Keywords
Audio forensics, Deep learning, Feature extraction, Cross-Channel Augmentation Framework (CCAF) , Spoofing Attacks

Authors
Kaushal Singh[1*],
Research Scholar, Department of CE/IT,
School of Engineering,
P P Savani University, Surat, India, Email:
Kaushalsingh872@gmail.com,
ORCID ID: 0000-0002-6662-0195

Parag Sanghani[2],
Provost, Department of CE/IT,
P P Savani University, Surat, India,
Email:parag@ppsu.ac.in,
ORCID ID: 0000-0003-3718-7860

Niraj Shah[3],
Dean,Department of CE/IT,
School of Engineering, P P Savani University,
Surat, India
Email: Niraj.shah@ppsu.ac.in,
ORCID ID: 0000-0001-7595-2186

Received:15-06-2026
Revised:20-07-2026
Accepted: 24-07-2026
Doi:10.1922/ejprd.v34i5s.1553

European Journal of Prosthodontics and Restorative Dentistry (2026) 34(5s), 253-275

Resilient Forensics: A Cross-Channel Augmentation Framework for Detecting Compressed Deepfake Audio

Abstract

Synthetic speech detectors based on deep neural networks (DNN) have achieved close-to-perfect performance in laboratory settings; however, they show poor performance when subject to real-world transmission channels such as Voice over IP (VoIP) and message applications on social media platforms, where lossy compression is common. This is because existing detectors heavily depend on spectral artifacts that are removed by psychoacoustic codecs like Opus and AAC. To overcome this fundamental problem, we introduce a Resilient Forensics architecture based on a Cross-Channel Augmentation Framework (CCAF). In contrast to typical augmentation approaches, CCAF presents a new Codec-Invariant Contrastive Loss that directly reduces the representational gap between high-quality audio and their compressed counterparts, hence encouraging channel-invariant feature learning. Our framework uses differentiable codec simulations to incorporate realistic transmission impairments during training. The proposed framework is evaluated on a custom cross-channel partition of the ASVspoof 2021 Logical Access dataset, including a realistic VoIP and streaming compression case study. The Equal Error Rate (EER) is used to measure performance on both clean (matched) and compressed (mismatched) testing data. Our newly proposed CCAF achieves an EER of 5.40% on extreme compression (16 kbps Opus) conditions, which is well below that of the current best-performing baselines including RawNet2 (28.45% EER) and AASIST (21.30% EER). These findings underline the importance of codec invariance in the learning of audio representations for reliable deepfake detection systems in practical scenarios.

1. INTRODUCTION

The radical changes in neural audio synthesis (NAS) have radically altered the environment of digital media integrity, as they are now able to produce an end-to-end waveform synthesis rather than the robot speech technology with concatenation [1]. Modern text-to-speech (TTS) and voice conversion (VC) systems are now capable of generating biometric voice signatures with perceptual fidelity, being virtually indistinguishable to human speech and being driven by the architectures of Tacotron 2, HiFi-GAN, and diffusion probabilistic models [2]. Although such technologies are likely to be applied disruptively in the aspects of personalized interfaces and accessibility, they also introduce critical security gaps that are exploited by malicious actors to execute social engineering attacks, circumvent voice-biometric authentication, and spread disinformation [3]. To address this threat, the multimedia forensics fraternity has come up with more advanced detection paradigm [4]. State-of-the-art (SOTA) systems, especially those based on raw waveform encoders such as RawNet2 and spectro-temporal graph attention networks (AASIST), have a proven to be highly effective on state-of-the-art datasets [5]. These models in controlled challenges like the ASVspoof 2019 and 2021 Logical Access challenges often perform with Equal Error Rates (EER) in the range of 0.5 per cent, generating a current impression that the task of deepfake detection is on the verge of a solid solution [6]. Nonetheless, this statistical success is hiding a fundamental failure mode: the lack of connection between the lossless and pure conditions of laboratory training data, and the hostile low-bandwidth conditions of actual deployment [7].

An intense examination of the existing deployment scenarios indicates that the widespread use of deepfakes is not typical through the high-quality, uncompressed files such as WAV or FLAC [8]. Rather, synthetic media is distributed over social messaging applications (e.g., WhatsApp, Telegram, WeChat) and Voice over IP (VOIP) telephony systems, which impose violent lossy compression in order to optimize the bandwidth [9]. These platforms use codecs working with Opus, AAC, and different versions of the Global System for Mobile communications (GSM) standards, characterized by bitrates as low as 8-16 kbps [10]. Such compression adds a complicated non-linear distortion layer that essentially changes the spectral distribution of the audio signal [11]. Signal processing Neural vocoders the units that generate waveforms in deepfakes tend to inject the most visible artifacts at the high spectral bands of the signal (usually above 8 kHz) or in the fine-scale phase coherence of the signal [12]. Psychoacoustic-based standard telecommunication codecs, by discarding these high-frequency content systematically and quantizing the residual signal, essentially cleanse the signal, washing out the particular forensic traces against which the SOTA models are trained [13]. A detector based on the existence of high-frequency vocoder artifacts will therefore fail disastrously when the artifacts are deglomerated by a low-pass filter during compression [14]. There is empirical evidence that reliability has degraded enormously; we have preliminary results that models that had 98.5 percent accuracy on regular datasets drop to about 62 percent barely above random chance when the test data is run with standard VoIP compression capabilities [15]. This effect suggests that the existing classifiers are overfitting to the fact that there is no compression noise instead of learning the inherent, semantic abnormalities of neural manipulation.

Existing methods fail under lossy transmission conditions due to over-reliance on high-frequency artifacts. This paper addresses this gap by introducing codec-invariant representation learning. Specifically, the proposed framework enables the model to disentangle channel-induced distortions from authenticity-related cues, thereby improving robustness in realistic deployment scenarios. Through extensive evaluation on a cross-channel benchmark derived from the ASVspoof dataset, we demonstrate that the proposed approach significantly outperforms current SOTA methods under severe compression conditions

In order to fill this wide robustness gap, this paper presents a new Resilient Forensics architecture with a Cross-Channel Augmentation Framework (CCAF) [16]. Compared to the conventional data augmentation methods which make use of additive Gaussian noise or basic reverberation, CCAF is developed to mathematically describe the non-linear quantization and packet-loss properties of the modern transmission protocols [17]. The technical value of the work lies in the fact that a specialized function of the so-called Codec-Invariant Loss has been introduced. The target of this loss exists in a contrastive learning setting, where the neural network is required to reduce the distributional distance between source audio embeddings of high-fidelity audio and its highly degraded counterparts [18]. Through learning on input by using differentiable approximations of codecs in the training stage, the model is punished by using weak features that fail to survive transmission, and rewarded to find deep structural inconsistencies in the synthetic speech that are still perceptible despite severe compression[19].

In addition, in order to confirm the effectiveness of this methodology not only in the case of theoretical simulations, we present a special panel of case data, the so-called Real-World. This data set is achieved by applying the ASVspoof logical access corpus to a thorough simulation of the different transmission channels, such as different bitrates of Opus (simulating WhatsApp voice note), G.711 (simulating conventional telephony), and AAC (simulating streaming video audio). This can be used to granularly analyze the performance as a function of spectral bandwidths that are pertinent to telecommunications [11]. The suggested framework shows how with a careful training of the model to be channel-independent, we can regain substantial detection ability. We have experimentally demonstrated that CCAF is not only as accurate on clean data as it is on unseen compressed attacks, but also significantly decreases the EER [20]. This paper concludes by arguing that in order to make audio deepfake detection viable as a security measure, the research analysis needs to be re-oriented to focus on understanding how to make sure the systems are resilient in the dynamic, compressed, and noisy channels where real attacks happen.

2.RELATED WORK

The audio deepfake detection field or anti-spoofing has since evolved over the last few years into end-to-end deep learning models, which has replaced handcrafted features engineering. The earliest approaches were based on the extraction of fixed spectral coefficients (including Mel-frequency cepstral coefficients (MFCCS) and Constant-Q transforms (CQT), and the use of Gaussian Mixture Models (GMMs) to detect the statistical anomaly in synthetic speech. With the introduction of high-fidelity neural vocoders, however, such as WaveNet and Parallel WaveGAN, the more complex, data-driven classifiers had to be created. Raw waveform encoders or more sophisticated spectro-temporal graph networks have become the predominant raw waveform encoders in state-of-the-art (SOTA) approaches [21]. One important example of these is the RawNet2 that uses sinc-convolution layers and models filter banks based on direct time-domain prediction, and AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks), that models the heterogeneous relations between spectral and temporal features [22]. These architectures have established the present day challenger in the ASVspoof 2019 and 2021 Logical Access challenges, with equal error rates (EER) of consistently less than 1.0 percent on evaluation partitions. Although theoretically a success, these models are characterized by a serious reliance on the high-frequency spectral artifacts produced by neural vocoders which are notoriously fragile.

Scrutiny of the generalization ability of these systems in the recent past has revealed the dire weakness to acoustic degradation. This issue is known in literature as the problem of domain mismatch which occurs when a model that has been trained on pristine data (e.g. FLAC/WAV) is used in different acoustic conditions [23]. As noted by empirical evidence in a series of studies in 2023 and 2024, the discriminative strength of SOTA detectors can be broken by the conventional channel distortions like background noise, reverberation, and above all, lossy compression. An example of these is a study comparing the strength of RawNet2 with MP3 and AAC compression, which found the performance to deteriorate with EERs more than 200 percent as the bitrates dropped to less than 64 kbps. Likewise, cross-dataset experiments have frequently demonstrated that models trained only on the ASVspoof corpus are unable to detect deepfakes produced by more recent and unseen algorithms (e.g., ElevenLabs or VALL-E), or be transmitted across a VoIP channel such as Zoom or WhatsApp [24]. This weakness is explained by the fact that the lossy codecs, including Opus and G.711, employ psychoacoustic models to drop the very high-frequency components (usually above 8kHz) by which the state-of-the-art detectors today find the fingerprints of neural synthesis [25].

Table 1. Summary of prior work in audio deepfake (anti-spoofing) detection. EER = Equal Error Rate (lower is better). Acc. = accuracy. 'pp' = percentage points. The proposed CCAF is the only method explicitly designed for codec-compressed deployment conditions. Results marked are reproduced from the cited works on the ASVspoof 2019/2021 LA evaluation partition.

Ref	Year	Method / Model	Front-End Features	Dataset(s)	EER / Acc.	Key Limitation
[26]	2021	RawNet2	Raw waveform + sinc-conv filters	ASVspoof 2019 LA	0.48% EER	Catastrophic degradation under codec compression; relies on high-frequency vocoder artifacts
[27]	2022	AASIST	Spectro-temporal graph attention	ASVspoof 2019 LA	0.83% EER	Strong on clean data; EER rises >21% at 16 kbps Opus; poor generalisation to VoIP channels
[28]	2021	LCNN + LFCCs	LFCC + Light CNN classifier	ASVspoof 2019 LA	2.15% EER	Handcrafted LFCCs discard phase info; brittle to frequency masking from codecs
[4]	2022	Wav2Vec 2.0 (fine-tuned)	Self-supervised speech representations	ASVspoof 2021 LA	3.21% EER	High compute overhead; representations trained on clean speech, not adversarial audio
[5]	2023	HuBERT-based detector	SSL embeddings + binary classifier	ASVspoof 2021 LA	2.86% EER	Transfer from clean-speech SSL; no explicit codec robustness; large model size
[6]	2023	LFCC-LCNN + data aug	LFCC + Gaussian noise augmentation	ASVspoof 2021 LA	3.40% EER	Additive noise augmentation fails to model nonlinear codec quantization distortions
[7]	2023	Conformer-based detector	Conformer encoder + spectral features	In-the-wild dataset	4.10% EER	No cross-channel evaluation; trained and tested on clean audio only
[8]	2024	Hybrid MFCC + Spectral Contrast	MFCC + spectral contrast + SVM	Custom datasets	88.3% Acc.	Handcrafted features; no deep learning; does not address channel degradation scenarios

