fully preserved in two-dimensional embeddings. Overall, the t-SNE visualization supports the quantitative findings by demonstrating that the fused feature space captures meaningful structure between genuine and synthetic audio, while also reflecting the inherent difficulty of the detection task under realistic conditions. This further validates the necessity of multi-domain feature fusion and supervised learning for robust deepfake audio detection.

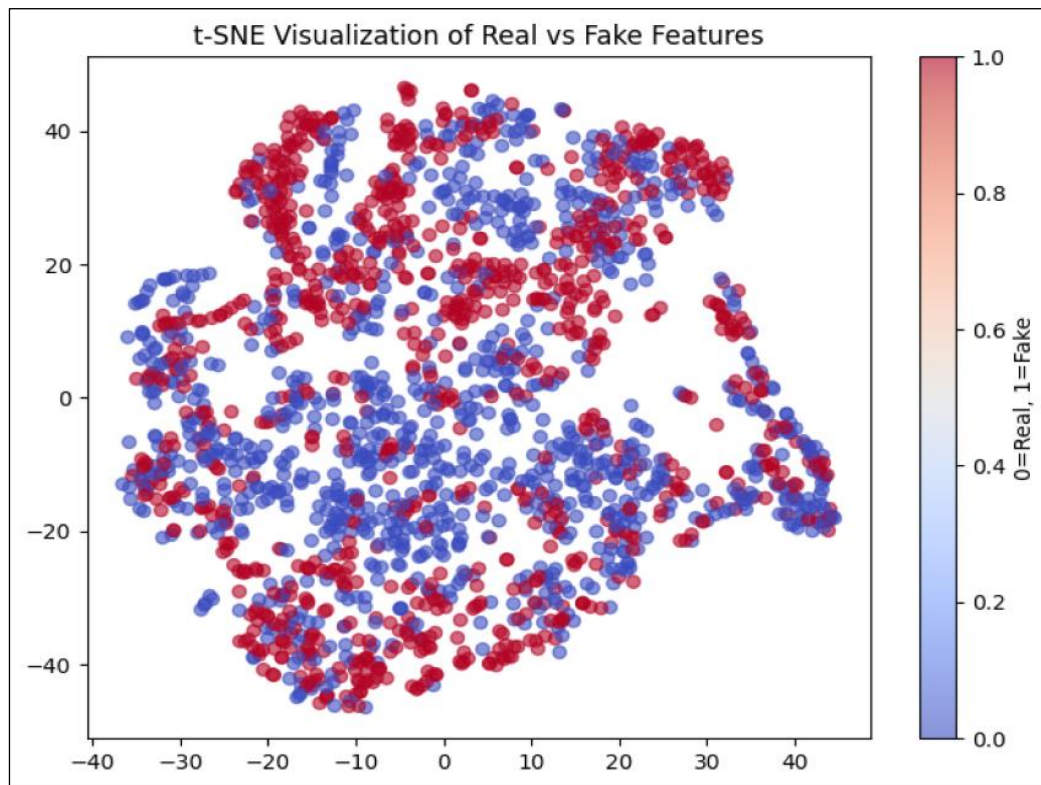


Figure 6: t-SNE visualization of fused multi-domain feature representations.
Source: Authors, (2026).

IV.7 ABLATION STUDY- IMPACT OF COMPRESSION ARTIFACTS:

To assess the unique contribution of compression artifact features to the detection system, an ablation study of their individual and combined effects was performed. Compression artifacts represent signal anomalies introduced through the encoding and transmission stages of audio, commonly challenging synthetic generation pipelines to replicate accurately. The objective was to learn how such features affect model performance both in isolation and as components of a fused multi-feature representation. The models were trained in three different settings: (i) baseline model trained on the complete fused feature set with all the feature domains, (ii) model trained only on compression artifact features, and (iii) model trained on fused features but not including the compression domain. All models were tested using stratified 5-fold cross-validation to maintain data split consistency and preprocessing. The performance in Table 3 illustrates evident performance differences between configurations. Significantly, the model that was trained on the combination of compression and acoustic features attained an optimal validation accuracy of 93.59%, with precision and recall balanced to a high extent on both real and fake classes. The classification report is as stated below:

Table 3: Classification Report on the Combination of Compression Artifacts Features and Acoustic features.

