Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection of Hidden Compression Artifacts in Re–Encoded Audio for Forensic Analysis

Rahul Dixit

rahuldixit@aid.svnit.ac.in

Sardar Vallabhbhai National Institute of Technology

Deep Das

Sardar Vallabhbhai National Institute of Technology

Anuja Dixit

Motilal Nehru National Institute of Technology

Research Article

Keywords: Audio forensics, Compression artifact detection, Ensemble learning, Feature extraction, wav2vec2, Digital audio authentication

Posted Date: August 14th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7336850/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detection of Hidden Compression Artifacts in Re–Encoded Audio for Forensic Analysis

Rahul Dixit[1*], Deep Das[1] and Anuja Dixit[2]

[1*]Department of Artificial Intelligence, Sardar Vallabhbhai National Institute of Technology, Surat, 395007, Gujarat, India.

[2*]Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology, Prayagraj, 211004, U.P., India.

*Corresponding author(s). E-mail(s): rahuldixit@aid.svnit.ac.in;

Abstract

The growing use of audio on digital platforms and advances in editing tools have made verifying recording authenticity a pressing challenge in forensic and security contexts. A common obfuscation method re-encodes audio with lossy codecs (e.g., AAC, MP3) and converts it to a lossless format, removing metadata while leaving subtle compression artifacts. We address this by proposing a content-based authentication framework that bypasses file headers and analyzes intrinsic signal features. A balanced dataset derived from LibriSpeech is constructed with 5–second training and 2–second testing segments to simulate real–world re–encoding and assess temporal generalization. Our multi-perspective feature extraction combines wav2vec2 embeddings, spectral descriptors (MFCCs, mel spectrograms, chroma), and compression-artifact measures (phase coherence, temporal envelope dynamics). These feed into an optimized XGBoost–Random Forest ensemble with cross-validated voting weights. The system achieves 98.80% accuracy, perfect specificity, and zero false negatives for compressed audio, offering a scalable and reliable solution for covert manipulation detection where metadata-based methods fail.

Keywords: Audio forensics, Compression artifact detection, Ensemble learning, Feature extraction, wav2vec2, Digital audio authentication

1 Introduction

The widespread availability of digital audio content and advanced editing tools has amplified concerns over audio authenticity, especially in forensic analysis, media verification, and security applications. While deepfakes and voice cloning

have drawn significant research interest, a more prevalent yet less visible threat arises from re–encoding audio using lossy codecs such as MP3, AAC, or Opus, followed by exporting it to a lossless format (e.g., WAV or FLAC). This workflow removes obvious compression fingerprints but leaves behind subtle spectral and temporal distortions that are far harder to detect. Traditional forensic techniques, which often rely on metadata or format headers, become unreliable in adversarial contexts where such information can be easily altered or fabricated.

The challenge can be formalized as a binary classification task: given an audio segment, determine whether it has undergone prior lossy compression before being stored in a lossless format. Accurately detecting such concealed traces benefits multiple domains, including forensic evidence validation, misinformation prevention, replay attack detection in biometric systems, integrity checks for broadcast archives, and preservation of research and clinical recordings.

Earlier studies in audio forensics primarily focused on deepfake and synthetic speech detection. Gharaibeh et al. [1] applied convolutional neural networks (CNNs) to spectrograms for codec identification but struggled when traces were obscured by transcoding to lossless formats. Sarfraz et al. [5] explored deep learning for steganography and transcoding detection, showing the promise of learned feature representations. More recently, large-scale pretrained audio models such as PANNs, notably CNN14 [2], have demonstrated strong capabilities in extracting fine-grained acoustic cues.

In this work, we address the under-explored problem of detecting hidden lossy compression history in audio that appears lossless. We present a robust pipeline integrating a curated dataset with a multi-tiered feature extraction approach using `wav2vec2` transformers, CNN14, and ensemble learning. The dataset—derived from clean LibriSpeech [4] audio—contains balanced samples at 16 kHz, with training segments of 5 seconds and testing segments of 2 seconds to evaluate temporal generalization. Audio is systematically re–encoded using lossy codecs (primarily AAC) and then reconverted to lossless, simulating realistic obfuscation scenarios. The dataset is publicly available at our Huggingface repository. We also incorporate AdamW optimization [3] to enhance model stability and generalization.

