Scuola IMT Alti Studi Lucca

Towards reliable audio deepfake attribution and model recognition: a multi-level autoencoder-based framework

Questa è la versione sottoposta a revisione paritaria (postprint) della seguente opera:

Original
Towards reliable audio deepfake attribution and model recognition: a multi-level autoencoder-based framework / Di Pierno, A., Guarnera, L., Allegra, D., Battiato, S.. - (2025), pp. 101-109. (MM '25: The 33rd ACM International Conference on Multimedia Dublin, Ireland October 27 - 28, 2025) [10.1145/3746265.3759668].

Availability:
This version is available at: 20.500.11771/35718

Publisher:
Association for Computing Machinery

Published
DOI:10.1145/3746265.3759668

Terms of use:

This publication is made accessible in accordance with the terms for deposit in the institutional repository, as defined by the IMT School for Advanced Studies Lucca's Open Access Policy. (https://library.imtlucca.it/sites/default/files/regolamento-policy-open-access-imtlib_0.pdf).

Si prega di consultare le pagine informative dell'editore relative alle politiche di autoarchiviazione.

(Article begins on next page)

05 October 2026

Towards Reliable Audio Deepfake Attribution and Model Recognition: A Multi-Level Autoencoder-Based Framework

Andrea Di Pierno
IMT School of Advanced Studies
Lucca, Tuscany, Italy
andrea.dipierno@phd.unict.it

Luca Guarnera
University of Catania
Department of Mathematics and Computer Science
Catania, Sicily, Italy
luca.guarnera@unict.it

Dario Allegra
University of Catania
Department of Mathematics and Computer Science
Catania, Sicily, Italy
dario.allegra@unict.it

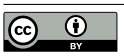
Sebastiano Battiato
University of Catania
Department of Mathematics and Computer Science
Catania, Sicily, Italy
sebastiano.battiato@unict.it

Abstract

The proliferation of audio deepfakes poses a growing threat to trust in digital communications. While detection methods have advanced, attributing audio deepfakes to their source models remains an underexplored yet crucial challenge. In this paper we introduce LAVA (Layered Architecture for Voice Attribution), a hierarchical framework for audio deepfake detection and model recognition that leverages attention-enhanced latent representations extracted by a convolutional autoencoder trained solely on fake audio. Two specialized classifiers operate on these features: *Audio Deepfake Attribution* (ADA), which identifies the generation technology, and *Audio Deepfake Model Recognition* (ADMR), which recognize the specific generative model instance. To improve robustness under open-set conditions, we incorporate confidence-based rejection thresholds. Experiments on ASVspoof2021, FakeOrReal, and CodecFake show strong performance: the ADA classifier achieves F1-scores over 95% across all datasets, and the ADMR module reaches 96.31% macro F1 across six classes. Additional tests on unseen attacks from ASVpoof2019 LA and error propagation analysis confirm LAVA's robustness and reliability. The framework advances the field by introducing a supervised approach to deepfake attribution and model recognition under open-set conditions, validated on public benchmarks and accompanied by publicly released models and code. Models and code are available at https://github.com/adipiz99/LAVA-framework.

CCS Concepts

• **Applied computing** → **Computer forensics**; • **Security and privacy** → *Domain-specific security and privacy architectures*; • **Computing methodologies** → *Neural networks*.



This work is licensed under a Creative Commons Attribution 4.0 International License.
DFF '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2047-5/2025/10
https://doi.org/10.1145/3746265.3759668

Keywords

Audio deepfakes; Deepfake attribution; Model recognition; Open-set recognition; Neural networks; Autoencoders; Attention mechanisms; Digital forensics; Synthetic speech; Deepfake detection

ACM Reference Format:
Andrea Di Pierno, Luca Guarnera, Dario Allegra, and Sebastiano Battiato. 2025. Towards Reliable Audio Deepfake Attribution and Model Recognition: A Multi-Level Autoencoder-Based Framework. In *Proceedings of the 1st Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media (DFF '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746265.3759668

1 Introduction

The rise of synthetic media generation techniques, particularly those based on deep learning, has led to the widespread emergence of *deepfakes*, manipulated audio, video, or images that convincingly imitate real individuals [1, 21]. Among these, audio deepfakes have attracted increasing attention due to their potential to impersonate voices in high-stakes contexts such as voice authentication, political communication, or disinformation campaigns. For instance, fraudsters once used AI-generated speech to impersonate a company executive and steal $35 million from a bank[1]. In another case, a fake voice of President Biden was used in a robocall to mislead voters ahead of the New Hampshire primaries[2].

While audio synthesis offers valuable benefits in fields such as accessibility, entertainment, and human-computer interaction, it also introduces serious risks to security and public trust. In particular, the proliferation of audio deepfakes fosters a growing form of *impostor bias* [2], in which the authenticity of genuine audio is increasingly questioned. This erosion of trust impacts critical domains including journalism, legal evidence, and personal communication. While recent studies have begun to explore audio deepfake attribution [12, 15, 24, 31], they often tackle the problem from alternative perspectives, such as attacker identification, pipeline inference,

[1] https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/, Last accessed: 21 June 2025
[2] https://www.reuters.com/world/us/fake-biden-robo-call-tells-new-hampshire-voters-stay-home-2024-01-22/, Last accessed: 20 June 2025

or unsupervised clustering, typically in closed-set conditions. Attribution plays a fundamental role in digital forensics, enabling investigators to infer the underlying technology or model family used to produce a fake. However, the task is particularly challenging due to the diversity and constant evolution of generation methods [7], and the subtle nature of the artifacts they leave behind.
Since binary detection of audio deepfakes (real vs. deepfake) has already been extensively studied in the literature [33], in this paper we shift our focus toward the attribution of synthetic content. In particular, we propose a multi-level architecture for audio deepfake attribution, called LAVA (Layered Architecture for Voice Attribution). Specifically, LAVA is built upon a deep convolutional autoencoder trained exclusively on fake audio, which extracts compact latent representations reused across two task-specific classifiers:

- **Level 1: Audio Deepfake Attribution (ADA)**: Given an audio deepfake sample A_i, the i-th input to be analyzed, the objective is to attribute it to its source manipulation technology by selecting among known generation methods such as *ASVspoof2021*, *FakeOrReal*, or *CodecFake*.
- **Level 2: Audio Deepfake Model Recognition (ADMR)**: identifies the specific generator model, the same task addressed by Guarnera et al. [8, 9] in the deepfake image domain for model attribution. In our framework, this level is activated only when the first-level classifier (ADA) attributes the sample to the CodecFake dataset, as it is the only dataset that includes labeled generator classes. The input audio A_i is then processed by a dedicated classifier that assigns it to one of the six known codec classes.