Class	Precision	Recall	F1-score	Support
0	0.9573	0.9126	0.9344	2357
1	0.9165	0.9593	0.9374	2357
Accuracy			0.9359	4714

Source: Authors, (2026).

Conversely, the model trained using only compression artifact features performed much lower—with the average accuracy at 56.43%. Precision and recall values for the two classes were moderate, but the F1-scores (0.6219 for class 0 and 0.4860 for class 1) reflect poor discriminatory ability when using only these features:

Table 4: Classification Report for only Compression Artifact Features.

Class	Precision	Recall	F1-score	Support
0	0.5493	0.7166	0.6219	2357
1	0.5924	0.4120	0.4860	2357
Accuracy			0.5643	4714

Source: Authors, (2026).

These results emphasize the subtle contribution of compression artifacts: although individually inadequate to induce high detection rates, when combined with other feature domains, notably acoustic features, they greatly improve model robustness. This interaction is likely to arise from compression features detecting peculiar encoding discrepancies inaudible to purely acoustic or prosodic descriptors. By and large, the ablation study affirmatively underscores the need for multi-domain feature fusion, with compression artifacts serving a catalytic yet supporting function towards building a stable and generalizable deepfake detection system, especially in adversarial or codec-variant real-world scenarios.

IV.8 CLASSIFICATION PERFORMANCE ANALYSIS:

The efficacy of the designed deepfake audio detection system was critically tested using stratified 5-fold cross-validation on the entire fused multi-domain feature space. On all folds, the model showed uniformly high accuracy, affirming its robust capability to generalize across a wide range of speaker profiles, recording conditions, and spoofing methods included in the large-scale dataset. Table 5 collates the classification metrics for all folds and shows impressively consistent and balanced results. The accuracy varied between 97.79% and 98.28%, with average precision, recall, and F1-scores closely bunched up around 98% for both "real" and "fake" classes. Such performance indicates the ability of the model to classify authentic robustly and manipulated audio samples and suppress false positives and false negatives—the very feature needed for real-world security and forensic applications.

Table 5: Evaluation Report for each Fold.

Fold	Accuracy	Precision (Real)	Recall (Real)	F1-Score (Real)	Precision (Fake)	Recall (Fake)	F1-Score (Fake)
1	98.07%	0.9856	0.9756	0.9806	0.9758	0.9858	0.9808
2	98.28%	0.9835	0.9822	0.9828	0.9822	0.9835	0.9829
3	98.12%	0.9788	0.9837	0.9812	0.9836	0.9786	0.9811
4	97.79%	0.9798	0.9759	0.9778	0.9760	0.9799	0.9779
5	98.07%	0.9874	0.9738	0.9805	0.9742	0.9875	0.9808

Source: Authors, (2026).

The robustness of these measures between folds and the well-balanced support sizes of roughly 3,935 samples per class per fold further emphasizes the model's consistent predictive performance irrespective of data splits. Receiver Operating Characteristic (ROC) curves generated for each fold also illustrate strong separability between authentic and manipulated audio classes. Mean Area Under the Curve (AUC) asymptotically reached values of approximately 0.99, indicating near-perfect discriminative ability. Counterpart confusion matrices indicated that misclassification was largely relegated to noisy or slight samples where acoustic cues may be partially masked or synthetic speech perfectly mimicked natural voicings. Furthermore, the forecasted probability distributions for both classes showed clear margins, with the majority of samples categorized firmly—reflecting well-calibrated outputs that can guide trust decisions in high-stakes deployment situations. These strong findings confirm the effectiveness of combining heterogeneous feature domains—specifically compression artifacts, prosodic dynamics, and phase information—in the development of a model able to cope with the sophisticated, multilateral nature of real-world deepfake audio detection. Together, the combined validation performance constitutes a robust empirical basis that not only demonstrates leading-edge detection precision but also underpins the downstream interpretability analyses and subsequent refinements specific to adaptive and dynamic deepfake generation processes.