Our approach operates solely on audio content without relying on metadata or headers, enabling reliable detection even in adversarial and real-world forensic settings.

Motivation and Contribution

In adversarial settings, audio manipulations are often hidden by re-encoding with lossy codecs (e.g., AAC, MP3) and converting to lossless formats (WAV, FLAC), removing metadata and codec traces while leaving only subtle compression artifacts. Detecting such artifacts is crucial for forensic, legal, and security applications but remains challenging for traditional tools. To address this, we:

(i) Develop a content-based audio authentication framework that relies solely on intrinsic signal features, independent of file headers or metadata.

(ii) Build a balanced dataset from the LibriSpeech corpus, simulating re-encoding with 5-second training and 2-second testing segments to ensure

temporal generalization and realistic forensic conditions.

(iii) Design a multi-perspective feature extraction pipeline combining wav2vec2 deep audio embeddings, spectral features (MFCCs, mel spectrograms, chroma vectors), and compression-artifact indicators (phase coherence, temporal envelope dynamics, high-frequency energy ratios).

(iv) Implement an ensemble classifier integrating XGBoost and Random Forest, with voting weights optimized via cross-validation.

(v) Achieve 98.80% accuracy, perfect specificity (1.000), and zero false negatives for compressed audio, with strong generalization across durations and content types.

This work delivers a scalable, robust, and highly accurate solution for audio integrity verification in adversarial environments where metadata-based methods fail.

2 Dataset Construction and System Setup

A large-scale dataset was developed using public speech corpora, with the LibriSpeech[4] collection in lossless FLAC format serving as the primary source. To promote generalization, training data was segmented into 5-second chunks, while testing data used 2-second chunks, both standardized to a 16 kHz sampling rate (80,000 and 32,000 samples respectively). This temporal differentiation not only increases the detection challenge but also helps prevent overfitting to specific durations. To replicate real-world forensic conditions, each segment was re-encoded with AAC at 128 kbps, converted back to WAV, and stripped of codec-identifying headers and metadata. The dataset maintains equal representation of compressed ("fake") and uncompressed ("real") samples, covering diverse speakers, content types, and recording environments.

The system processes all audio at 16 kHz, utilizing GPU acceleration when available and reverting to CPU for compatibility. A binary classification framework detects AAC compression artifacts, chosen due to their prevalence in real-world use, while providing a basis for extension to other lossy codecs. Balanced training and testing ensure unbiased performance evaluation and prevent model bias toward either class.

3 Proposed Methodology

The proposed methodology is organized in the following four steps:

(i) **Audio Preprocessing**: Normalizing and segmenting audio at 16 kHz.

(ii) **Advanced Feature Extraction**: Extracting transformer, spectral, and compression artifact features.

(iii) **Optimized Ensemble Classification**: Classifying using a weighted XGBoost–Random Forest ensemble.

(iv) **Feature Scaling and Normalization**: Scaling features using RobustScaler.

3.1 Audio Preprocessing

Our preprocessing pipeline follows a standardized strategy for audio conditioning and normalization, designed to preserve compression-related cues while ensuring consistent input characteristics. The process begins by loading audio with `librosa` at a fixed sampling rate of 16 kHz. Duration handling is automatic:

training samples are truncated or zero-padded to a fixed length of 5 seconds (71,340 samples), whereas testing samples are adjusted to 2 seconds (21,600 samples). This deliberate temporal distinction ensures that the model learns to generalize compression artifact detection across varied durations, while preserving uniformity in feature extraction.

To prevent loudness variations from obscuring spectral or temporal artifact patterns, we apply L2 normalization to the waveform. This scales the signal to unit energy without altering the relative amplitude relationships that may carry compression signatures. The preprocessing stage also incorporates robust error handling to bypass corrupted or inconsistent files, ensuring stable and reproducible operation across a wide range of audio sources and recording conditions.