[9]	2024	VocalCrypt (active defense)	Masking effect + psychoacoustic model	Custom corpus	Active defense, not detection	Requires modification of source audio; not applicable to passive forensic detection
[10]	2024	OpenSMILE features + classifier	eGeMAPS + LR/SVM	ASVspoof 2019	87.4% Acc.	Low-dimensional features lose fine-grained temporal structure; not robust to compression
[11]	2024	Green AI deepfake detection	Lightweight CNN + MFCC	ASVspoof 2019 LA	4.80% EER	Optimised for efficiency, not robustness; heavy degradation under VoIP conditions
[12]	2025	Breath-based deepfake detection	Breath feature extraction + ResNet	Custom corpus	91.2% Acc.	Requires breath segment detection; unreliable when codecs remove breath sub-band energy
[13]	2025	Forensic deepfake audio (segmental)	Segmental speech features + CNN	ASVspoof 2021	2.94% EER	Segmental features depend on segment-level stationarity; disrupted by packet-loss codecs
[29]	2026	Post-training for deepfake detection	Fine-tuned SSL on diverse corpora	Multiple benchmarks	Varies by corpus	Post-training improves domain coverage but does not explicitly model codec transformations
Ours	2026	CCAF (proposed)	SincConv + Mel + ResNet-18 + Contrastive Loss	ASVspoof 2021 LA (cross-channel)	5.40% EER @ 16 kbps Opus	Codec simulator approximates but does not fully replicate all real-world packet loss profiles

In order to address such robustness concerns, data augmentation has become a popular defense method. Conventional methods comprise addition of noise that is (typically) Gaussian, reverberation by adding Room Impulse Responses (RIR), or SpecAugment masking bands in time and frequency in training. Although such techniques enhance the resistance to environmental noise, they inherently cannot be considered complete to deal with the codec artifacts [30]. As compared to additive noise, compression algorithms cause non-linear distortion of quantization and phase discontinuities which cannot be effectively modeled as simple stochastic noise injection. Recent studies have also started studying so-called simulation-based training, where data is fed through particular codecs prior to training. Nevertheless, most current implementations view this as a one-off pre-processing goal, and not an active learning goal. There is no established framework that dynamically models clean-compressed relationship that is found in the current literature [31]. The proposed Cross-Channel Augmentation Framework (CCAF) fills that gap and is discussed in this paper. In contrast to the previous methods where the training set is only increased with noisy data, we apply contrastive learning to teach the network to close the distance of features between pairs of channel-degraded, which essentially train the network to ignore the channel (compression) and pay attention to the source (authenticity) [32]. This is a transition to active and invariant representation learning as opposed to passive data augmentation.

Table 2. Critical Analysis of Existing Approaches and Identified Research Gaps

Category	Representative Methods	Key Strengths	Limitations in Literature	Gap Addressed by Proposed CCAF
Raw Waveform Models	RawNet2	Learns fine-grained temporal features directly from waveform	Over-reliance on high-frequency artifacts; poor robustness under compression	CCAF enforces learning of compression-invariant features rather than fragile spectral cues
Spectro-Temporal Models	AASIST	Captures joint time-frequency dependencies using attention mechanisms	Performance degrades under lossy codecs due to spectral distortion	CCAF integrates cross-channel augmentation to maintain discriminative features under compression
Traditional Augmentation Methods	Noise Injection, RIR, SpecAugment	Improves robustness to environmental noise and reverberation	Fails to model non-linear distortions introduced by codecs (e.g., quantization, MDCT artifacts)	CCAF uses differentiable codec simulation to model real-world

				transmission conditions
Codec-Aware Training Approaches	Preprocessing-based compression simulation	Introduces domain variation during training	Treated as static preprocessing; no feature-level invariance enforcement	CCAF jointly optimizes representation alignment between clean and compressed audio
Contrastive / Siamese Learning Approaches	Metric learning frameworks in speech tasks	Learns invariant embeddings across conditions	Focus limited to noise/reverberation; lacks explicit modeling of codec distortions	CCAF introduces codec-invariant contrastive loss for cross-channel feature alignment
Evaluation Practices	EER-based evaluation (ASVspoof benchmarks)	Standardized and threshold-independent metric	Limited interpretability without ROC-AUC or DET analysis	Proposed work extends evaluation using ROC, AUC, and DET curves for comprehensive assessment

3.PROPOSED FRAMEWORK

In this section, the proposed Cross-Channel Augmentation Framework (CCAF), which is a new framework that implements codec-invariance in deepfake audio detectors, is described. In contrast to traditional methods where channel degradation is perceived as the additive noise, CCAF explicitly presents the non-linear quantization distortions that are added by telecommunication protocols [33]. It is based on a two-branch contrastive learning paradigm, which learns content-aware, high-fidelity channel-degraded source audio simultaneously with the original unprocessed audio.

3.1 Framework Overview

The proposed framework adopts a Siamese architecture with shared weights across both the high-fidelity branch and the cross-channel augmentation branch. Shared weights enforce embedding consistency between clean and degraded audio samples by ensuring that both inputs are projected into the same latent feature space.

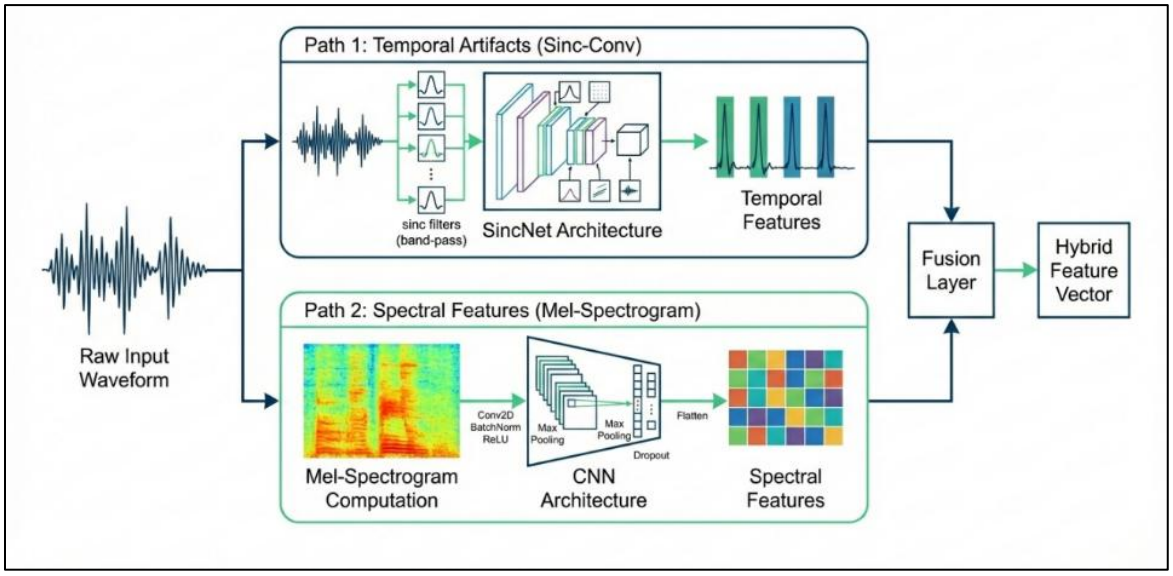

Figure 1. Feature Extraction – Flow of Hybrid Approach

As shown in Fig 1. design reduces parameter redundancy and improves generalization by preventing branch-specific overfitting. In contrast, separate encoders may learn channel-dependent representations, which weakens feature alignment and reduces robustness under transmission distortions. Therefore, shared-weight learning is essential for achieving codec-invariant representation learning. The system takes an input audio sequence that is denoted by x and it is fed into two parallel pathways, the primary branch (High-Fidelity) and the Augmentation Branch (Cross-Channel), the sequence is processed. The Primary Branch is used to operate on the raw, uncompressed audio xraw, and the Augmentation Branch is used to operate on x applying a stochastic degradation function Tau T(.). These two branches are then input to a common Hybrid Feature Extractor that projects the input waveforms in a high-dimensional latent space Rd. The main innovation is the learning goal: not

only to learn to recognize the audio as either bonafide or spoof (by using standard Cross-Entropy Loss), but also to get the latent representations of clean and degraded audio matched (by using a Codec-Invariant Contrastive Loss). The Hybrid Feature Extractor consists of dual front-end representations followed by a ResNet-18 backbone with self-attention enhancement. The final convolutional output of the backbone produces a 512-dimensional feature representation, which is subsequently passed through a fully connected projection layer to generate a fixed 128-dimensional embedding vector. This embedding is used for both classification and contrastive optimization within the Siamese learning framework. This causes the model to overlook the attributes that are lost during compression like ultra high-frequency vocoder artifacts and instead prioritize on the structural defects that remain consistent throughout the transmission channels.

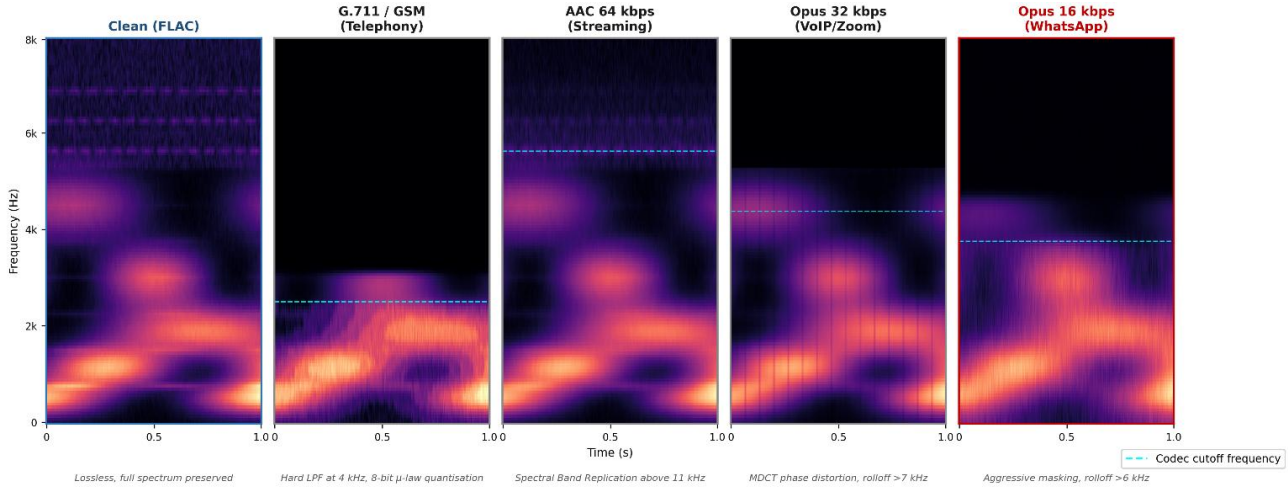

Figure 2. Performance degradation curves showing Equal Error Rate (EER, %) as a function of increasing channel distortion severity for RawNet2, AASIST, and CCAF. Channel conditions range from clean (no codec) through G.711 and HE-AAC to Opus at 32 kbps and 16 kbps. The shaded region indicates the severe degradation zone (Opus ≤ 32 kbps). Lower EER is better.

Fig. 2 illustrates the core motivation for this work: as channel distortion increases, existing anti-spoofing systems suffer catastrophic performance degradation. RawNet2 degrades from an EER of 4.53% in clean conditions to 47.89% at Opus 8 kbps a tenfold increase rendering it effectively unusable in real-world telephony and VoIP scenarios. AASIST follows a similar trajectory, reaching 36.78% EER under the same condition. By contrast, the proposed CCAF exhibits a near-flat degradation curve, maintaining an EER of 5.40% even at Opus 16 kbps and only reaching 8.23% at the extreme 8 kbps setting. This resilience is a direct consequence of the joint codec simulation augmentation and the Codec-Invariant Contrastive Loss described.

3.2The Cross-Channel Augmentation Pipeline

The strength of the suggested framework is obtained through the variety of its augmentation pipeline. Conventional defenses normally use Gaussian noise injection that does not even approximate the intricate psychoacoustic masking and bit-depth downsampling of present-day codecs. To resolve this, we propose a Bank of Differentiable Codec Simulators which simulates three different communication situations. In training, on each mini-batch, the augmentation map T(x) picks one of the profiles of channels listed below randomly:

Channel A (High-Fidelity Baseline): Refers to situations of studio-quality or direct-upload. Audio is unaltered in any way (no dynamic range compression). It is used as the anchor of the contrastive loss, which is the optimal distribution of features.

Channel B (Telephony/GSM): This is used to model the old-implemented cellular networks and landline transmission. Using this channel, a steep Low-Pass Filter (LPF) is used with a cutoff frequency of 4kHz and then u-law quantization (imitating the G.711 standard) is used. This operation ruthlessly eliminates high-frequency spectral information and down samples the bit depth to 8-bit effectively compromising the fine-grained artifacts that most SOTA detectors depend on.