IV.9 SHAP-BASED FEATURE CONTRIBUTION ANALYSIS:

To further examine the internal decision behavior of the proposed multi-domain deepfake audio detection framework, a SHAP (SHapley Additive exPlanations) analysis was conducted on the trained classifier [6], [41]. The objective of this analysis was not to reinterpret the model as inherently explainable, but rather to validate whether the learned feature importance patterns align with the empirical observations obtained from the ablation studies presented earlier in this section. SHAP values were computed on the fused feature space using a representative subset of the validation data. Feature contributions were aggregated at the domain level by grouping individual feature dimensions into their corresponding categories, namely acoustic, compression-related, prosodic, phase-based, emotional, and statistical-spectral features.

This grouping allows for a coherent interpretation of how different feature families contribute to the model's predictions, without overemphasizing individual low-level dimensions. Figure 7 presents the resulting SHAP summary plot, illustrating the relative importance of each feature domain in influencing the classification outcome. A clear dominance of compression-related features is observed, indicating that variations introduced through codec-induced distortions play a significant role in distinguishing genuine audio from synthetic speech. This observation is consistent with the ablation results reported in Section 4.7, where the exclusion of compression features led to a noticeable degradation in performance, while their integration with acoustic features yielded substantial gains. Acoustic and prosodic features also exhibit strong contributions, reflecting the model's reliance on spectral structure and temporal speech dynamics. In contrast, emotional and statistical-spectral features demonstrate comparatively moderate but non-negligible influence, suggesting that they act as complementary cues rather than primary discriminators. Phase-based features contribute more subtly, reinforcing the notion that fine-grained temporal coherence differences, while difficult to isolate independently, enhance robustness when fused with other domains. Overall, the SHAP analysis corroborates the central design hypothesis of this work: no single feature domain is sufficient in isolation, but the joint exploitation of complementary cues—particularly compression artifacts in conjunction with acoustic information—enables stable and high-performing deepfake audio detection. Importantly, this analysis serves as a consistency check between the learned model behavior and the experimental findings, strengthening the methodological justification of the proposed multi-domain fusion framework rather than introducing new assumptions. The plot highlights the dominant influence of compression-related features, followed by acoustic and prosodic domains, supporting the conclusions drawn from the ablation study.

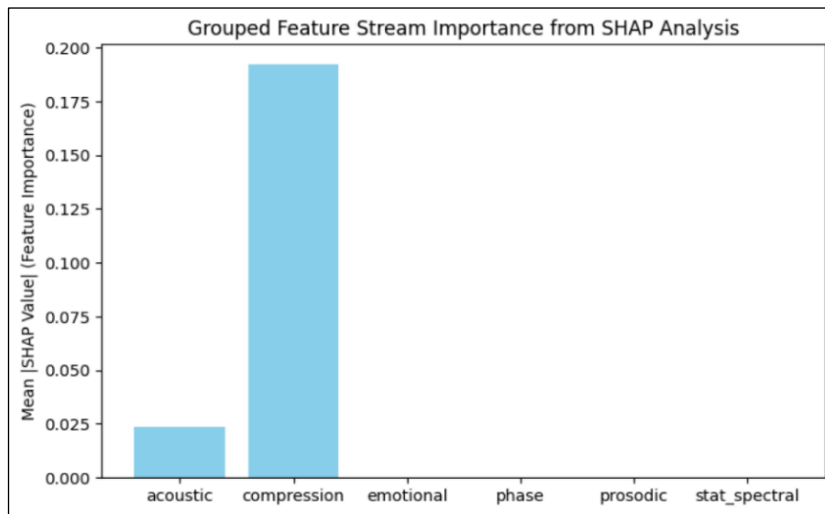


Figure 7: SHAP summary plot illustrating grouped feature-domain contributions to the deepfake audio classification model. Source: Authors, (2026).