In detail, we distinguish between:

- **Deepfake Attribution**: the task of assigning a fake audio sample to a known generation method or synthesis pipeline (e.g., a dataset or manipulation technology);
- **Deepfake Model Recognition**: the task of identifying the specific generator model, defined by its architecture and parameters, responsible for synthesizing the audio, among a known set of alternatives.

Both classifiers share the same encoder backbone and include an attention module to reweight salient features in the latent space. The system incorporates a confidence-based rejection threshold to abstain from uncertain classifications, thus improving robustness under open-set conditions. To rigorously evaluate our architecture, we perform experiments on three publicly available datasets, *ASVspoof2021* [13], *FakeOrReal* [22], and *CodecFake* [28]. We measure classification performance using standard metrics, such as accuracy and F1-score, and conduct detailed ablation studies on the attention modules. We also introduce two complementary tests:

- An **error propagation analysis**, which quantifies how misclassifications at the ADA Level affect downstream ADMR decisions;
- A **generalization test**, evaluating both classifiers on synthetic audio from ASVspoof2019 LA [26], a dataset not seen during training but semantically close to ASVspoof2021.

Finally, we compare our method with recent state-of-the-art approaches in ADMR tasks, and show that LAVA achieves competitive or superior results across multiple settings.
Our main contributions are as follows:

- LAVA, a modular multi-level architecture for audio deepfake attribution, leveraging an autoencoder trained solely on fake audio.
- A framework based on two levels: Audio Deepfake Attribution and Audio Deepfake Model Recognition.
- An attention mechanism, integrated into each classifier, proved to be effective through controlled ablation studies.
- A rejection strategy based on confidence thresholds, enabling the system to reject out-of-distribution inputs.

The remainder of the paper is organized as follows. Section 2 reviews related work on audio deepfake detection and attribution. Section 3 presents our proposed architecture and training strategy. Section 4 outlines datasets and experimental settings. Section 5 reports empirical results, including ablations and robustness tests. Section 6 provides an in-depth analysis of the results, examines comparisons with prior work, and highlights the strengths of the proposed architecture, particularly its hierarchical design, attention-based encoding, and robustness to open-set conditions, while also discussing potential limitations. Section 7 concludes the paper and outlines future directions.

2 Related works

2.1 Audio Deepfake Detection

Recent years have seen a growing interest in detecting synthetic audio, driven by the increasing realism of text-to-speech (TTS) and voice conversion (VC) systems. A common strategy involves converting waveforms into time-frequency representations such as Mel-Frequency Cepstral Coefficients (MFCCs) or Constant-Q Cepstral Coefficients (CQCCs), which are then used as input to either convolutional [25] or dense [10] neural architectures for binary classification. End-to-end models such as RawNet [11] and x-vector-based systems [4] have demonstrated strong performance on raw inputs. However, these methods typically focus on binary classification (real vs. fake), and struggle to generalize across unknown synthesis methods [14]. Recent studies [16, 33] have also highlighted generalization as a major open challenge, especially in cross-dataset or open-set scenarios. To address this, some works have begun to explore self-supervised features [20, 24], showing promising results for more robust detection and transferability.

2.2 Attribution in Multimedia Forensics

Attribution is well-established in image forensics, where techniques identify source devices or editing tools [3, 17]. In audio, fewer works address generator attribution. CodecFake [28] introduces a benchmark for codec-based manipulation detection. Klein et al. [12] explore the classification of spoofing system components through both end-to-end and two-stage learning strategies. Müller et al. [15] evaluated various models for audio deepfake attribution but observed rapid performance degradation in open-world setups. Yan et al. [30] introduced the ADA dataset for audio deepfake attribution and proposed the CRML method, which enhances representation separation for open-set classification by leveraging multi-center learning. More recent proposals include the STOPA dataset [7],