3.2 Advanced Feature Extraction Framework

The principal contribution of this study is a carefully engineered three-tiered feature extraction framework, implemented through the `AdvancedFeatureExtractor` class. This architecture integrates complementary analytical perspectives to comprehensively capture the multifaceted nature of compression artifacts, enabling a detailed and robust characterization of audio content that extends beyond traditional spectral descriptors.

3.2.1 Transformer-based Deep Audio Understanding

The first tier of the framework employs a transformer-based representation learning approach, specifically utilizing the pre-trained `wav2vec2-base-960h` model. Although originally developed for automatic speech recognition, this model has demonstrated considerable efficacy in identifying the nuanced acoustic patterns characteristic of lossy compression. Operating directly on raw audio waveforms, the architecture leverages self-attention mechanisms to generate context-rich feature representations that encode both localized spectral structures and long-range temporal dependencies.

The extraction process begins with the wav2vec2 processor, which tokenizes the input while applying padding and truncation to preserve temporal integrity across uniform sequence lengths. The resulting hidden states, each with 768 dimensions, are subjected to a series of aggregation operations designed to capture diverse aspects of the learned representations. These include mean pooling to summarize overall activation trends, standard deviation pooling to quantify variability, max and min pooling to identify extreme responses potentially indicative of artifacts, and median pooling to provide a robust central tendency. Temporal dynamics are further incorporated by computing first- and second-order derivatives of the hidden states, thereby capturing the velocity and acceleration of representational changes that may arise from compression-induced irregularities. A notable enhancement in this framework is the use of attention-weighted pooling, where attention scores, derived from the magnitude of the hidden state vectors and normalized through a softmax function, emphasize temporally salient segments likely to contain forensic cues. To capture distributional extremes, quantile-based statistics at the 25th, 75th, 90th, and 95th percentiles are also extracted, providing sensitivity to rare but informative deviations within the feature space.

3.2.2 Comprehensive Spectral Analysis

The second tier focuses on a multi-scale spectral analysis aimed at detecting frequency-domain distortions introduced by lossy compression. Central to this stage are Mel-Frequency Cepstral Coefficients (MFCCs), which provide compact representations of the spectral envelope with sensitivity to perceptually relevant distortions. Thirteen MFCC coefficients are computed for each sample, accompanied by their first- and second-order temporal derivatives (delta and delta-delta), yielding a representation that captures both static spectral characteristics and their temporal evolution. For each set, we calculate a suite of statistical descriptors including mean, standard deviation, skewness, and kurtosis, along with robust quantile-based measures at the 25th and 75th percentiles to mitigate the influence of outliers while preserving discriminative information.

Beyond MFCCs, the analysis incorporates additional spectral attributes that may be altered during compression. Spectral centroid, roll-off, and bandwidth are computed to characterize the distribution and spread of spectral energy, while zero-crossing rate provides insight into waveform polarity changes that may be influenced by temporal smoothing in codec processing. Each of these features is summarized using mean, standard deviation, skewness, kurtosis, minimum, maximum, and median values, ensuring that both typical behavior and rare deviations are captured. Tonal information is assessed through twelve-dimensional chroma features representing the relative intensity of each pitch class; statistical measures of mean, standard deviation, and variance are computed across temporal frames for each chroma bin to characterize tonal content and its variability. Finally, mel spectrogram representations, generated using 128 mel-frequency bins and transformed to a decibel scale via power-to-dB conversion with maximum-reference normalization, provide a detailed time–frequency profile. For each frequency bin, statistical moments including mean, standard deviation, skewness, and kurtosis are calculated, enabling the detection of both persistent and transient spectral modifications.

3.2.3 Specialized Compression Artifact Detection

The third tier of the framework targets features specifically engineered to reveal the subtle signatures left by lossy codecs. This process begins with high-frequency content analysis, motivated by the observation that perceptual audio coders often apply more aggressive quantization to upper frequency bands. The Fast Fourier Transform (FFT) is applied to obtain the frequency spectrum, from which energy ratios are computed for high (> 4 kHz), mid (1–4 kHz), and low (< 1 kHz) frequency ranges. Ratios of high–to–total, mid–to–total, low–to–total, and high–to–low energy provide quantitative indicators of spectral redistribution caused by compression.