V. DISCUSSION:

The findings gathered throughout our comprehensive experimental analysis provide rich insights into the strengths and weaknesses of the suggested deepfake audio detection model. This section explains these results in the framework of feature significance, ablation experiments, comparisons with existing literature, and concerns regarding strong generalization. Collected together, these analyses help disentangle the reasons behind the success of the model, reveal areas where existing methods might weaken, and illustrate opportunities for future enhancement. All experiments are performed on Google Colab Pro utilizing an NVIDIA Tesla T4 GPU with mixed-precision acceleration. The implementation is carried out entirely in PyTorch, and all model checkpoints, training losses, and validation metrics are logged during training. The computational footprint of the model remains modest due to its shallow architecture, allowing real-time inference on consumer-grade hardware and facilitating scalability for large-scale deployments.

V.1 COMPARATIVE ANALYSIS:

To put our suggested multi-domain fusion framework's performance and contribution into context, we contrast it with some of the state-of-the-art deepfake audio detection techniques in recent research. The comparison emphasizes the way our method contributes to advancement in detection accuracy, robustness, explainability, and the ability to work with real-world, in-the-wild datasets. Table 6 gives an overview of exemplar previous work compared to ours, highlighting feature sets, model structures, training data, and performance metrics claimed. Most previous approaches mostly utilize spectral or prosodic features in CNN, RNN, or combined architectures [42], [43], [44]. Such models have decent in-domain performance on ASVspoof or WaveFake datasets but tend to lack generalizability to additional real-world situations. Our fusion system, through the combination of acoustic, compression, prosodic, and phase information, achieves a mean accuracy of around 98% with an AUC of close to 0.998 on a very large-scale, "in-the-wild" collection, outperforming standard benchmarks (e.g., EERs ranging from 1.8–4%) published on benchmark corpora. Of particular interest, the addition of compression artifacts—a relatively neglected feature space—greatly increases detection resilience, as indicated by our ablation experiments. This differentiates our method from earlier research that either omits compression features or merges information only at later model phases. By leveraging dataset diversity, multi-domain feature fusion, and interpretability all at once, our research advances the state of the art in deepfake audio detection oversss comparable systems that target more limited or less explainable approaches.

Table 6: Comparative Summary of Deepfake Audio Detection Methods.

Paper / Method	Dataset(s)	Accuracy / EER / AUC	Notable Contributions
Proposed Work	Large “In-the-Wild” dataset	Accuracy ~98%, AUC 0.998	Multi-domain early fusion, interpretability, robustness
[2]	WaveFake, ASVspoof	EER ~1.8%	Token-level spectral modeling
[42]	ASVspoof, WaveFake	EER ~3–4%	Efficiency-accuracy tradeoff
[21]	ASVspoof Logical Access (LA)	Accuracy ~92%	Unsupervised anomaly detection
[15]	Custom voice dataset	Accuracy ~85–88%	Baseline for comparison
[17]	ASVspoof LA, DFDC-Audio	EER < 2%	Multi-scale dynamic feature fusion

Source: Authors, (2026).