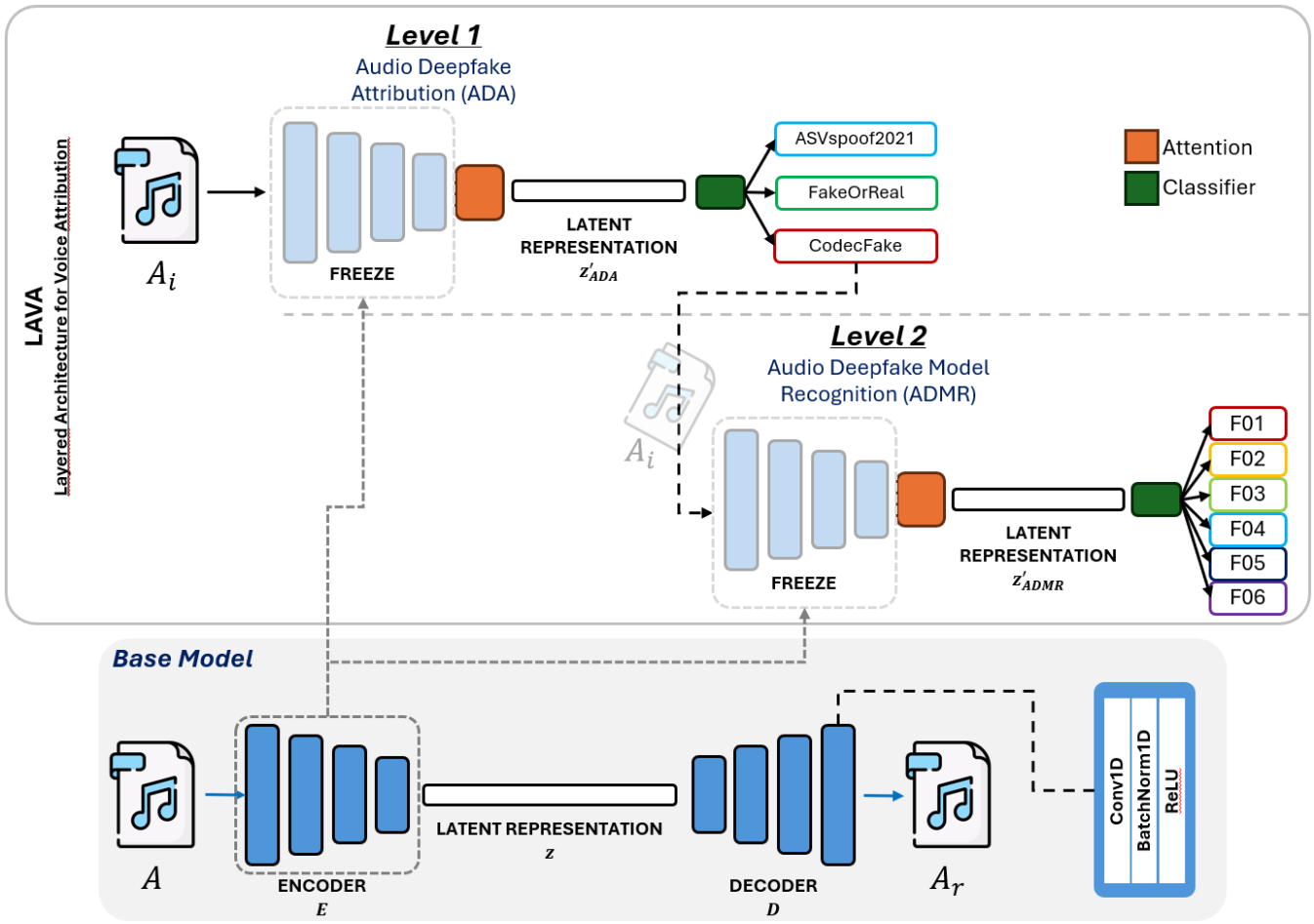


Figure 1: Overview of the LAVA framework. At the bottom, the *base model* is a deep convolutional autoencoder trained to reconstruct fake audio inputs A by minimizing the discrepancy between the original waveform A and its reconstruction A_r, using a smoothed L1 loss. Once trained, the decoder is discarded and the encoder E is reused as a frozen backbone for all subsequent classification tasks. At the top, an input audio sample A_i is processed by the encoder to obtain a latent representation z, which is then passed through an attention module. In Level 1 (ADA), the resulting attended representation z'_{ADA} is used to classify the sample into one of three dataset categories: *ASVspoof2021 (ASV)*, *FakeOrReal (FoR)*, or *CodecFake (Codec)*. If the sample is attributed to *CodecFake* and the classifier confidence exceeds a predefined rejection threshold, the sample is forwarded to Level 2 (ADMR). Here, the same encoder and attention module are reused to produce a second attended representation z'_{ADMR}, which is then classified into one of six codec-specific classes (*F01–F06*). To build a robust attribution model, a threshold strategy was applied at each level of decision making: whenever the confidence associated with a prediction drops below a predefined threshold (different for each level), the corresponding sample is discarded and marked as "unknown".

designed to benchmark source tracing under systematic generation variation, and TADA [24], a training–free method leveraging SSL embeddings and k-NN clustering for generator grouping. Wang et al. [27] proposed attribution enhancement strategies to amplify synthesis-specific traces, and Negroni et al. [18] reformulated attribution as a source verification problem using similarity learning. Neri et al. [19] propose a dual-branch CNN that processes MFCC and GTCC features in parallel to attribute synthetic speech to its generation algorithm, achieving high accuracy on a closed-set dataset from the IEEE Signal Processing Cup.

2.3 Our Contribution

Differently from prior work, we propose a unified multi-level architecture tailored to deepfake audio attribution. It employs a shared autoencoder trained solely on fake samples to encode latent representations, followed by specialized classifiers for audio deepfake attribution (Level 1) and audio deepfake model recognition (Level 2). Attention mechanisms further refine these embeddings. To our knowledge, this is the first attempt to structure audio deepfake attribution as a hierarchical multi-task pipeline with built-in interpretability.

3 Proposed approach

In this section, we describe the multi-level architecture of our proposed framework for audio deepfake attribution and model recognition, and the datasets used for training and evaluation. We also discuss the rationale behind the ablation studies and our evaluation protocol. Figure 1 illustrates an overview of the proposed approach.

3.1 Datasets

We use three publicly available datasets in our experiments:

- **CodecFake** [28]: A synthetic dataset composed of audio generated through six different speech codecs. It includes both real and fake utterances and is primarily used for fine-grained model-level attribution. The six fake classes differ in their compression strategies, architecture complexity, and training paradigms. For instance, *SoundStream (F01)* [34] and *EnCodec (F04)* [6] are real-time neural codecs with transformer-based bottlenecks, while *FuncCodec (F03)* [5] and *AcademicCodec (F06)* [32] represent lightweight or academic baselines. *SpeechToknizer (F02)* [35] focuses on token-based speech modeling, and *AudioDec (F05)* [29] employs diffusion-based reconstruction. These differences result in diverse signal characteristics and artifact patterns, making CodecFake suitable for evaluating model-level attribution capabilities.
- **ASVspoof2021** [13]: A benchmark dataset for spoofing detection containing both bonafide and spoofed utterances, generated using a variety of synthesis techniques.
- **FakeOrReal (FoR)** [22]: A curated dataset designed for training and evaluating deepfake detection and attribution systems, containing real and synthetic audio segments.