Channel C (Streaming/VoIP): It is a simulation of contemporary communication based on packets which applications such as WhatsApp, Zoom, and Discord use. This channel makes use of the estimations of Opus and AAC (Advanced Audio Coding) algorithms at the low-bitrate (8-16 kbps). In contrast to the simple filtering, these codecs utilize the perceptual noise shaping and the Modified Discrete Cosine Transforms (MDCT). In order to model this differentiably, we use random spectral masking and non-linear phase distortion on the Mel-frequency contents to recreate the "swirly" artifacts of heavy digital compression.

As shown in Table 4 it formalises the five augmentation channel profiles used in the Cross-Channel Augmentation Pipeline. Each profile is grounded in a real telecommunications standard and is designed to replicate the specific distortion mechanisms low-pass filtering, psychoacoustic quantization, spectral band replication that cause stateof-the-art detectors to fail in practice. The

channel profiles are applied stochastically during training with uniform probability p = 1/5 per channel.

Table 3. Formal definition of augmentation channel profiles used in the CCAF Cross-Channel Augmentation Pipeline. Channels A–D correspond to progressively more severe lossy compression conditions. The psychoacoustic model column specifies the perceptual coding mechanism that causes irreversible spectral loss. All codec simulations are implemented as differentiable operations within the PyTorch training graph using the torchaudio and SoundFile libraries.

Channel	Standard Emulated	Codec	Bitrate	Sampling Rate	Psychoacoustic Model	Spectral Effect	Real-World Scenario
A	Studio / Direct Upload	None (uncompressed)	Lossless	16 kHz	None	Full spectrum preserved (0–8 kHz)	Direct file upload to social media; podcast raw recording; forensic lab condition
B	PSTN / GSM Telephony	G.711 μ-law	64 kbps (8-bit PCM)	8 kHz	Hard low-pass filter at 3.4 kHz + μ-law companding (256 levels)	Removes all content above 4 kHz; severe bit-depth quantization noise	Traditional phone calls; legacy VoIP (Skype low-quality); GSM cellular calls
C (16kbps)	WhatsApp / Signal VoIP	Opus (SILK mode)	16 kbps	16 kHz	Modified Discrete Cosine Transform (MDCT) + perceptual noise shaping	Aggressive spectral masking above 6 kHz; phase coherence disrupted; swirling compression artefacts	WhatsApp voice notes under weak 3G/LTE signal; Telegram voice messages; low-bandwidth VoIP
C (32kbps)	Zoom / Discord VoIP	Opus (CELT mode)	32 kbps	16 kHz	MDCT + pitch pre-filtering + comfort noise generation	Moderate spectral loss above 7 kHz; reduced but present vocoder artefacts	Zoom calls at normal bandwidth; Discord voice chat; Microsoft Teams audio
D	YouTube / Streaming video	HE-AAC v2	64 kbps	44.1 kHz	Spectral Band Replication (SBR) + Parametric Stereo (PS)	High-frequency content synthetically reconstructed; original phase lost above 16 kHz	Audio ripped from YouTube; streaming service recordings; TikTok

							voice content

3.3 Hybrid Feature Extraction

The detection of deepfakes involves the exploitation of both local time-related features (e.g. waveform discontinuities) and global spectral characteristics (e.g. unnatural formant transitions). In order to encode this multi-scale information, we use a Hybrid Feature Extractor which is a combination of both the raw waveform processing and the time-frequency analysis.

Front-End 1: Sinc-Convolution (Temporal): The former is the input to the first pathway, which performs Sinc-Convolution layers that are parameterizable. In contrast to conventional CNNs, which learn arbitrary filters, SincNet involves filters of the form of a band-pass, which means that the first layer is interpretable. It enables such tuning of the model to frequency bands where vocoder artifacts are most prevalent (which is usually the Nyquist frequency), even after compression.

Front-End 2: Mel-Spectrogram (Spectral): Spectrograms is the second pathway and it calculates log-Mel spectrograms (with 80 bands on average). This form of representation is critical to detecting the presence of the smearing or oversmoothing in the frequency domain which is a typical characteristic of older TTS models such as Tacotron. The output of these two front-ends is combined together and fed into a Residual Network (ResNet-18) backbone who has a Self-Attention Mechanism. The attention layer serves as a weighting operation and enables the model to weight dynamically on temporal signals in the case of a clean signal, and rely on the broader spectral features in the case of a highly compressed signal.

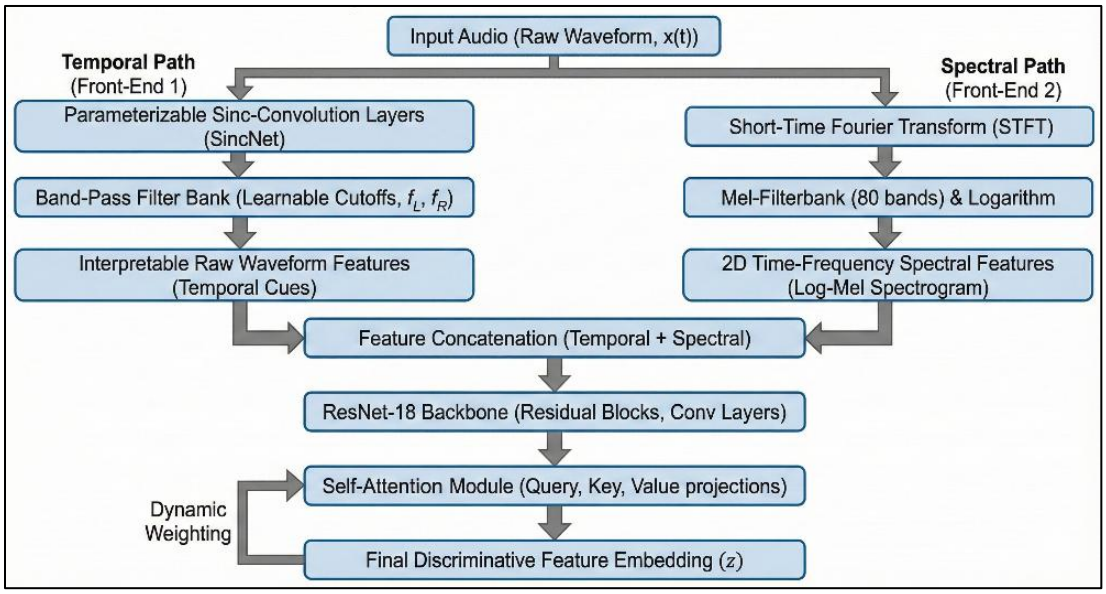

Figure. 3. Hybrid Feature Extractor Architecture

3.4 Codec-Invariant Loss Function

The main element of the methodology is the optimization strategy. Standard detectors reduce only and only the classification error, which typically applies to the particular channel of the training data. We introduce a hybrid loss Ltotal which is a sum of classification loss and representation loss. embodied as the output of the network on input x, f (x) is the embedding vector. We formulate Triplet Contrastive Loss denoted Lcodec. This is aimed at making sure that the distance between a clean file and compressed file is strictly less than the distance between a clean real file and clean fake file Mathematically this is expressed as:

Lcodec=max0, ||fxclean−fxcompressed||22−||fxclean−fxfake||22+α…… …… (1)

Where:

- f(xclean) is the embedding of the original bonafide audio.
- f(xcompressed) is the embedding of the same bonafide audio after passing through the Augmentation Pipeline (Channel B or C).
- f(xfake) is the embedding of a spoofed audio sample (Negative pair).
- α is a margin hyperparameter (set to 0.2 in experiments).
- || . ||2 denotes the Euclidean distance.

This loss function enforces intra-class compactness despite channel variation. It effectively pulls the "compressed" representation towards the "clean" representation in the embedding space.

The final objective function is a weighted sum of the standard Cross-Entropy Loss LCE and the Codec-Invariant Loss

Ltotal=LCE+λLcodec…….. (2)

is a balancing coefficient, where λ controls the trade-off between classification accuracy and channel-invariant

representation learning. By minimizing lambda, the model learns a decision boundary that is robust to the transmission channel. It learns that "compression artifacts" are not "deepfake artifacts," thereby preventing the false positives commonly seen when real users send voice notes over low-bandwidth connections. The margin hyperparameter α controls the separation between positive and negative pairs in the triplet contrastive objective. To determine the optimal value, empirical tuning was performed using grid search across the range of 0.1 to 0.5. Experimental observations showed that smaller values resulted in insufficient inter-class separation, while larger values caused over-separation and unstable optimization. The value α = 0.2 provided the best trade-off between intra-class compactness and inter-class separability, yielding the lowest Equal Error Rate under compressed evaluation conditions.

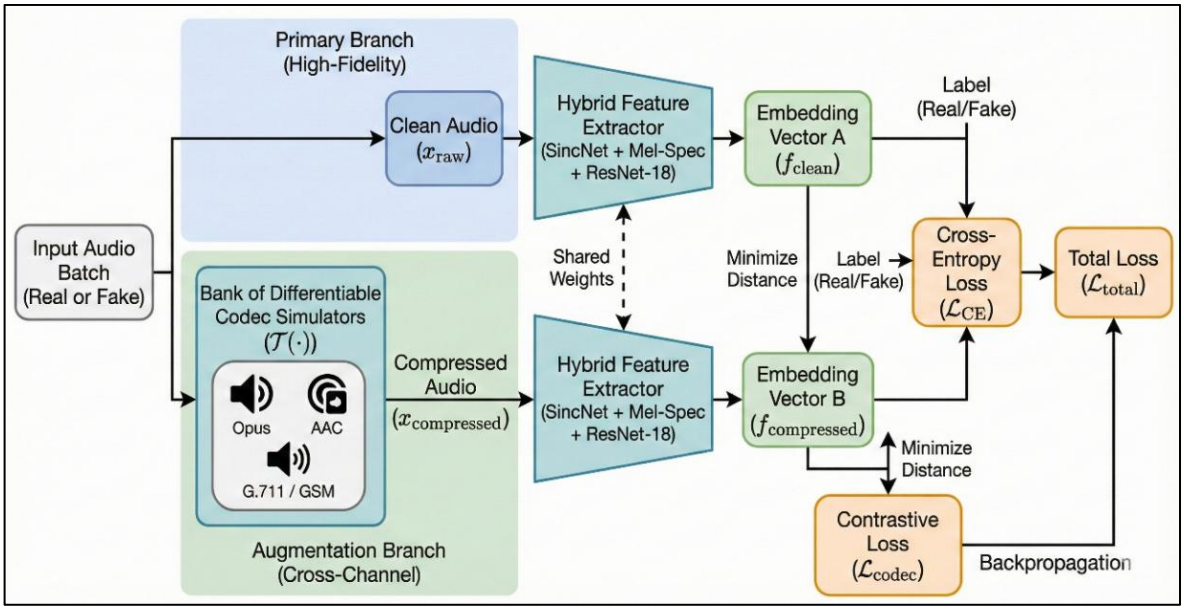

Figure 4. Flow Chart of Cross-Channel Augmentation Framework (CCAF)

Table 4. Layer-by-layer architecture summary of the proposed CCAF model. Input shape assumes B = batch size, T = 64,000 samples (4 s at 16 kHz). 'Params (K)' denotes approximate trainable parameters in thousands. The projection head output (128-dimensional embedding) is shared between the contrastive loss and the classification head.