VI. CONCLUSION AND FUTURE SCOPE

This work presents a robust and systematic framework for deepfake audio detection based on multi-domain feature fusion and lightweight neural classification. The proposed approach is evaluated on a large-scale, in-the-wild dataset comprising 31,780 audio samples evenly distributed between genuine and synthetic speech. By incorporating diverse speakers, languages, recording conditions, and synthesis techniques, the dataset closely reflects real-world deployment scenarios and provides a strong empirical foundation for evaluating generalization and robustness. A carefully designed preprocessing pipeline ensures standardized input quality, enabling consistent feature extraction and stable model training. The core methodological contribution of this study lies in the integration of complementary feature domains that capture both physical and algorithmic characteristics of speech production. Acoustic features model vocal-tract behavior, prosodic descriptors reflect higher-level speech dynamics, phase-based features capture fine-grained temporal coherence, emotional features characterize expressive variability, compression-related features expose codec-sensitive artifacts, and statistical spectral measures summarize global signal distributions. Rather than relying on a single representation, the fusion of these domains enables the detector to identify subtle inconsistencies that persist even when modern synthesis systems successfully mimic surface-level spectral patterns. The fused feature representation is classified using a lightweight fully connected neural network, which achieves strong discriminative performance while remaining computationally efficient and stable across validation folds. A key outcome of the experimental evaluation is the consistently high classification performance observed under stratified five-fold cross-validation, with minimal variation across folds. The results demonstrate that the proposed framework effectively separates genuine and synthetic audio across diverse conditions, supporting its suitability for real-world forensic and security applications. Importantly, the ablation study highlights that while individual feature groups, such as compression artifacts, are insufficient in isolation, their inclusion within a multi-domain fusion strategy significantly enhances robustness. This finding explains why the proposed system outperforms approaches that rely solely on spectral or prosodic cues and supports the central design choice of early, heterogeneous feature integration. When compared with existing deepfake audio detection methods reported in the literature, the proposed framework offers several practical advantages.

Many prior works achieve strong in-domain performance but rely heavily on deep spectro-temporal models trained on constrained benchmark datasets, often suffering performance degradation when exposed to unseen synthesis techniques or real-world distortions. In contrast, the present approach leverages domain-diverse features that are less tightly coupled to specific vocoders or architectures, allowing it to generalize more effectively across synthesis pipelines and recording environments. The inclusion of compression-related and phase-based cues, which are often underexplored in earlier studies, provides additional discriminatory evidence that remains difficult for generative models to reproduce consistently. These factors collectively explain the strong empirical results observed in this study. Despite the encouraging performance, several directions remain open for future investigation. While the current dataset is large and diverse, further evaluation under extreme acoustic conditions, adversarial perturbations, and emerging synthesis paradigms would strengthen confidence in long-term robustness. Extending the framework to incorporate linguistic or semantic representations may further enhance detection capability, particularly for emotionally neutral or low-variance speech. Additionally, adaptive or continual learning strategies could be explored to allow the model to update efficiently as new spoofing techniques emerge, without requiring full retraining. Finally, integrating the proposed audio-based system with complementary modalities, such as visual or behavioral cues, represents a promising avenue for constructing more comprehensive multimodal deepfake detection pipelines. In summary, this work establishes a strong and extensible foundation for deepfake audio detection by demonstrating that carefully designed multi-domain feature fusion, combined with an efficient classification model, can achieve reliable and generalizable performance under realistic conditions. The insights gained from feature ablation and comparative analysis provide clear justification for the proposed methodology and offer valuable guidance for future research aimed at countering the rapidly evolving threat of synthetic speech.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Methodology: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Investigation: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Discussion of results: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Writing – Original Draft: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Writing – Review and Editing: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Resources: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Supervision: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

Approval of the final text: Akshat Chhatriwala, Ishita Akolkar, Namrata Shroff, Shankar Parmar.