All audio samples are converted to mono and resampled at 16 kHz to ensure uniformity across datasets. This sampling rate balances perceptual quality with computational efficiency and is commonly adopted in speech processing literature [23]. Waveforms are normalized by their peak absolute amplitude and trimmed or zero-padded to a fixed length of 3 seconds (i.e., 48,000 samples).

As regards the first level (ADA) we use 75,000 fake samples evenly drawn from the three datasets, as shown in Table 1.

Table 1: Distribution of samples per dataset across training, validation, and test splits for ADA.

Split	CodecFake	ASVspoof2021	FakeOrReal	Total
Training	15,000	15,000	15,000	**45,000**
Validation	5,000	5,000	5,000	**15,000**
Testing	5,000	5,000	5,000	**15,000**
Total	**25,000**	**25,000**	**25,000**	**75,000**

The second level (ADMR) is trained on 313,282 fake samples from CodecFake, distributed as specified in Table 2.

Unlike CodecFake, the ASVspoof2021 and FakeOrReal datasets do not include fine-grained labels specifying the exact generation model or codec used to synthesize each audio sample. They are organized as binary classification datasets with labels indicating only whether an utterance is real or fake. As a result, they are unsuitable for training or evaluating the ADMR classifier, which requires detailed ground truth annotations at the model level. For this reason, only CodecFake is used at Level 2 of the attribution pipeline.

Table 2: Class-wise distribution of CodecFake samples across training, validation, and test sets.

Split	F01	F02	F03	F04	F05	F06	Total
Training	31,329	31,329	31,329	31,325	31,328	31,328	**187,968**
Validation	10,443	10,443	10,443	10,442	10,443	10,443	**62,657**
Testing	10,443	10,443	10,443	10,442	10,443	10,443	**62,657**

3.2 Proposed Autoencoder

At the core of our architecture lies a convolutional autoencoder trained exclusively on fake audio samples. This design is based on the assumption that deepfakes, despite their variability, share generation-specific artifacts that can be encoded more effectively when real samples are excluded. Once training is complete, the decoder D is discarded and only the encoder E is kept and used in both ADA and ADMR modules. The encoder consists of a stack of convolutional layers, listed in Table 3, designed to progressively compress the input waveform into a latent representation z of shape $(256, T')$, where T' depends on the temporal downsampling rate. The encoder is trained by minimizing a Smoothed L1 Loss between the input x and its reconstruction $\hat{x}$:

$$\mathcal{L}_{\text{smooth-L1}}(x, \hat{x}) = \begin{cases} 0.5 \cdot (x - \hat{x})^2 / \beta, & \text{if } |x - \hat{x}| < \beta \\ |x - \hat{x}| - 0.5 \cdot \beta, & \text{otherwise} \end{cases} \tag{1}$$

where β is a threshold hyperparameter. We use $\beta = 0.0001$ to emphasize small reconstruction errors while maintaining robustness to outliers.

Table 3: Architecture of the encoder stack used to generate latent representations.

Layer	Type	In Channels	Out Channels	Kernel Size	Stride / Padding
1	Conv1D	1	32	9	2 / 4
2	BatchNorm1D	-	32	-	-
3	ReLU	-	-	-	-
4	Conv1D	32	64	9	2 / 4
5	BatchNorm1D	-	64	-	-
6	ReLU	-	-	-	-
7	Conv1D	64	128	9	2 / 4
8	BatchNorm1D	-	128	-	-
9	ReLU	-	-	-	-
10	Conv1D	128	256	9	2 / 4
11	BatchNorm1D	-	256	-	-
12	ReLU	-	-	-	-

After training, all encoder weights are frozen except for the final convolutional layer, which remains trainable to support task-specific adaptation. The encoder thus acts as a shared feature extractor for both levels of the LAVA framework.

To enhance feature saliency, an attention mechanism is applied to the latent representation z, defined as:

$$z' = z \odot \sigma(\text{Conv1D}(z)) \tag{2}$$

where σ is the sigmoid activation, $\odot$ denotes element-wise multiplication and z' (called z'_{ADA} for Level 1 and z'_{ADMR} for Level 2) represents the reweighted latent representation obtained by modulating each channel of z according to its learned relevance through the attention mechanism, as shown in Table 4. The latent representation z'_{ADA} or z'_{ADMR} is then passed through a classifier composed of adaptive average pooling, two fully connected layers with ReLU activation, and a final linear layer that outputs logits for either 3 (ADA) or 6 (ADMR) classes.

Table 4: Architecture of attention mechanism and classification head.

Component	Layer	Details
Attention	Conv1D	In: 256, Out: 256, Kernel Size: 1
	Sigmoid	Element-wise activation over channels
Classifier	AdaptiveAvgPool1D	Output shape: (256, 1)
	Flatten	Output shape: (256)
	Linear + ReLU	In: 256, Out: 128
	Linear (Output)	In: 128, Out: 3 (ADA) or 6 (ADMR)

This modular design enables encoder reuse while maintaining classifier independence. Additionally, the shared latent space fosters consistency across attribution levels and enables clearer task separation, which facilitates analysis and debugging of individual components.