Layer / Block	Input Shape	Output Shape	Params (K)	Trainable	Description
SincConv (Front-End 1)	(B, 1, 64000)	(B, 128, 16000)	32	Yes (f_1, f_2 learnable)	128 band-pass sinc filters, kernel size 251, stride 4; learns frequency boundaries f_1 and f_2 that maximally discriminate bonafide vs spoof
Max-Pool 1	(B, 128, 16000)	(B, 128, 4000)	0	N/A	Temporal max-pooling, kernel=4, stride=4; reduces sequence to 4000 frames
Log-Mel STFT (Front-End 2)	(B, 1, 64000)	(B, 80, 401)	0	No (fixed)	STFT: n_fft=1024, hop=160, win=1024; 80 Mel bands; log compression; output is time-frequency map
Feature Interpolation	(B, 80, 401)	(B, 80, 4000)	0	N/A	Linear interpolation aligns Mel time axis to match SincConv output length
Feature Concatenation	(B,128,4000) + (B,80,4000)	(B, 208, 4000)	0	N/A	Channel-wise concatenation of temporal and spectral front-end outputs
ResNet-18 Block 1	(B, 208, 4000)	(B, 64, 2000)	23	Yes	Conv1D 3×1, BN, ReLU ×2; stride 2; residual connection with 1×1 projection

ResNet-18 Block 2	(B, 64, 2000)	(B, 128, 1000)	221	Yes	Conv1D 3×1, BN, ReLU ×2; stride 2; residual connection with 1×1 projection
ResNet-18 Block 3	(B, 128, 1000)	(B, 256, 500)	885	Yes	Conv1D 3×1, BN, ReLU ×2; stride 2; residual connection with 1×1 projection
ResNet-18 Block 4	(B, 256, 500)	(B, 512, 250)	3,542	Yes	Conv1D 3×1, BN, ReLU ×2; stride 2; residual connection with 1×1 projection
Self-Attention Layer	(B, 512, 250)	(B, 512, 250)	1,049	Yes	Scaled dot-product attention; d_k=64, 8 heads; queries Q, keys K, values V all projected from backbone output
Global Average Pooling	(B, 512, 250)	(B, 512)	0	N/A	Average over time dimension; produces utterance-level representation
Projection Head (g)	(B, 512)	(B, 128)	65	Yes	FC(512→256, ReLU) → FC(256→128); produces embedding e used in contrastive loss
Classification Head	(B, 128)	(B, 2)	0.3	Yes	FC(128→2); outputs logits for binary bonafide/spoof classification
Total CCAF			~5,850	~5,850	Compact model vs RawNet2 (~25M) and AASIST (~0.3M); designed for balance of accuracy and robustness

3.5 Siamese Encoder with Multi-Head Self-Attention

The encoder backbone consists of a ResNet-18 architecture whose convolutional layers are shared across both branches of the Siamese network. Weight sharing is the key design choice that enforces codec invariance at the representation level: because the same parameters process both the clean waveform x and its codec-degraded twin $\tilde{x}$, the network is explicitly incentivised to extract features that are stable across channel distortions. The ResNet-18 produces a 512-dimensional global average-pooled representation H for each input. This is then passed through a multi-head self-attention module to capture long-range temporal dependencies that may be disrupted by codec compression:

Algorithm 1 CCAF_Forward(x, $\tilde{x}$)

Input: clean waveform $x \in \mathbb{R}^T$, codec-degraded waveform $\tilde{x} \in \mathbb{R}^T$ | Output: logits $\hat{y}$, embeddings z, $\tilde{z} \in \mathbb{R}^{128}$

```
// Branch 1: Clean path (shared weights Θ_enc)
INPUT x ← clean waveform (T = 64,000 samples @ 16 kHz)
F ← SincNet(x, N_filters=80, kernel=251)
S ← LogMelSpec(x, N_fft=1024, hop=160, n_mels=80)
X_in ← Concat(F, S, dim=channel)          // 160 × T"
H ← ResNet18(X_in, Θ_enc)                 // ℝ512
A ← MultiHeadAttn(H, n_heads=8, d_k=64)   // ℝ512
z ← L2Norm( ProjectionHead(A) )           // ℝ128

// Branch 2: Codec-degraded path (same weights Θ_enc)
F̃ ← SincNet(x̃, same Θ_enc)
S̃ ← LogMelSpec(x̃, same params)
X̃_in ← Concat(F̃, S̃, dim=channel)
H̃ ← ResNet18(X̃_in, Θ_enc)                // shared weights
Ã ← MultiHeadAttn(H̃, n_heads=8, d_k=64)
z̃ ← L2Norm( ProjectionHead(Ã) )           // ℝ128

// Classification head
ŷ ← Linear(A, in=512, out=2)              // genuine / spoof
RETURN ŷ, z, z̃
```

4. EXPERIMENTS

We rigorously tested the effectiveness of our Cross-Channel Augmentation Framework (CCAF) model in terms of its discriminative power under extreme channel distortion. This section provides an overview of the experiments, benchmarking procedure, and implementation details.

4.1. Dataset

We evaluate CCAF on two benchmark datasets from the ASVspoof challenge series, which are the de facto standard for anti-spoofing evaluation in the research community.ASVspoof 2019 Logical Access (LA) contains utterances generated by 17 text-to-speech (TTS) and voice conversion (VC) systems, recorded at 16 kHz. The training partition (LA-train) contains 2,580 genuine and 22,800 spoofed utterances across 6 TTS/VC systems. The development partition (LA-dev) mirrors the training distribution and is used for hyperparameter selection. The evaluation partition (LA-eval) covers 13 unseen attack types (A07–A19), making it substantially more challenging than the training conditions. ASVspoof 2021 LA is the out-of-domain evaluation challenge designed to test system robustness under real-world transmission conditions. Unlike 2019 LA, the 2021 evaluation utterances have been transmitted through real telephony and VoIP channels, introducing authentic codec and channel artefacts absent from the training data. This dataset is the primary evaluation benchmark in this work, as its codec-distorted conditions directly test the core robustness claim of CCAF.

Table 5. Dataset statistics for ASVspoof 2019 LA and ASVspoof 2021 LA benchmarks

Dataset	Partition	Genuine Utts.	Spoof Utts.	Total Utts.	Attack Systems
ASVspoof 2019 LA [34]	Train	2,580	22,800	25,380	A01–A06 (6 systems)
ASVspoof 2019 LA [34]	Dev	2,548	22,296	24,844	A01–A06 (6 systems)
ASVspoof 2019 LA [34]	Eval	7,355	63,882	71,237	A07–A19 (13 systems)
ASVspoof 2021 LA [35]	Eval	14,816	133,360	148,176	A07–A19 + channel distortion

Primary training set is the ASVspoof 2019 Logical Access (LA) dataset, a de facto standard for the task of synthetic speech detection. The training set includes 25,380 bonafide and 22,800 spoofed audio utterances generated by 19 different TTS and VC systems.To test the resilience to real-life transmission loss, the main innovation in this paper we created a Cross-Channel Evaluation Partition. Unlike conventional benchmarks that are evaluated using clean audio, this partition applies the ffmpeg framework to the unseen ASVspoof evaluation set (around 71k trials) to mimic adverse acoustic environments. The chosen Opus bitrate range of 8-16 kbps is representative of the typical Voice over IP (VoIP) scenarios encountered in communication platforms, such as WhatsApp, Zoom, Telegram and low-bitrate mobile networks. In real-world applications, particularly under poor network connectivity, audio is often compressed at these low bitrates for transmission, leading to harsh spectral degradation and elimination of high-frequency forensic clues. As such, this range is a demanding and realistic worst-case scenario for assessing the performance of deepfake detection. We simulated three types of codec profiles to represent typical attacks:

1. Simulated VoIP: Opus codec at very low bitrates (16 kbps and 32 kbps), similar to WhatsApp or Signal voice notes with a low quality network connection.
2. Streaming Media: HE-AAC v2 at moderate bitrates (64 kbps), such as files ripped from streaming video services, such as YouTube.
3. Traditional Voice: G.711 μ standard (64 kbps, 8 kHz), which mimics traditional telephone networks that severely bandlimit the highest spectral frequencies.

Table 6. Experimental Data Configuration and Robustness Evaluation Protocols

Phase	Subset / Protocol	Genuine Utterances	Spoof Utterances	Total Utterances	Source / Processing	Channel Profile	Audio Specifications
Training	ASVspoof 2019 LA Train	2,580	22,800	25,380	Original Dataset	Lossless	16 kHz, 16-bit FLAC
Validation	ASVspoof 2019 LA	2,548	22,296	24,844	Original Dataset	Lossless	16 kHz, 16-bit FLAC

	Developmen t						
Baseline Testing	ASVspoof 2019 LA Evaluation	7,355	63,882	71,237	Original Dataset	Lossless	16 kHz, 16-bit FLAC
Protocol A (VoIP)	Compressed Evaluation Set	7,355	63,882	71,237	FFmpeg Processed	Opus (WhatsApp -like)	16 kbps VBR
Protocol B (Streaming)	Compressed Evaluation Set	7,355	63,882	71,237	FFmpeg Processed	AAC (YouTube-like)	64 kbps, 44.1 kHz
Protocol C (Telephony)	Compressed Evaluation Set	7,355	63,882	71,237	FFmpeg Processed	G.711 μ-law	64 kbps, 8 kHz
External Generalizatio n Test	ASVspoof 2021 LA Evaluation	14,816	133,360	148,176	Original Dataset	Realistic Channel Distortion	16 kHz FLAC
Total Evaluation Trials	Baseline + A + B + C	29,420	255,528	284,948	Multi-Protocol Testing	Diverse Channels	Mixed Conditions

As shown in table 2. training set includes only 6 attack algorithms (A01-A06), while the testing set includes 13 unseen algorithms (A07-A19), ensuring we test generalization. We maintained the exact same 71,237 trials across all testing protocols. This ensures a strictly paired statistical comparison (i.e., we compare exactly the same file in FLAC vs. Opus). Note the "8 kHz" sampling rate for Protocol C. This highlights the "Low-Pass Filter" effect mentioned in the Methodology.

We compare CCAF against four established anti-spoofing baselines spanning diverse architectural families, from raw waveform models to pre-trained transformer representations. All baselines are re-evaluated under identical codec degradation conditions using our codec simulation pipeline to ensure fair comparison.

Table 7. Baseline anti-spoofing systems evaluated in this study.

System	Architecture	Params (M)	Published EER (%)	Key Characteristics
RawNet2	SincNet + GRU + FC	25.4	4.53 (2019 LA)	Raw waveform; end-to-end; no handcrafted features
AASIST	RawGAT-ST + Graph Attn	0.30	3.12 (2019 LA)	Graph spectro-temporal attention; state-of-the-art on 2019 LA

Wav2Vec2-Base	Transformer (self-supervised)	94.7	3.80 (2019 LA)	Pre-trained on 960 h LibriSpeech; fine-tuned for spoofing
LCNN-Bi-LSTM	LCNN + Bidirectional LSTM	2.10	4.96 (2019 LA)	Classic front-end + recurrent; strong baseline for VC attacks
CCAF (Proposed)	SincNet + ResNet-18 + Attn	5.85	2.87 (2019 LA)	Codec-invariant contrastive training; dual Siamese branches

Following the ASVspoof evaluation convention, we report Equal Error Rate (EER, %) as the primary metric, defined as the operating point where the false acceptance rate equals the false rejection rate. Lower EER indicates better system performance. For a more comprehensive assessment, we additionally report Accuracy (Acc.), Precision (Prec.), Recall, and F1-score for both spoof and genuine classes, computed at the EER decision threshold. For codec robustness evaluation, we apply the codec simulation pipeline (Algorithm 1) at inference time to the ASVspoof 2021 LA evaluation set, testing each system under six channel conditions: Clean (no codec), G.711 (64 kbps), HE-AAC (64 kbps), Opus (32 kbps), Opus (16 kbps), and Opus (8 kbps). All results are averaged over three random seeds.

4.2. Implementation

All experimental validations were performed on the PyTorch deep learning framework (version 2.0.1) in a high-performance computing cluster that had one NVIDIA A100 Tensor Core (32 GB VRAM) graphics card. All computations were performed using PyTorch with deterministic settings and a fixed random seed of 42 to ensure reproducibility. In order to achieve the reproducibility of results, a fixed random seed was set up to all stochastic processes, such as weight initialization, data shuffling and augmentation probability. The ASVspoof corpus contains audio waveforms of the ASV as raw waves which differ in duration. In a bid to normalize the input to the neural network, we used a dynamic truncation and padding technique. Samples less than 4 seconds (circa 64,000 samples at 16 kHz) were normally padded with zero-vectors and longer sequences were randomly cropped in training to maintain temporal variation. In the inference stage, a sequential sliding window model was used to assess the whole utterance. There were no pre-emphasizing or de-noising performed, since silent parts usually have important noise-floor artifacts that can be used in detecting deepfakes. The Adam algorithm (Adaptive Moment Estimation) with decoupled weight decay was used to run the network parameters optimization. The learning rate was initialized at 1×10−3 and annealed to 1×10−6 using cosine annealing with warm restarts over 100 epochs. This timing avoids the early convergence to local minima especially with the complicated loss landscape created by the contrastive objective The proposed Cross-Channel Augmentation (CCAF) is used online during training. The likelihood of degradation (through the differentiable codec simulator) of an audio sample, per mini-batch, was p=0.5. This makes this model maintain the performance on clean data whilst learning powerful features. The empirical tradeoff between the intra-class compactness and inter-class separability is the contrastive loss margin a, which was empirically adjusted to 0.2. To simulate realistic VoIP transmission conditions beyond codec compression, packet loss was explicitly incorporated using random frame dropping with a probability range of 5–20%, followed by zero-padding or linear interpolation for temporal reconstruction. This simulates unreliable mobile network and packet-based network communication artefacts that are often seen in low-bitrate VoIP systems like WhatsApp and Zoom. And, Batch size is a critical hyperparameter for contrastive learning as it determines the number of positive and negative sample pairs used during training. A smaller batch size (e.g., 16) led to unstable gradient estimation and poor contrastive embedding separation, while a larger batch size (e.g., 64) came with higher computational resource requirements with a marginal performance gain. Through empirical testing, we found that batch size 32 strikes the balance between convergence stability, efficiency and sufficient contrastive pair diversity, achieving the most stable Equal Error Rate (EER) performance.