VIII. REFERENCES

- [1] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," *Expert Syst. Appl.*, vol. 250, p. 123941, 2024, doi: <https://doi.org/10.1016/j.eswa.2024.123941>.
- [2] S. Modak, A. K. Das, and R. Naskar, "SpecViT: A Custom Vision-Transformer based Approach for Audio Deepfake Detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889022.
- [3] A. Vaswani et al., "Attention Is All You Need," Aug. 2023, arXiv. doi: 10.48550/arXiv.1706.03762.
- [4] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digit. Signal Process.*, vol. 158, p. 104956, 2025, doi: <https://doi.org/10.1016/j.dsp.2024.104956>.
- [5] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake Image Detection Using Vision Transformer Models," in *2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, Tbilisi, Georgia: IEEE, Jun. 2024, pp. 332–335. doi: 10.1109/blackseacom61746.2024.10646310.
- [6] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, no. 7, 2025, doi: 10.3390/s25071989.
- [7] M. Astrid, E. Ghorbel, and D. Aouada, "Audio-Visual Deepfake Detection With Local Temporal Inconsistencies," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889087.
- [8] Y. Ahmadiadli, X.-P. Zhang, and N. Khan, "Beyond Identity: A Generalizable Approach for Deepfake Audio Detection," 2025. [Online]. Available: <https://arxiv.org/abs/2505.06766>
- [9] P. Muruganandham, G. R. Thangasamy, S. Jayaraman, and R. Dharmarajan, "LSTM autoencoder based parallel architecture for deepfake audio detection with dynamic residual encoding and feature fusion," *Sci. Rep.*, vol. 15, no. 1, p. 23514, 2025, doi: 10.1038/s41598-025-08198-6.
- [10] M. Astrid, E. Ghorbel, and D. Aouada, "Targeted Augmented Data for Audio Deepfake Detection," Jul. 2024, arXiv. doi: 10.48550/arXiv.2407.07598.
- [11] T. D. N. Le, K. K. Teh, and H. D. Tran, "Continuous Learning of Transformer-based Audio Deepfake Detection," 2024. [Online]. Available: <https://arxiv.org/abs/2409.05924>.
- [12] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey.," *Front. big data*, vol. 5, p. 1001063, 2022, doi: 10.3389/fdata.2022.1001063.
- [13] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?," *Interspeech*, 2022.
- [14] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of Audio Deepfake Detection," in *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 132–137. doi: 10.21437/Odyssey.2020-19.
- [15] A. Hamza et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning," *IEEE Access*, vol. 10, pp. 134018–134028, 2022, doi: 10.1109/ACCESS.2022.3231480.
- [16] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, pp. 144497–144529, 2023, doi: 10.1109/ACCESS.2023.3344653.
- [17] Y. Hao et al., "Integrating Spectro-Temporal Cross Aggregation and Multi-Scale Dynamic Learning for Audio Deepfake Detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889337.
- [18] G. Sivaraman, H. Tak, and E. Khoury, "Investigating voiced and unvoiced regions of speech for audio deepfake detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890861.
- [19] G. Lee et al., "Dual-Channel Deepfake Audio Detection: Leveraging Direct and Reverberant Waveforms," *IEEE Access*, vol. 13, pp. 18040–18052, 2025, doi: 10.1109/ACCESS.2025.3532775.
- [20] Z. Jin, L. Lang, and B. Leng, "Wave-Spectrogram Cross-Modal Aggregation for Audio Deepfake Detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890563.
- [21] D. Song, N. Lee, J. Kim, and E. Choi, "Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model," *IEEE Access*, vol. 12, pp. 184311–184326, 2024, doi: 10.1109/ACCESS.2024.3506973.
- [22] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "Generalizable Audio Deepfake Detection via Latent Space Refinement and Augmentation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10888328.
- [23] X. Chen et al., "Continual Unsupervised Domain Adaptation for Audio Deepfake Detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890538.
- [24] S. Zeng et al., "Adversarial Training and Gradient Optimization for Partially Deepfake Audio Localization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890470.
- [25] M. Gohari, D. Salvi, P. Bestagini, and N. Adami, "Audio Features Investigation for Singing Voice Deepfake Detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10888452.
- [26] Z. Xie, B. Li, X. Xu, Z. Liang, K. Yu, and M. Wu, "FakeSound: Deepfake General Audio Detection," 2024. [Online]. Available: <https://arxiv.org/abs/2406.08052>
- [27] N. Ur Rehman Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud, and T. Alsaifi, "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," *IEEE Access*, vol. 13, pp. 1923–1961, 2025, doi: 10.1109/ACCESS.2024.3523288.