3.3 Level 1 - ADA

The aim of the first level (ADA) is to determine the synthesis technology used to generate a given fake audio sample A_i. Specifically, the frozen encoder E (pretrained as part of the base autoencoder and fine-tuned only in its final convolutional layer) processes A_i to produce a latent representation z. This representation is refined through an attention mechanism to yield z'_{ADA}. The latent representation z'_{ADA} is then classified into one of three dataset categories: *ASVspoof2021 (ASV)*, *FakeOrReal (FoR)*, or *CodecFake (Codec)*. This stage serves as the entry point to the LAVA pipeline. A confidence-based rejection mechanism (Section 3.6) is applied to the softmax output: if the maximum confidence score is below the threshold τ_{ADA}, the sample is rejected and labeled as unknown.

3.4 Level 2 - ADMR

In the second stage (ADMR) attribution proceeds only if the output of the ADA classifier corresponds to the *CodecFake* class and the associated confidence score exceeds the threshold τ_{ADA}. The same encoder E and attention mechanism are reused to extract an attended latent representation z'_{ADMR} from the original input A_i. This refined embedding is then processed by a second classifier to attribute the sample to one of six codec-specific generation classes (*F01–F06*). A second rejection threshold τ_{ADMR} is applied at this level: if the classifier's confidence is below this threshold, the sample is rejected and labeled as "unknown". This mechanism serves to limit the propagation of erroneous predictions from Level 1 and prevent misclassification of samples that, although routed to ADMR, deviate from known latent patterns. Together, the modularity of E, the attention refinement, and the hierarchical rejection thresholds contribute to the pipeline's robustness in open-set attribution scenarios.

3.5 Ablation Settings

To evaluate the impact of the attention mechanism, we train and test each classifier both with and without attention. The same training setup is used for all configurations: 50 epochs, batch size of 16, and early stopping based on validation loss.

3.6 Rejection Threshold

To improve the system's robustness and its generalization capabilities, we adopt a rejection mechanism based on confidence scores. For each classifier, we compute a rejection threshold during training, defined as the minimum confidence score required for a prediction to be accepted. Specifically, this threshold is not arbitrarily fixed but is derived from the distribution of softmax confidence scores observed on the training set. For each training sample, we record the softmax confidence associated with its predicted class. These values are then sorted in descending order, and the threshold is set at the percentile that ensures at least 85% classification accuracy on the training data. This ensures that only predictions made with sufficient confidence are accepted. At test time, predictions with a confidence score above this threshold are accepted as valid class predictions. Conversely, if the predicted class confidence falls below the threshold, the sample is rejected and assigned to an "unknown" class. This implies that the input is considered inconsistent with any of the known classes seen during training. The rejection mechanism is applied independently in both the ADA and ADMR classifiers and plays a crucial role in enabling open-set attribution and limiting error propagation across stages in the hierarchical architecture.

4 Experimental Setup

In this section, we describe the training protocol, evaluation metrics, and implementation details used in our experiments. We also outline the baseline setup for the ablation studies introduced in Section 3.

4.1 Autoencoder Pretraining

The shared encoder E, used in all classifiers, is derived from a deep convolutional autoencoder trained exclusively on fake audio samples from the CodecFake dataset. The dataset comprises 313,282 samples, evenly distributed across six codec classes. We employed a training/validation split of 80/20, resulting in 250,625 samples for training and 62,657 for validation. While the pretraining samples are drawn from the same dataset later used in the ADMR task, the autoencoder is trained solely for reconstruction without using generator labels. This ensures that representation learning remains disentangled from the downstream classification objectives. The autoencoder was trained for a maximum of 50 epochs using the Adam optimizer with a learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, and batch size of 16. Early stopping was based on validation loss. The reconstruction objective is a Smoothed L1 Loss (Eq. 1) with $\beta = 0.0001$, which balances robustness to outliers with sensitivity to small deviations. The best model was selected at epoch 44, achieving a training loss of 0.0074 and a validation loss of 0.0018 (Figure 2).

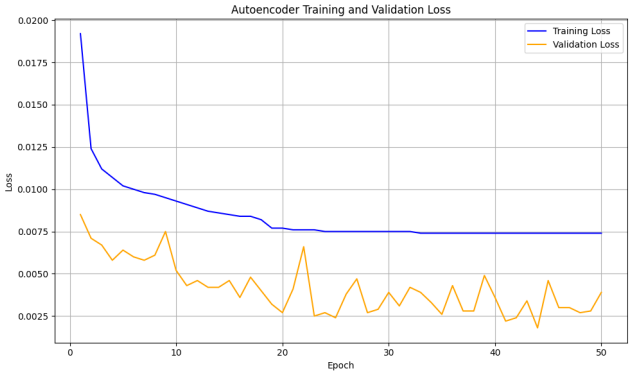


Figure 2: Training and Evaluation Loss of the autoencoder

These values confirm that the encoder learned a compact and high-fidelity representation of fake audio samples, capturing generation-specific artifacts while filtering out irrelevant variance. The decoder D is discarded after training, and the encoder is used as a frozen backbone in all downstream classifiers (with the exception of its final convolutional layer, which remains trainable).

4.2 Training Procedure

All classifiers are trained independently using the Adam optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. We adopt a batch size of 16 and a maximum of 50 training epochs with early stopping based on validation loss to prevent overfitting. The encoder weights are frozen during training, except for the final convolutional layer, which remains trainable to allow mild task-specific adaptation. All experiments are performed on a workstation equipped with an NVIDIA RTX A6000 GPU (48GB VRAM) and a 32-core AMD Ryzen Threadripper PRO 3975WX CPU. Datasets are preprocessed and stored as normalized waveforms of fixed length (3 seconds, 16 kHz, mono), as detailed in Section 3.

4.3 Evaluation Protocol

Each classifier is evaluated on the dedicated test split described in Section 3. We report standard classification metrics including Accuracy, Precision, Recall, and F1-score, both per class and as macro averages in Section 5.