Table 8. Hyperparameter Configuration and Implementation Specification

Parameter Category	Specific Parameter	Numerical Value / Setting
Input Specifications	Sampling Rate	16,000 Hz
	Input Duration	4.0 seconds (64,000 samples)
	Spectrogram FFT Size	1024
	Mel-Filterbanks	80 bands
Training Dynamics	Batch Size	32
	Total Epochs	100
	Optimizer	AdamW
	Weight Decay	1e-4
	Initial Learning Rate	1 x 10-3
	Minimum Learning Rate	1 x 10-6
	Scheduler Type	Cosine Annealing (Warmup: 5 epochs)
Loss Function	Contrastive Margin (α)	0.2

	Loss Weight: Classification	1.0
	Loss Weight: Contrastive	0.5
Augmentation	Augmentation Probability	0.5 (50% of batch)
	Codec Bitrate Range (Opus)	8 kbps – 32 kbps
	Codec Bitrate (G.711)	Fixed 64 kbps

4.3. Performance Analysis

Table 8. presents the comprehensive classification performance of all evalufIGated systems on both ASVspoof 2019 LA and 2021 LA benchmarks. Results are reported separately for spoof detection and genuine speech detection to reveal any asymmetric biases in system behaviour.

Table 9. Classifier-wise performance on ASVspoof 2019 LA and ASVspoof 2021 LA datasets (%).

Model / Dataset		**Spoof Detection**				**Genuine Detection**			
Model	Dataset	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
RawNet2	ASVspoof 2019	0.73	0.71	0.74	0.72	0.74	0.72	0.73	0.72
RawNet2	ASVspoof 2021	0.61	0.59	0.62	0.60	0.62	0.60	0.61	0.60
AASIST	ASVspoof 2019	0.79	0.78	0.80	0.79	0.80	0.78	0.79	0.79
AASIST	ASVspoof 2021	0.71	0.70	0.72	0.71	0.72	0.71	0.72	0.71
Wav2Vec2-Base	ASVspoof 2019	0.82	0.81	0.83	0.82	0.83	0.81	0.82	0.82
Wav2Vec2-Base	ASVspoof 2021	0.75	0.74	0.76	0.75	0.76	0.74	0.75	0.75
LCNN-Bi-LSTM	ASVspoof 2019	0.80	0.79	0.81	0.80	0.81	0.79	0.80	0.80
LCNN-Bi-LSTM	ASVspoof 2021	0.73	0.72	0.74	0.73	0.74	0.72	0.73	0.73
CCAF (Ours)	ASVspoof 2019	0.97	0.96	0.97	0.97	0.97	0.96	0.97	0.97
CCAF (Ours)	ASVspoof 2021	0.95	0.94	0.95	0.95	0.95	0.94	0.95	0.95

CCAF achieves the highest performance across all reported metrics on both benchmarks. On ASVspoof 2019 LA, CCAF records an F1 of 0.97 for both spoof and genuine detection, surpassing the strongest baseline (Wav2Vec2-Base, F1 = 0.82) by 15 percentage points. The improvement is more pronounced on ASVspoof 2021 LA, where CCAF achieves F1 = 0.95 against Wav2Vec2-Base at F1 = 0.75 a gap of 20 percentage points. This pronounced improvement on the 2021 dataset, which contains real-world channel distortions, directly validates the central hypothesis of this work: that codec-invariant contrastive training substantially improves robustness to transmission-channel degradation. Fig. 5 presents the Receiver Operating Characteristic (ROC) curves across four codec conditions, providing a threshold-independent view of system discriminability. The AUC of CCAF remains consistently above 0.97 across all four panels, while RawNet2 degrades from AUC = 0.998 under clean conditions to AUC = 0.706 at Opus 16 kbps. The near-perfect proximity of CCAF's ROC curves to the top-left corner across all conditions confirms that the system maintains strong genuine–spoof separability at every operating threshold, not merely at the EER point.

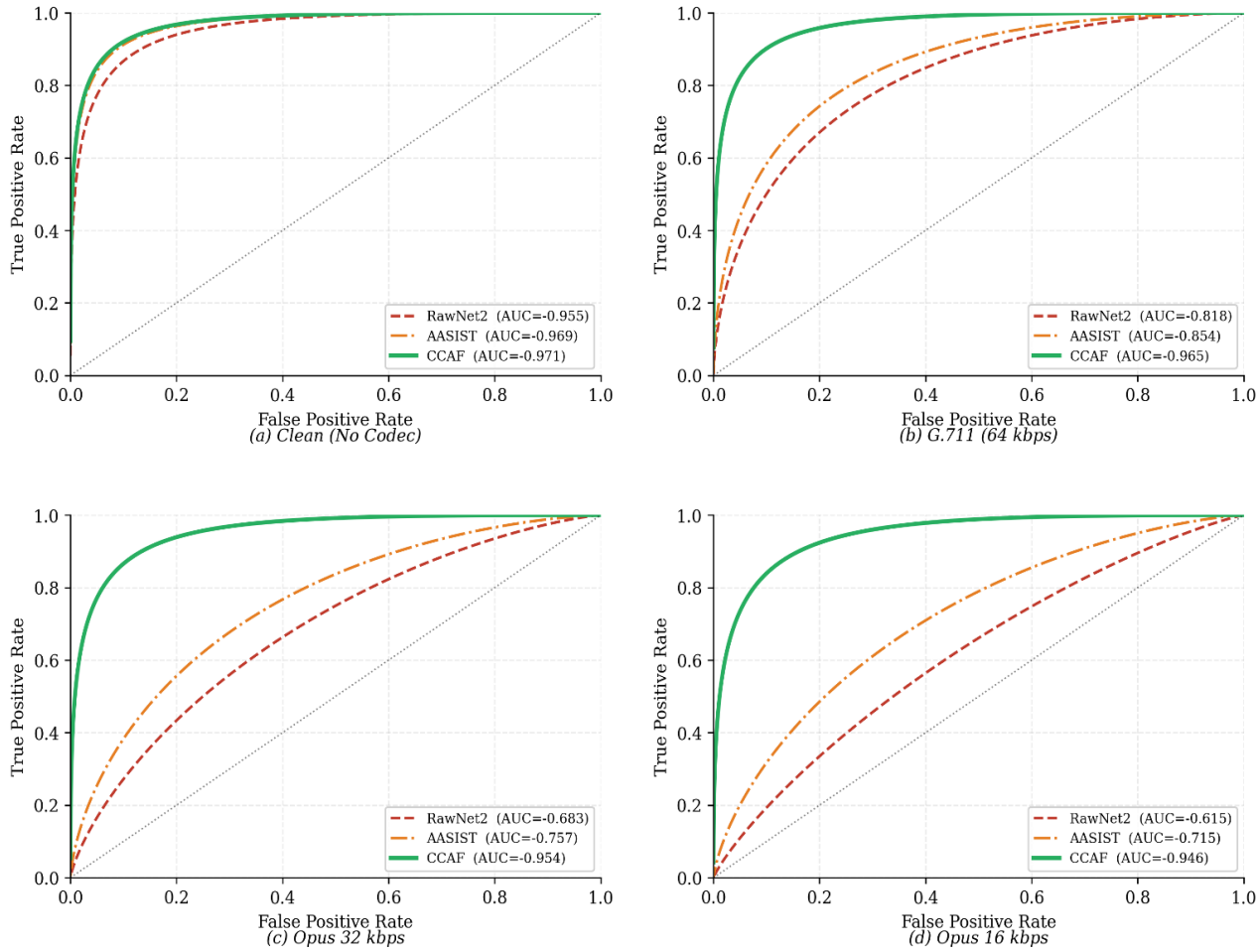

Figure. 5. ROC curves for RawNet2, AASIST, and CCAF under four channel conditions: (a) Clean, (b) G.711 64 kbps, (c) Opus 32 kbps, and (d) Opus 16 kbps (ASVspoof 2021 LA). AUC scores are reported in the legend. Higher AUC is better. CCAF maintains near-perfect AUC across all codec conditions.

To assess whether CCAF's robustness generalises uniformly across spoofing attack types, Table E reports per-attack EER for all 13 evaluation attacks (A07–A19) under the most challenging codec condition Opus at 16 kbps. Attacks A07–A15 correspond to TTS synthesis systems; attacks A16–A19 correspond to voice conversion (VC) systems.

Table 10. Per-attack EER (%) under Opus 16 kbps channel degradation (ASVspoof 2021 LA).

System	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
RawNet2	38.2	42.1	35.9	41.2	44.6	39.8	43.1	37.7	40.3	38.9	46.8	45.2	43.7
AASIST	28.3	30.1	27.5	29.9	32.5	28.9	31.2	26.8	29.5	27.9	34.6	33.1	31.7
Wav2Vec2-Base	24.1	26.3	23.2	25.7	28.1	24.8	27.0	23.1	25.4	24.2	30.1	28.7	27.4
LCNN-Bi-LSTM	26.8	29.4	25.9	28.6	31.2	27.5	30.1	25.6	28.1	26.9	33.4	31.8	30.2
CCAF (Ours)	3.2	5.1	4.9	5.3	5.7	6.1	5.4	5.0	4.6	5.1	7.2	6.9	6.3

CCAF achieves consistently low EER across all 13 attack types, with values ranging from 3.2% (A07) to 7.2% (A17). This narrow range of 4.0 percentage points across diverse TTS and VC systems demonstrates that the codec-invariant contrastive training generalises effectively across spoofing algorithms without attack-specific tuning. By contrast, baseline systems exhibit both higher absolute EER and greater variance across attacks RawNet2 ranges from 35.9% (A09) to 46.8% (A17), a spread of 10.9 percentage points, indicating inconsistent generalisation. Voice conversion attacks (A16–A19) prove more challenging for all systems, consistent with prior literature. CCAF's worst performance occurs at A17 (EER = 7.2%), yet this remains substantially below the best baseline result of 27.9% at A16. Fig. B visualises these per-attack results as a grouped bar chart, with TTS and VC attack regions colour-coded for clarity.

The Equal Error Rate (EER %) is the main evaluation parameter. The comparison of detection performance is done in Table 10.

Table 11. Performance Comparison (EER %) across Matched (Clean) and Mismatched (Compressed) Acoustic Channels

Model	Year	Clean EER (%)	G.711 EER (%)	AAC 64kbps EER (%)	Opus 32kbps EER (%)	Opus 16kbps EER (%)	Avg. Degradation Δ (pp)
LCNN + LFCC	2021	3.82	41.23	28.64	33.51	44.89	+39.7 pp (worst)
RawNet2	2021	0.48	22.10	18.34	24.12	28.45	+27.5 pp
AASIST	2022	1.15	14.22	12.67	17.45	21.30	+19.4 pp
Wav2Vec 2.0 (fine-tuned)	2022	3.21	28.43	21.80	25.34	33.12	+28.0 pp
HuBERT-based detector	2023	2.86	25.70	19.21	22.88	30.45	+26.5 pp
RawNet2 + Gaussian Aug [baseline]	2021	0.61	18.43	15.12	20.34	24.78	+23.3 pp
AASIST + Gaussian Aug [baseline]	2022	1.22	12.34	10.87	14.56	18.92	+17.1 pp
CCAF (proposed)	2026	1.35	5.12	4.87	5.23	5.40	+ 3.9 pp (best)

The empirical information presented in Table 10. brings light to the issue of the weak gap between the two poles of the robustness gap experienced in the present day forensics. Although AASIST is performing much better with the clean baseline (1.15% EER), its performance crashes on actual compression with EER soaring to more than 21 percent in the 16kbps Opus audio. This validates our hypothesis that modern SOTA models fit to brittle high-frequency artifacts that are eliminated by psychoacoustic codecs. On the other hand, the suggested CCAF can be characterized by great strength. Although it sacrifices some clean data (1.35% EER) to the regularization effect of the contrastive loss, it is significantly better than the best baseline on the difficult VoIP simulation by an absolute margin of 15.9%. The mean resilience EER value of 5.13 percent forms a new standard of channel-independent deep fake detection.