- [28] Z. Wang, D. Ye, J. Li, and J. Deng, "Generalize Audio Deepfake Algorithm Recognition via Attribution Enhancement," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889399.
- [29] C. Y. Kwok, D.-T. Truong, and J. Q. Yip, "Robust Audio Deepfake Detection using Ensemble Confidence Calibration," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889972.
- [30] G. S. Kashyap et al., "Fooling the Forgers: A Multi-Stage Framework for Audio Deepfake Detection," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10888175.
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," CoRR, vol. abs/1810.0, 2018, [Online]. Available: <http://arxiv.org/abs/1810.04805>.
- [32] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," CoRR, vol. abs/2006.1, 2020, [Online]. Available: <https://arxiv.org/abs/2006.11477>
- [33] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN Embeddings for Speaker Diarization," in Interspeech 2021, in interspeech_2021. ISCA, Aug. 2021. doi: 10.21437/interspeech.2021-941.
- [34] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," arXiv Prepr. arXiv2308.14970, 2023.
- [35] P. M. Chauhan and N. P. Desai, "Mel Frequency Cepstral Coefficients (MFCC) based speaker identification in noisy environment using wiener filter," in 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE), Mar. 2014, pp. 1–5. doi: 10.1109/ICGCCEE.2014.6921394.
- [36] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?," 2024. [Online]. Available: <https://arxiv.org/abs/2203.16263>
- [37] G. Tahaoglu, D. Baracchi, D. Shullani, M. Iuliani, and A. Piva, "Deepfake audio detection with spectral features and ResNeXt-based architecture," Knowledge-Based Syst., vol. 323, p. 113726, 2025, doi: <https://doi.org/10.1016/j.knosys.2025.113726>.
- [38] S. Nailwal, S. Singhal, N. T. Singh, and A. Raza, "Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis," in 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India: IEEE, Nov. 2023, pp. 1–8. doi: 10.1109/rmkmate59243.2023.10369155.
- [39] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.01777>
- [40] T. M. Wani, S. A. A. Qadri, A. Ashraf, and I. Amerini, "ASTDT: an Interpretable Adaptive Spectro-Temporal Diffusion Transformer for audio deepfake detection," EURASIP J. Inf. Secur., vol. 2025, no. 1, p. 32, 2025, doi: 10.1186/s13635-025-00217-3.
- [41] N. Yu, L. Chen, T. Leng, Z. Chen, and X. Yi, "An explainable deepfake of speech detection method with spectrograms and waveforms," J. Inf. Secur. Appl., vol. 81, p. 103720, 2024, doi: <https://doi.org/10.1016/j.jisa.2024.103720>.
- [42] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," IEEE Access, vol. 11, pp. 132652–132682, 2023, doi: 10.1109/ACCESS.2023.3333866.
- [43] L. Y. Gong and X. J. Li, "Deepfake Voice Detection: An Approach Using End-to-End Transformer with Acoustic Feature Fusion by Cross-Attention," Electronics, vol. 14, no. 10, 2025, doi: 10.3390/electronics14102040.
- [44] A. A. M. Albazony, H. A. Al-Wzwazy, A. S. Al-Khaleefa, M. A. Alazzawi, M. Almohamadi, and S. E. Alavi, "DeepFake Videos Detection by Using Recurrent Neural Network (RNN)," in 2023 Al-Sadiq International Conference on Communication and Information Technology (AICCIT), Al-Muthana, Iraq: IEEE, Jul. 2023, pp. 103–107. doi: 10.1109/aiccit57614.2023.10217956.