To assess the reliability and robustness of our attribution framework, we report not only standard classification results but also two additional evaluations: an error propagation test and a generalization test. The former simulates the full inference pipeline to quantify how misclassifications in the initial attribution stage (ADA) affect downstream model recognition (ADMR). The latter evaluates the system's capacity to handle unseen data. These analysis are reported in Section 5.

4.4 Ablation Protocol

To isolate the contribution of the attention mechanism, we train and evaluate versions of each classifier with the attention module removed. All other components and hyperparameters remain unchanged. Performance is compared against the full-attention variants to quantify the impact of feature reweighting on attribution accuracy.

5 Results and Discussion

5.1 Evaluation Metrics

We evaluate our models using standard classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive overview of classification performance across both balanced and imbalanced class distributions.

5.2 Experimental Results

As shown in Tables 5 and 6, our models achieve strong performance across all attribution levels when using the attention mechanism. The ADA module achieves an overall accuracy of 96.21%. Most notably, the ADMR model reaches a macro-average F1-score of 96.31% across six codec classes, demonstrating the architecture's effectiveness for fine-grained attribution. These results confirm that the integration of attention modules, as demonstrated in Section 5.5, helps the model focus on salient latent features, improving class separability in the encoded space.

5.3 Error Propagation Test

We simulate the full inference pipeline by feeding each sample through the ADA classifier and forwarding it to the ADMR classifier only when the ADA prediction corresponds to *CodecFake*. This setup reflects real-world deployment, where upstream errors affect downstream attribution. The evaluation was conducted on 21,000 samples, including 15,000 fake (5,000 per dataset's test set) and 6,000 randomly selected real audio (2,000 per dataset). Real samples were never seen during training and serve as hard negatives to simulate out-of-distribution inputs. The ADA classifier achieves a 26.82% error rate demonstrating effective rejection of anomalous inputs. Among the 5,386 samples classified as *CodecFake*, the ADMR classifier introduces additional errors, with a 35.37% misclassification rate. These results highlight the impact of early-stage decisions in the LAVA pipeline and confirm the importance of robust attribution at both levels.

Table 5: Audio Deepfake Attribution (ADA) results

Dataset	Precision	Recall	F1-Score
CodecFake	0.9749	0.9568	0.9658
ASVspoof2021	0.9402	0.9720	0.9558
FakeOrReal	0.9724	0.9576	0.9649
Accuracy		0.9621	
Macro Avg	0.9625	0.9621	0.9622
Weighted Avg	0.9625	0.9621	0.9622

Table 6: Audio Deepfake Model Recognition (ADMR) results

Class	Precision	Recall	F1-Score
F01	0.9975	0.9980	0.9978
F02	0.9016	0.9325	0.9168
F03	0.9921	0.9959	0.9940
F04	0.9789	0.9880	0.9835
F05	0.9778	0.9741	0.9760
F06	0.9319	0.8907	0.9108
Accuracy		0.9632	
Macro Avg	0.9633	0.9632	0.9631
Weighted Avg	0.9633	0.9632	0.9631

Table 7: Performance comparison with and without attention for ADA and ADMR tasks

Level	Model	Precision	Recall	F1-Score	Accuracy
L1 ADA	With Attention	0.9625	0.9621	0.9622	0.9621
	Without Attention	0.9104	0.9079	0.9082	0.9079
L2 ADMR	With Attention	0.9633	0.9632	0.9631	0.9632
	Without Attention	0.8412	0.8256	0.8172	0.8256

5.4 Generalization Test

To evaluate the architecture's capacity for generalization, we tested both classifiers on 20,000 synthetic samples from ASVspoof2019 LA [26], a dataset not used during training, but semantically close to ASVspoof2021, although it includes different spoofing techniques, codec chains, and non-overlapping speaker identities. In the ADA task, 28.82% of the samples were correctly rejected as unknown. Most of the remaining samples were attributed to ASVspoof2021 (64.18%), a behavior consistent with the similarity between the two datasets. The rejection mechanism proved effective in isolating anomalous inputs whose confidence scores did not match any known class distribution. In the ADMR task, the model achieved 81.28% accuracy, demonstrating strong rejection capabilities even under open-set conditions. This result suggests that the latent space learned by the autoencoder, combined with attention-based reweighting, enables robust handling of unseen synthesis techniques that share latent similarities with known codecs.

5.5 Ablation Studies

To assess the importance of attention mechanisms, we trained variants of both classifiers with the attention modules removed. As shown in Table 7, performance consistently dropped across all tasks. The impact was especially severe in the ADMR classifier (Table 7), whose accuracy fell from 96.32% to 82.56%. A closer inspection of class-level metrics (Table 8) reveals significant performance degradation, particularly for class F06, whose recall dropped to 42.63%. This underscores the value of attention for fine-grained attribution, where subtle feature differences must be preserved and leveraged for reliable classification.

Table 8: ADMR results *without attention*.

Class	Precision	Recall	F1-Score
F01	0.9605	0.9497	0.9551
F02	0.6372	0.8926	0.7436
F03	0.8687	0.8962	0.8823
F04	0.8389	0.8973	0.8671
F05	0.8771	0.8914	0.8842
F06	0.8650	0.4263	0.5711
Accuracy		0.8256	
Macro Avg	0.8412	0.8256	0.8172
Weighted Avg	0.8412	0.8256	0.8172

6 Discussion

The proposed LAVA architecture demonstrates strong attribution performance across both coarse-grained (technology-level) and fine-grained (model-level) tasks. The integration of attention mechanisms proves consistently beneficial, particularly in the more challenging ADMR task, where subtle generator-specific artifacts must be isolated in a shared latent space. The error propagation analysis reveals the importance of accurate predictions at early stages of the pipeline: misclassifications in ADA significantly affect ADMR outcomes, validating the hierarchical structure's sensitivity to upstream decisions. Our generalization test on ASVspoof2019 LA confirms that the model can extrapolate beyond its training data. In the ADA stage, the model correctly rejects 28.82% of the samples as unknown and assigns 64.18% to ASVspoof2021, behavior consistent with the distributional similarity between the two datasets. In the ADMR task, the system achieves 81.28% accuracy, despite the test set being entirely unseen during training. This highlights the effectiveness of the rejection mechanism in filtering anomalous inputs and the model's capacity to generalize under open-set conditions.