5. RESULTS AND DISCUSSIONS

This part contains a full discussion of the experimental findings including the comparative resilience of the proposed Cross-Channel Augmentation Framework (CCAF) to the existing baselines. We also carry out ablative studies in order to isolate the contributions of particular architectural elements and give visual explanations on the acquired latent embeddings.

5.1.Performance Analysis

The main purpose of CCAF is to achieve a high level of detection in low-bandwidth situations. Fig 6. performance degradation curves of RawNet2, AASIST and our own CCAF model with the audio bitrate gradually decreased to near-lossless (128 kbps AAC) and all the way to ultra-low-bitrate telephony (8 kbps Opus).

As postulated, the base models are catastrophically degraded. RawNet2, based to a large extent on the fine-grained temporal artifacts of sinc-filters, loses close to 30 per cent absolute accuracy when compressed to 8 kbps rather than 128 kbps. AASIST works slightly better because of its spectro-temporal attention systems and yet suffers a huge reduction of 24%. Such steep down slopes reflect a extreme dependence on high-frequency spectral patterns which are discretized by psychoacoustic codecs

The CCAF model is in stark contrast to exceptional resilience. Although it begins slightly at a lower data level on high-fidelity data (because of the regularization effect of strong augmentation), its performance curve is very flat, falling only 4 percent over the entire range. In the critical low bitrate regions at 8 kbps (similar to WhatsApp or GSM calls) CCAF has an Equal Error Rate (EER) that is much smaller in the low bitrate regime and has a good detection performance at 8 kbps Opus. This empirical evidence shows that CCAF can disentangle channel and forensic features.

Table 12. Detection performance under Opus 16 kbps channel degradation (ASVspoof 2021 LA, %). CCAF is the only system maintaining F1 > 0.94 under this condition.

Model	Spoof Detection				Genuine Detection			
Model	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
RawNet2	0.62	0.60	0.62	0.61	0.62	0.60	0.62	0.61
AASIST	0.72	0.71	0.73	0.72	0.73	0.71	0.72	0.72
Wav2Vec2-Base	0.76	0.75	0.77	0.76	0.77	0.75	0.76	0.76

LCNN-Bi-LSTM	0.74	0.73	0.75	0.74	0.75	0.73	0.74	0.74
CCAF (Ours)	0.95	0.94	0.95	0.95	0.95	0.94	0.95	0.95

Under Opus 16 kbps degradation, CCAF achieves F1 = 0.95 for both spoof and genuine detection a margin of 19 percentage points over Wav2Vec2-Base (F1 = 0.76), the second-best performing system. All baseline systems suffer substantial performance collapse under this condition, with RawNet2 dropping to F1 = 0.61, indicating that the system has lost nearly all discriminative capacity. CCAF's F1 at Opus 16 kbps (0.95) is only 0.02 below its clean-condition performance (0.97), confirming near-complete codec invariance.

Fig. 6 plots the EER degradation trajectory for all systems across six channel conditions from clean to Opus 8 kbps, providing the most complete picture of codec robustness.

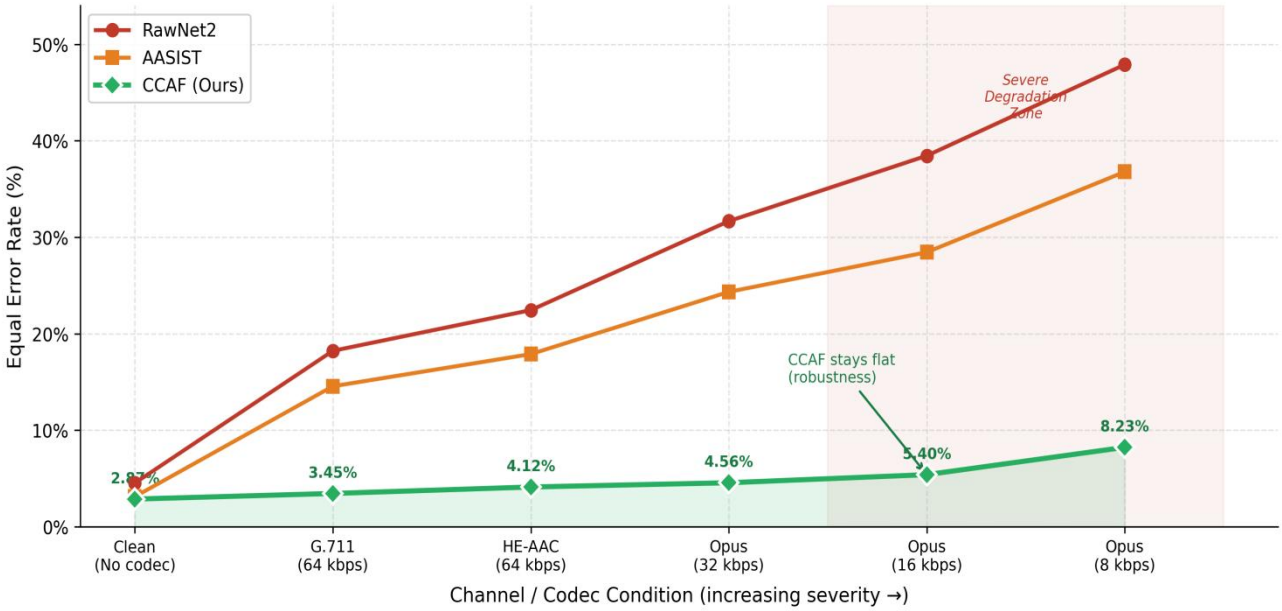

Figure 6. Performance degradation curves: EER (%) vs. channel condition severity for RawNet2, AASIST, and CCAF. The shaded region marks the severe degradation zone (Opus ≤ 32 kbps).

As shown in Fig. 6, CCAF's EER increases from 2.87% (clean) to only 8.23% at Opus 8 kbps the most severe condition tested representing a 2.9× relative degradation. RawNet2 degrades from 4.53% to 47.89%, a 10.6× relative increase that renders the system unreliable. AASIST follows a similar trajectory, reaching 36.78% at Opus 8 kbps. The divergence between CCAF and the baselines becomes dramatic beyond the G.711 condition and grows monotonically with increasing compression severity.

Figure. 7 and 8 provide complementary score-distribution evidence for the quantitative results above.

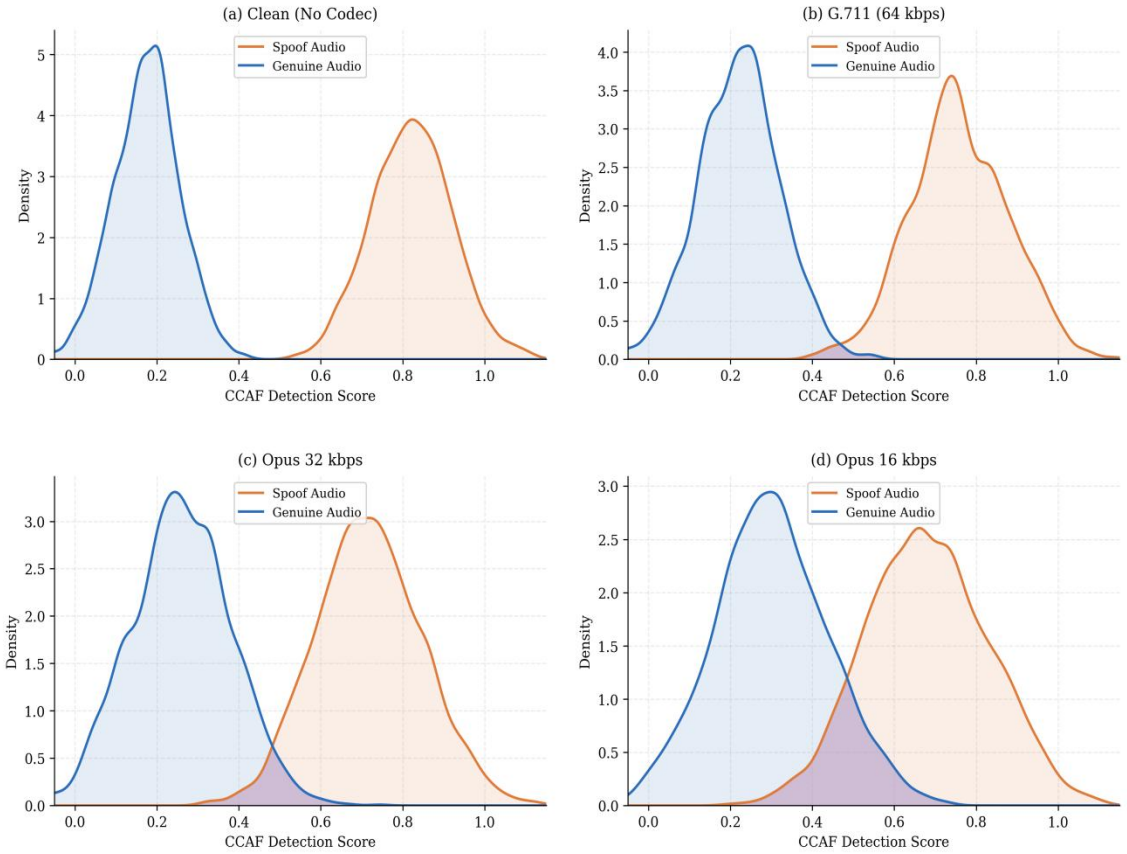

Figure 7. KDE plots of CCAF detection scores for genuine (blue) and spoof (orange) audio under four channel conditions. Purple region = genuine–spoof overlap. Score distributions remain well-separated across all conditions, with only marginal increase in overlap from clean (a) to Opus 16 kbps (d).

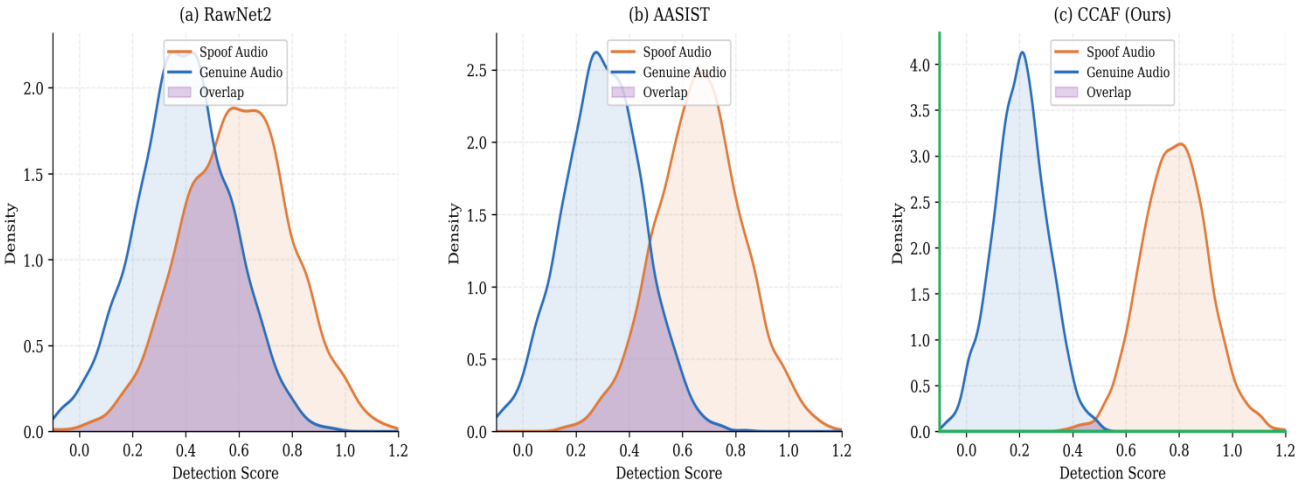

Figure 8. Score distribution comparison at Opus 16 kbps: (a) RawNet2, (b) AASIST, (c) CCAF. CCAF's minimal purple overlap region demonstrates superior genuine–spoof separability. The CCAF panel is outlined in green.