6.1 Discussion on Prior Works

While several recent works have addressed tasks related to audio deepfake attribution, none of them perform the same two-level, supervised attribution that LAVA targets. Existing approaches typically focus on either clustering-based identification, attacker recognition, or vocoder tracing, often in closed-set or in-domain conditions. As such, a direct, level-by-level comparison is not possible. Nonetheless, we provide an overview of representative methods that address similar goals from different perspectives. Müller et al. [15] propose attacker-level neural embeddings trained on ASVspoof2019 and report high accuracy (97.10%) in a closed-set speaker identification task. Although they explore clustering in an out-of-domain setting by holding out some identities, their method is not designed for generator attribution and lacks any rejection mechanism, key features in forensic attribution tasks. Klein et al. source tracing system [12] focuses on reconstructing generation pipelines using acoustic and vocoder model inference. Their goal is to determine the transformation chain behind a spoofed signal,

rather than attributing it to a known model. Their reported 84.6% accuracy refers to closed-set vocoder classification in a highly controlled environment, without open-set evaluation or supervised class-level attribution. Recent methods have also begun exploring open-set scenarios. The TADA framework [24] uses self-supervised embeddings and k-NN to cluster audio samples based on generator identity. However, unlike LAVA, TADA does not rely on predefined class labels and does not perform supervised attribution; rather, it attempts to associate each sample with a latent model identity. Its unsupervised clustering nature makes it fundamentally different and not directly comparable. Another recent study, ReTA [31], introduces a strategy for rejection threshold adaptation in open-set deepfake attribution. While the results on SFR and DFAD datasets are promising, the lack of code, use of uncommon benchmarks, and reliance on static ResNet features without encoder-decoder or attention mechanisms limit comparability with LAVA. In summary, LAVA introduces a unified and structured attribution pipeline that:

- supports both in-domain and out-of-domain evaluation;
- integrates a confidence-based rejection mechanism for open-set robustness;
- enables both technology-level (ADA) and model-level (ADMR) supervised attribution with strong generalization to previously unseen data.

This dual-level attribution design is especially important in forensic contexts, where investigators must not only detect synthetic content but also trace its exact origin in terms of the underlying synthesis technology and architecture. These distinctions position LAVA as a reliable and scalable tool for forensic analysis of audio deepfakes in real-world conditions.

7 Conclusions

In this work, we introduced LAVA, a novel multi-level framework for audio deepfake attribution and model recognition grounded in a shared convolutional autoencoder trained exclusively on synthetic audio. The architecture supports both technology-level attribution (ADA) and fine-grained model recognition (ADMR), and leverages attention mechanisms to enhance performance, while the modular design improves transparency and task specialization.. A key feature of LAVA is its ability to operate in open-set conditions via a confidence-based rejection mechanism that prevents overconfident misclassification of unfamiliar inputs—an essential requirement for forensic deployment. Experimental results on (*CodecFake*, *FakeOrReal*, and *ASVspoof2021*) show that LAVA achieves high attribution performance across tasks, with consistent improvements brought by the attention mechanism and stable behavior under distributional shifts. Our generalization test on the unseen ASVspoof2019 LA dataset confirmed the system's robustness, with 81.28% accuracy in ADMR and a well-calibrated rejection behavior in ADA. Error propagation analysis emphasized the importance of robust upstream decisions, while ablation studies confirmed the crucial role of attention in capturing model-specific artifacts. Compared to recent approaches, LAVA provides a unique combination of supervised attribution, modularity, and rejection-aware generalization. Unlike prior work that focuses on unsupervised clustering, speaker identity, or vocoder pipeline tracing, LAVA offers direct attribution of synthetic audio to both the generation technology and the underlying model, addressing a pressing need in forensic audio analysis. Its hierarchical and modular design, with clearly defined decision stages, makes it a strong candidate for integration into real-world forensic workflows. Future research will extend the framework with additional attribution levels (e.g., family-level generalization), explore multimodal fusion with visual deepfake detectors, and investigate attribution-aware defenses for online content moderation and forensic auditing.

8 Acknowledgments

This study has been partially supported by SERICS (PE00000014), including its vertical project FF4ALL, under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU.

References

[1] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, Sara Mandelli, Gian Luca Marcialis, Marco Micheletto, Andrea Montibeller, Giulia Orrù, Alessandro Ortis, Pericle Perazzo, Giovanni Puglisi, Davide Salvi, Stefano Tubaro, Claudia Melis Tonti, Massimo Villari, and Domenico Vitulano. 2025. Deepfake Media Forensics: State of the Art and Challenges Ahead. In *Advances in Social Networks Analysis and Mining*, I-Hsien Ting, Reda Alhajj, Panagiotis Karampelas, and Min-Yuh Day (Eds.). Springer Nature Switzerland, Cham, 33–48.

[2] Mirko Casu, Luca Guarnera, Pasquale Caponnetto, and Sebastiano Battiato. 2024. GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. *Forensic Science International: Digital Investigation* 50 (2024), 301795. doi:10.1016/j.fsidi.2024.301795

[3] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2019. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. *arXiv:1812.02510* (2019). arXiv:1812.02510 [cs.CV]

[4] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*. ISCA. doi:10.21437/interspeech.2020-2650

[5] Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2023. FunCodec: A Fundamental, Reproducible and Integrable Open-source Toolkit for Neural Speech Codec. arXiv:2309.07405 [cs.SD]

[6] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High Fidelity Neural Audio Compression. arXiv:2210.13438 [eess.AS]

[7] Anton Firc, Manasi Chibber, Jagabandhu Mishra, Vishwanath Pratap Singh, Tomi Kinnunen, and Kamil Malinka. 2025. STOPA: A Database of Systematic VariaTion Of DeePfake Audio for Open-Set Source Tracing and Attribution. *arXiv:2505.19644* (2025).

[8] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2024. Mastering Deepfake Detection: A Cutting-edge Approach to Distinguish GAN and Diffusion-model Images. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 11, Article 343 (Sept. 2024), 24 pages. doi:10.1145/3652027

[9] Luca Guarnera, Oliver Giudice, Matthias Niessner, and Sebastiano Battiato. 2020. On the Exploitation of Deepfake Model Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[10] Eshika Jain and Amanveer Singh. 2024. Deepfake Voice Detection Using Convolutional Neural Networks: A Comprehensive Approach to Identifying Synthetic Audio. In *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*. 1–5. doi:10.1109/CCIS63231.2024.10931997

[11] Jee-weon Jung, Hee-Soo Kim, Hye-jin Kim, Sung-Hyun Yoon, and Ha-Jin Yu. 2019. RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification. In *Interspeech 2019*. 1268–1272.

[12] Nicholas Klein, Tianxiang Chen, Hemlata Tak, Ricardo Casal, and Elie Khoury. 2024. Source Tracing of Audio Deepfake Systems. In *Interspeech 2024*. ISCA, 1100–1104. doi:10.21437/interspeech.2024-1283

[13] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 2507–2522. doi:10.1109/taslp.2023.3285283

[14] Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize?. In *Interspeech*.

[15] Nicolas Müller, Franziska Diekmann, and Jennifer Williams. 2022. Attacker Attribution of Audio Deepfakes. In *Interspeech 2022*. 2788–2792. doi:10.21437/Interspeech.2022-129

[16] Nicolas M. Müller, Nicholas Evans, Hemlata Tak, Philip Sperl, and Konstantin Böttinger. 2024. Harder or Different? Understanding Generalization of Audio Deepfake Detection. *arXiv:2406.03512* (2024). arXiv:2406.03512 [cs.SD]

[17] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, Amit K. Roy-Chowdhury, and B. S. Manjunath. 2019. Detecting GAN generated Fake Images using Co-occurrence Matrices. *arXiv:1903.06836* (2019). arXiv:1903.06836 [cs.CV]

[18] Viola Negroni, Davide Salvi, Paolo Bestagini, and Stefano Tubaro. 2025. Source Verification for Speech Deepfakes. *arXiv:2505.14188* (2025).

[19] Michael Neri, Anna Ferrarotti, Luca De Luisa, Andrea Salimbeni, and Marco Carli. 2022. ParalMGC: Multiple Audio Representations for Synthetic Human Speech Attribution. In *2022 10th European Workshop on Visual Information Processing (EUVIP)*. 1–6. doi:10.1109/EUVIP53989.2022.9922861

[20] Orchid Chetia Phukan, Drishti Singh, Swarup Ranjan Behera, Arun Balaji Buduru, and Rajesh Sharma. 2024. Investigating Prosodic Signatures via Speech Pre-Trained Models for Audio Deepfake Source Attribution. In *arXiv preprint*. arXiv:2412.17796.

[21] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. 2024. On the Exploitation of DCT-Traces in the Generative-AI Domain. In *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3806–3812. doi:10.1109/icip51287.2024.10648013

[22] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 1–10. doi:10.1109/SPED.2019.8906599

[23] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5329–5333. doi:10.1109/ICASSP.2018.8461375

[24] Adriana Stan, David Combei, Dan Oneata, and Horia Cucu. 2025. TADA: Training-free Attribution and Out-of-Domain Detection of Audio Deepfakes. *arXiv:2506.05802* (2025). arXiv:2506.05802 [eess.AS]

[25] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language* 45 (2017), 516–535. doi:10.1016/j.csl.2017.01.001

[26] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *arXiv:1911.01601* (2020). arXiv:1911.01601 [eess.AS]

[27] Zhigang Wang, Dengpan Ye, Jingyang Li, and Jiacheng Deng. 2025. Generalize Audio Deepfake Algorithm Recognition via Attribution Enhancement. In *ICASSP 2025*. 1–5. doi:10.1109/ICASSP49660.2025.10889399

[28] Haibin Wu, Yuan Tseng, and Hung yi Lee. 2024. CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems. In *Interspeech 2024*. 1770–1774. doi:10.21437/Interspeech.2024-2093

[29] Yi-Chiao Wu, Israel D. Gebru, Dejan Marković, and Alexander Richard. 2023. Audiodec: An Open-Source Streaming High-Fidelity Neural Audio Codec. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5. doi:10.1109/icassp49357.2023.10096509

[30] Xinrui Yan, Jiangyan Yi, Jianhua Tao, and Jie Chen. 2022. Audio Deepfake Attribution: An Initial Dataset and Investigation. *arXiv:2208.10489* (2022).

[31] Xinrui Yan, Jiangyan Yi, Jianhua Tao, Yujie Chen, Hao Gu, Guanjun Li, Junzuo Zhou, Yong Ren, and Tao Xu. 2024. Reject Threshold Adaptation for Open-Set Model Attribution of Deepfake Audio. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 476–480. doi:10.1109/ISCSLP63861.2024.10800741

[32] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. HiFi-Codec: Group-residual Vector quantization for High Fidelity Audio Codec. arXiv:2305.02765 [cs.SD]

[33] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio Deepfake Detection: A Survey. *arXiv:2308.14970* (2023). arXiv:2308.14970 [cs.SD]

[34] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An End-to-End Neural Audio Codec. arXiv:2107.03312 [cs.SD]

[35] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. SpeechTokenizer: Unified Speech Tokenizer for Speech Large Language Models. arXiv:2308.16692 [cs.CL]