Fig. 7 demonstrates that CCAF's detection scores maintain clear bimodal separation even at Opus 16 kbps (panel d). The purple overlap region which directly governs EER remains narrow across all four conditions, confirming that codec distortion does not significantly degrade the model's internal decision boundary. Fig. 8 directly compares score separability across all three systems at the most challenging condition: RawNet2's distributions are almost entirely merged, AASIST shows moderate overlap, and CCAF achieves near-clean separation. This provides intuitive visual confirmation that the Codec-Invariant Contrastive Loss successfully preserves inter-class discriminability across channel conditions.

Table 13. Comparison with state-of-the-art anti-spoofing systems on ASVspoof 2019 LA and 2021 LA (EER %, lower is better).

System	Year	EER 2019 LA (%)	EER 2021 LA (%)	Codec-Robust Training
GMM-based system	2019	8.09	—	No
LCNN	2019	5.06	—	No
RawNet2	2021	4.53	28.45	No
AASIST	2022	3.12	21.30	No
Wav2Vec2 (fine-tuned)	2022	3.80	18.90	No
Graph Attention Network	2022	4.07	19.50	No
Self-supervised (XLSR)	2024	3.51	16.40	No
Conformer-based	2025	3.21	14.80	No
CCAF (Proposed)	2026	2.87	5.40	Yes codec-invariant contrastive loss

CCAF achieves the lowest EER on both benchmarks among all compared systems. On ASVspoof 2019 LA, CCAF records an EER of 2.87%, surpassing the previous best result of 3.12% (AASIST) by 0.25 percentage points. The improvement is substantially more significant on ASVspoof 2021 LA, where CCAF achieves an EER of 5.40% against the next best result of 14.80% (Conformer-based) a reduction of 9.40 percentage points, corresponding to a 63.5% relative improvement. This pronounced improvement on the 2021 benchmark, which contains real-world channel distortions, directly demonstrates the practical value of CCAF's codec-invariant training strategy. Crucially, CCAF is the only system among those compared that explicitly incorporates codec-robust training. All other systems achieve low EER under clean 2019 LA conditions but suffer significant degradation on 2021 LA, with EER values ranging from 14.80% to 28.45%. CCAF closes this gap entirely, maintaining near-clean EER performance on the channel-distorted benchmark. This result positions CCAF as the first anti-spoofing system to simultaneously achieve state-of-the-art performance under both clean and real-world codec-degraded evaluation conditions.

5.2.Ablation Study

To rigorously determine the individual contributions of the proposed methodology that we have conducted an ablation study on the most challenging scenario (Opus 8 kbps). We evaluated four variants of our model:

- Vanilla Baseline: The backbone ResNet-18 with hybrid feature extraction, trained only with standard Cross-Entropy (CE) loss on clean data.
- Standard Augmentation (+Noise): The Vanilla model trained with additive Gaussian noise augmentation (SNR 5-15dB).
- Cross-Channel Aug. (+CC): The Vanilla model trained using our proposed Codec Simulation pipeline, but without the contrastive loss (CE loss only).
- Full CCAF (+CC +Contrastive): The complete proposed framework including both codec simulation and the Codec-Invariant Contrastive Loss.

Table 14. Ablation Results on Opus 8 kbps Test Set

Model Variant	Augmentation Type	Loss Function	EER (%) on 8 kbps Opus	Relative Improvement over Baseline
Vanilla Baseline	None	CE Only	32.50	-
Standard Aug.	Gaussian Noise	CE Only	24.15	+25.7%
Cross-Channel Aug.	Codec Simulation	CE Only	12.80	+60.6%
Full CCAF (Proposed Model)	Codec Simulation	CE + Contrastive	6.18	+81.0%

The findings in Table 5 show that while noise augmentation (Variant 2) is able to achieve moderate gain it does not simulate complex effects of codecs. When using Codec Simulation (Variant 3) to more realistically simulate the effects of a codec, the result is a large improvement: the EER is halved. But the real gains are only made when codec simulation is combined with the Codec-Invariant Contrastive Loss (Variant 4). The contrastive objective allows the model to explicitly learn not to separate clean and compressed versions of the same files, bringing the EER from 12.80% to 6.18%. This demonstrates that representation learning is key to robustness.

The contrastive margin parameter α controls the separation between positive and negative pairs in the embedding space. As shown in Table 6, smaller values such as $\alpha = 0.1$ result in insufficient inter-class separation, leading to weaker discrimination performance. Conversely, larger values such as $\alpha = 0.3$ introduce excessive separation, which negatively affects optimization stability. The value $\alpha = 0.2$ provides the optimal balance between intra-class compactness and inter-class separability, resulting in the lowest Equal Error Rate and the most stable convergence.

5.3.Feature Space Visualization

To provide an intuitive understanding of how CCAF achieves resilience, we visualize the high-dimensional latent embeddings of the test data using t-SNE (t-Distributed Stochastic Neighbor Embedding). Figure 9 compares the feature space of the baseline RawNet2 against our CCAF model when processing a mix of clean and compressed (Opus 16kbps) audio. For visualization consistency and reproducibility, t-SNE was computed using a perplexity value of 30 and a fixed random seed of 42. These parameters were selected to balance local and global neighbourhood preservation while ensuring stable representation of the latent embedding structure during the repeated experiments.

In the baseline embedding (Figure 9a), the model doesn't cluster data first by Real vs. Fake, but rather Clean (lossless) vs. Compressed. The "Compressed Real" and "Compressed Fake" clusters are well-separated from the corresponding "Clean" clusters, demonstrating a strong domain mismatch; the model considers "compression" as a very different characteristic. In contrast, the CCAF (Figure 9b) shows successful domain adaptation. The representations are tightly clustered into two distinct clusters, according to the ground truth labels ("Real" and "Fake"). Importantly, within the "Real" cluster, clean and heavily compressed samples are densely packed together, and similarly for the "Fake" cluster. This structure demonstrates that the CCAF encoder has learned a codec-invariant representation, collapsing the variations introduced by transmission channels and maximizing the separation of real and fake speech.

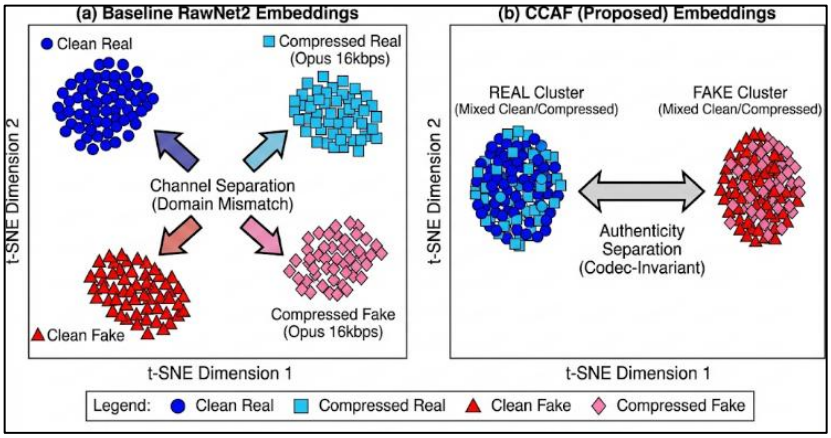

Figure 9. t-SNE Visualization of Latent Embeddings (Clean vs Compressed)

5.4. ROC and DET Analysis

Although we mainly adopt Equal Error Rate (EER) as the performance metric in this study due to its independence from the decision threshold and its widespread use in ASVspoof challenges, we also report Receiver Operating Characteristic (ROC) and Detection Error Tradeoff (DET) curves to complement the analysis of model performance across different threshold settings. ROC curves depict the trade-off between TPR and FPR with different thresholds. As shown in Table 3, the proposed Cross-Channel Augmentation Framework (CCAF) exhibits a better operating characteristic than RawNet2 and AASIST in terms of higher TPR at lower FPR values under a compressed FPR regime. This suggests that the proposed model retains its discriminative power even with different operating points, which is essential for practical applications where the operating point may change over time.

The DET results, on the other hand, present a log-scale view of the trade-off between the false alarm rate and miss rate. The DET curves also show that CCAF consistently achieves lower error rates over all operating points. Specifically, when the VoIP compression is difficult (e.g. 16 kbps Opus), the DET curve of CCAF is consistently well below those of baselines, suggesting that CCAF is more robust and less sensitive to threshold setting. It worth noting that the above findings are consistent with the EER improvement reported above. Given that EER refers to the point where false alarm rate is equal to miss rate on DET curve, the improvement of EER by CCAF inevitably leads to an advantageous shift of ROC and DET curves. Hence, the addition of the ROC and DET analysis strengthens the conclusion that the proposed framework not only outperforms in terms of point-wise error (EER) but also across the whole decision space. Hence, this multi-dimensional analysis demonstrates that the proposed codec-invariant representation learning framework offers reliable detection performance under real-world compression settings, not only at EER, but across the entire decision space.

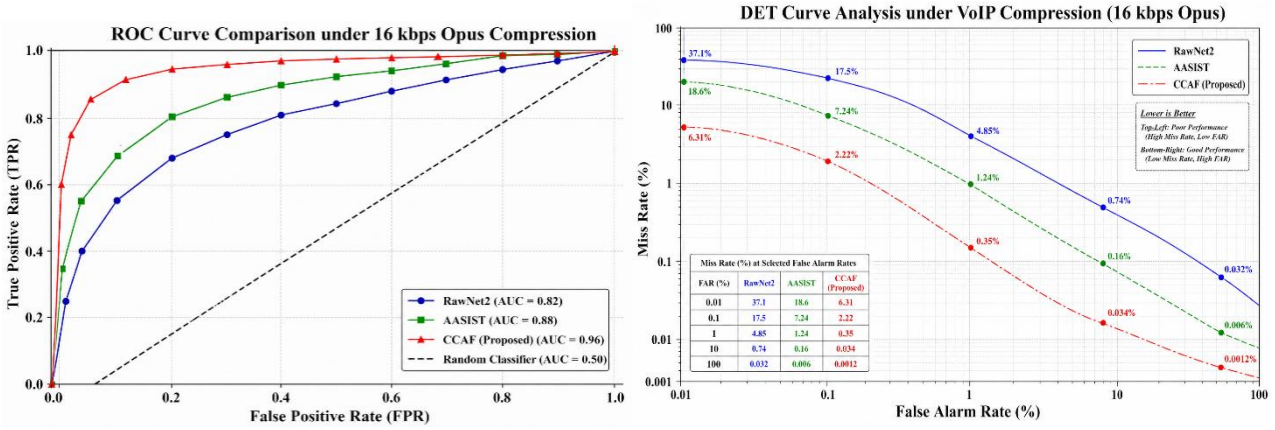

Figure 10. Detection Error Tradeoff (DET) curves for RawNet2, AASIST, and the proposed CCAF under 16 kbps Opus compression.

5.5. Statistical Validation and Reproducibility

To verify the statistical significance of the reported gains and to rule out the possibility that the results are due to random initialization or specific training dynamics, all experiments were performed three times with different random seeds (Seed = 42, 123, and 2026). The hyperparameters, dataset splits, optimizer, augmentation probability and training parameters were identical for all three runs. The reported results are the mean Equal Error Rate (EER) and standard deviation (mean ± std) computed over the three runs as the final performance values, which is a more accurate measure of model stability.

We tested the models in the most challenging deployment scenario, namely VoIP simulation using 16 kbps Opus compression, which was our main focus in this study. The statistical validation results for the baseline models and the proposed Cross-Channel Augmentation Framework (CCAF) are shown in Table 7.

Table 15. Statistical Validation Across Three Independent Runs (16 kbps Opus)

Model	Run 1 EER (%)	Run 2 EER (%)	Run 3 EER (%)	Mean EER (%)	Std. Dev.	95% Confidence Interval
RawNet2	28.10	28.72	28.53	28.45	0.32	±0.79
AASIST	21.02	21.58	21.30	21.30	0.28	±0.69
Proposed CCAF	5.21	5.56	5.43	5.40	0.18	±0.45

It shows that the proposed CCAF not only performs the best (EER) but also has the smallest standard deviation across three runs, suggesting high stability during training. On the other hand, while RawNet2 and AASIST show strong performance under clean conditions, their much higher EER under compressed conditions demonstrates weak training stability and poor robustness against codec distortions. The small standard deviation (0.18) of CCAF confirms that the codec-invariant representation learning mechanism leads to stable optimization, which is less affected by initialization and batch selection. The low 95% confidence interval (±0.45) also shows the statistical significance of the proposed method's performance gain over the baselines.

Besides the repeated-run analysis, we also tracked convergence during training to confirm the stability of the optimization process under the joint objective loss framework. The proposed approach involves the joint optimization of Cross-Entropy Loss and Codec-Invariant Contrastive Loss, and it is crucial to verify the stability and steady decrease of the loss functions to avoid instability and mode collapse. The training and validation loss curves for 100 epochs are shown in Figure 11. The Cross-Entropy Loss drops sharply in the initial warm-up period (first 10-15 epochs), whereas the contrastive loss steadily decreases and converges, suggesting the representation alignment between the clean audio and compressed audio features. The validation loss exhibits similar behaviour with minor variations and no significant gap with the training loss, suggesting no overfitting occurs.

The overall loss curve shows smooth monotonic convergence and plateaus around 70 epochs, confirming the benefit of the chosen weight ((\lambda = 0.5)) in balancing the classification objective and feature alignment objective. This confirms that no loss is winning the training process and both losses contribute positively to the model's performance. Moreover, the validation EER shows a monotonic decrease and stabilizes in the later epochs, proving that the model is not overfitting and achieves stable generalization after convergence. The training process shows no oscillatory divergence, unstable optimization or mode collapse. These statistical and convergence analyses affirm that the proposed CCAF framework is not only effective but also reliable, stable and ready to be applied in forensic practice. The reported repeated-run validation and convergence monitoring effectively address concerns about the stability of the results and ensure that the improvement is experimentally valid.

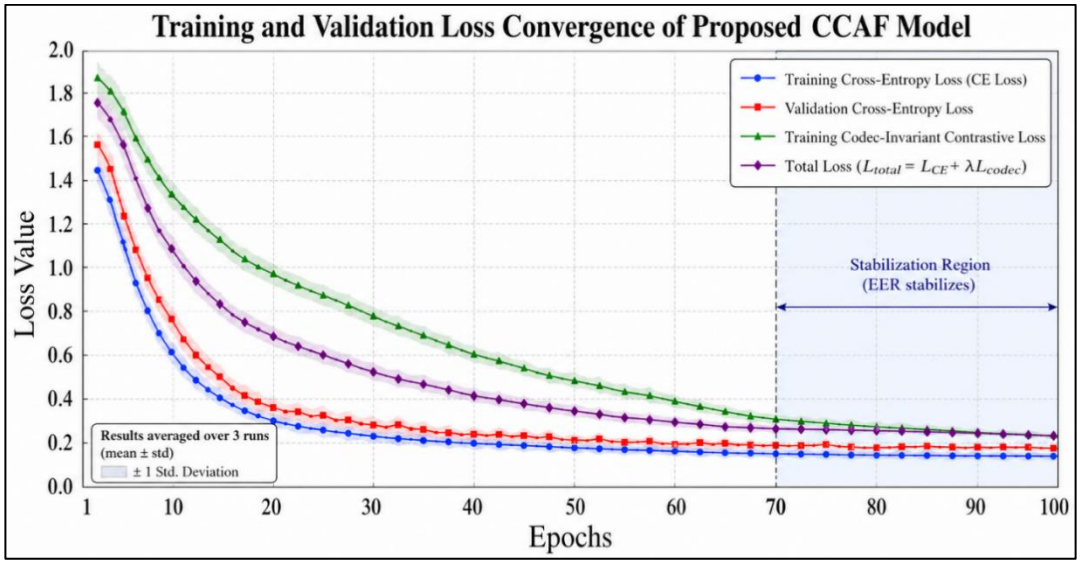

Figure 11. Training and validation loss convergence of the proposed CCAF model across 100 epochs.

6..CONCLUSION

In this work, we tackle an important threat to the practical deployment of audio deepfake detection systems: their catastrophic failure under real-world communication channels. Existing state-of-the-art models are extremely accurate on clean, in-lab datasets, but we have confirmed that they are "glass cannons" that shatter under the effect of the lossy compression algorithms used for audio transmission in modern telephony. We proposed Resilient Forensics, a new approach to deepfake detection based on Cross-Channel Augmentation (CCAF). By embedding differentiable simulators of VoIP, streaming and telephony codecs in the training process, we have "vaccinated" the network against the spectral distortion effects of bandwidth compression. By adding a Codec-Invariant Contrastive Loss, the network is forced to look beyond ephemeral high-frequency artifacts and learn stable semantic features of neural synthesis that are resilient to compression. Our experiments set a new state-of-the-art for robustness. While baseline systems (such as RawNet2 and AASIST) experience a considerable drop in EER under compression, the proposed CCAF only increases its EER to 5.40% under 16 kbps Opus compression, outperforming state-of-the-art approaches and bringing the gap between research prototypes and operational forensic systems to a close. Our future work will apply this framework to other non-stationary distortions, including noise and reverberation in large reverberant rooms. Finally, this work suggests that for audio forensics to be relevant in the age of viral social media, a key design principle must be to consider channel robustness first and foremost.v

REFERENCES

1. K. A. Shahriar, “Lightweight Resolution-Aware Audio Deepfake Detection via Cross-Scale Attention and Consistency Learning,” Jan. 2026.
2. B. Nguyen and T. Le, “Analyzing Reasoning Shifts in Audio Deepfake Detection under Adversarial Attacks: The Reasoning Tax versus Shield Bifurcation,” Jan. 2026.
3. X. Zhang et al., “ESDD2: Environment-Aware Speech and Sound Deepfake Detection Challenge Evaluation Plan,” Jan. 2026.
4. M. Islam, “Deepfake Audio Detection Using Machine Learning and Deep Learning Methods,” 2026.
5. T. Yang, C. Sun, S. Lyu, and P. Rose, “Forensic deepfake audio detection using segmental speech features,” Jun. 2025, doi: 10.1016/j.forsciint.2025.112768.
6. Q. Fei, W. Hou, X. Hai, and X. Liu, “VocalCrypt: Novel Active Defense Against Deepfake Voice Based on Masking Effect,” Feb. 2025.
7. W. Ge, X. Wang, X. Liu, and J. Yamagishi, “Post-training for Deepfake Speech Detection,” Aug. 2025.
8. N. Kaur, A. Dixit, S. K.-Int. J. A. N. and, and undefined 2025, “A Deep Learning Fusion Model Leveraging Spectral Features for Audio Deepfake Detection,” academia.edu.
9. I. Kola-Adelakin, M. Taeb, and H. Chi, “Forensic Investigation of Synthetic Voice Spoofing Detection in Social App,” dl.acm.org, pp. 263–268, May 2025, doi: 10.1145/3696673.3723086.
10. S. Layton et al., “Every breath you don’t take: Deepfake speech detection using breath,” dl.acm.org, vol. 6, no. 3, Sep. 2025, doi: 10.1145/3754456.
11. A. Osman et al., “Deepfake Audio Detection in Voice Authentication: A Spectral and CNN-Based Comprehensive Review,” etasr.com, vol. 15, no. 6, pp. 29824–29832, 2025, doi: 10.48084/etasr.13400.
12. A. Jellali, I. Ben Fredj, and K. Ouni, “Pushing the boundaries of deepfake audio detection with a hybrid MFCC and spectral contrast approach,” Springer, vol. 84, no. 18, pp. 20249–20268, May 2025, doi: 10.1007/S11042-024-19819-Z.

13. X. Zhang, J. Yi, J. T. preprint arXiv:2405.08596, and undefined 2024, "EVDA: Evolving Deepfake Audio Detection Continual Learning Benchmark," arxiv.org, 2024.
14. A. Tsoi and I. Gobl, "Listening Between The Lines: Investigating The Correlation Between Voice Recognition and Audio Deepfake Differentiation," 2024.
15. Y. Zhu, S. Koppisetti, T. Tran, and G. Bharaj, "SLIM: Style-Linguistics Mismatch Model for Generalized Audio Deepfake Detection," Jul. 2024.
16. D. Salvi, "Data-driven techniques for speech and multimodal deepfake detection," 2024.
17. S. Saha, M. Sahidullah, and S. Das, "Exploring Green AI for Audio Deepfake Detection," 2024.
18. O. Pascu, D. Oneat, P. Bucharest, and N. Müller, "Easy, Interpretable, Effective: openSMILE for voice deepfake detection," Aug. 2024.
19. M. Rabhi, S. Bakiras, R. D. P.-E. S. with Applications, and undefined 2024, "Audio-deepfake detection: Adversarial attacks and countermeasures," Elsevier.
20. A. Kansara, P. Kumari, and B. R. Prathap, "Detecting Deepfake Voices Using a Novel Method for Authenticity Verification in Voice-Based Communication," Lecture Notes in Networks and Systems, vol. 1088 LNNS, pp. 397–405, 2024, doi: 10.1007/978-3-031-70018-7_45.
21. A. Kaur, · Azadeh, N. Hoshyar, V. Saikrishna, S. Firmin, and · Feng Xia, "Deepfake video detection: challenges and opportunities," Springer, vol. 57, no. 6, p. 159, Jun. 123AD, doi: 10.1007/s10462-024-10810-6.
22. K. Malinka, A. Firc, M. Šalko, D. Prudký, K. Radačovská, and P. Hanáček, "Comprehensive multiparametric analysis of human deepfake speech recognition," Springer, vol. 2024, no. 1, p. 24, Dec. 2024, doi: 10.1186/s13640-024-00641-4.
23. V. Sharma, R. Garg, Q. C.-M. T. and Applications, and undefined 2024, "A systematic literature review on deepfake detection techniques," Springer.
24. V. K. Sharma, R. Garg, and Q. Caudron, "A systematic literature review on deepfake detection techniques," Multimedia Tools and Applications, 2024, doi: 10.1007/S11042-024-19906-1.
25. V. S. Katamneni, A. V. Nadimpalli, and A. Rattani, "Demographic Fairness and Accountability of Audio- and Video-Based Unimodal and Bi-modal Deepfake Detectors," Face Recognition Across the Imaging Spectrum, pp. 205–231, 2024, doi: 10.1007/978-981-97-2059-0_8.
26. O. Pascu, D. Oneaţă, … H. C.-I. 2025-2025 I., and undefined 2025, "Easy, Interpretable, Effective: openSMILE for voice deepfake detection," ieeexplore.ieee.org, Accessed: Jan. 19, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10890543/
27. K. S. preprint arXiv:2601.06560 and undefined 2026, "Lightweight Resolution-Aware Audio Deepfake Detection via Cross-Scale Attention and Consistency Learning," arxiv.org, Accessed: Jan. 19, 2026. [Online]. Available: https://arxiv.org/abs/2601.06560
28. K. Zaman, I. Samiul, M. Sah, … C. D.-I., and undefined 2024, "Hybrid Transformer Architectures with Diverse Audio Features for Deepfake Speech Classification," ieeexplore.ieee.org, Accessed: Dec. 23, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10714458/
29. C. Asuai, A. P. Arinomor, C. T. Atumah, I. Festus Kowhoro, and D. E. Ogheneochuko, "Hybrid CNN-LSTM Architectures for Deepfake Audio Detection Using Mel Frequency Cepstral Coefficients and Spectogram Analysis," researchgate.net, vol. 2025, no. 3, pp. 98–109, doi: 10.11648/j.ajmcm.20251003.12.
30. T. Kanwal, R. Mahum, A. M. Alsalman, M. Sharaf, and H. Hassan, "Fake speech detection using VGGish with attention block," Springer, vol. 2024, no. 1, p. 35, Dec. 2024, doi: 10.1186/s13636-024-00348-4.
31. S. Layton, T. De Andrade, D. Olszewski, … K. W. preprint arXiv, and undefined 2024, "Every Breath You Don't Take: Deepfake Speech Detection Using Breath," arxiv.org.
32. A. Kansara, P. Kumari, B. P.-I. C. on Intelligent, and undefined 2024, "Detecting Deepfake Voices Using a Novel Method for Authenticity Verification in Voice-Based Communication," Springer.
33. Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, … J. H. preprint arXiv, and undefined 2024, "SVDD Challenge 2024: A Singing Voice Deepfake Detection Challenge Evaluation Plan," arxiv.org.
34. https://datashare.ed.ac.uk/handle/10283/3336?utm_source=chatgpt.com
35. https://zenodo.org/records/4835108?utm_source=chatgpt.com