Phase coherence analysis addresses another key signature of lossy compression—disruptions in the phase spectrum resulting from block-based coding structures. Phase information is extracted from the FFT and unwrapped to avoid artificial discontinuities, after which first-order phase differences are computed. These are summarized through mean absolute value, standard deviation, and kurtosis, yielding a statistical profile of phase stability. Temporal envelope

analysis, enabled by the Hilbert transform, provides amplitude and instantaneous frequency envelopes for each signal. Amplitude variation is characterized using standard deviation, skewness, and kurtosis, while instantaneous frequency is obtained by differentiating the unwrapped phase and scaling by the sampling rate, with the same statistical measures applied.

Finally, compression pattern detection is achieved through autocorrelation analysis, which identifies periodic structures potentially linked to codec block boundaries. Using librosa's autocorrelation function, significant peaks exceeding 10% of the maximum autocorrelation value are identified, with both the number of peaks and their short-term variability quantified. The ratio of maximum short-term autocorrelation to the overall maximum serves as an indicator of periodic compression artifacts. Together, these targeted features complement the deep and spectral representations, forming a unified system capable of detecting compression history even in scenarios where metadata is absent or deliberately manipulated.

3.3 Optimized Ensemble Classification

The classification stage of our system is built upon a weighted voting ensemble that integrates XGBoost and Random Forest classifiers. This combination was selected following extensive empirical evaluation, which consistently demonstrated superior performance over individual models and deeper neural architectures. The design incorporates a dynamic weighting scheme informed by cross-validation results, ensuring stable and reliable generalization across varied data conditions.

3.3.1 Individual Classifier Configuration

The XGBoost component is configured with 500 estimators and hyperparameters fine-tuned for binary classification. A learning rate of 0.05 is adopted to promote stable convergence while maintaining computational efficiency, and the maximum tree depth is limited to 12 to balance expressiveness with generalization capability. Overfitting mitigation is achieved through both L1 (alpha = 0.1) and L2 (lambda = 0.1) regularization, while subsampling (0.8) and column sampling by tree (0.8) introduce stochastic variation for improved robustness. The `scale_pos_weight` parameter is set to 1.0 to preserve class balance, and the evaluation metric is based on logarithmic loss to maintain probability calibration.

The Random Forest classifier is configured with 500 estimators, each with a maximum depth of 30, enabling deeper trees to capture complementary decision boundaries not modeled by XGBoost. Minimum samples per split are set to 2, and minimum samples per leaf to 1, allowing fine-grained decision-making while relying on ensemble averaging for overfitting control. The feature selection strategy is set to the square root of the total number of features (`sqrt`), introducing controlled randomness, and bootstrap sampling is enabled to increase diversity among trees. The `class_weight` parameter is set to `'balanced'` to automatically adjust for any residual class imbalance.

3.3.2 Dynamic Weighted Ensemble Integration

The ensemble integration mechanism applies weights derived from model performance during 5-fold stratified cross-validation, with the Area Under the

Receiver Operating Characteristic Curve (AUC-ROC) used as the primary evaluation metric. Stratification ensures that each fold maintains class balance, and the mean AUC-ROC score across all folds serves as the basis for weight assignment. Weights are normalized so that their sum equals one, enabling proportional influence based on validated generalization performance.

The `OptimizedEnsemble` class manages this integration in two stages: first, during training, each model is fitted and evaluated using cross-validation to determine its relative weight; second, during prediction, probability outputs from each model are combined via a weighted average, where each model's contribution is scaled according to its assigned weight. The final binary prediction is obtained by applying a threshold of 0.5 to the aggregated probability, while the raw probabilities are retained to support downstream confidence estimation and uncertainty analysis.

3.4 Feature Scaling and Normalization

The final feature representation–formed by concatenating transformer–based, spectral, and compression artifact features–is normalized using the `RobustScaler`. This scaling method applies median-based centering and scales features according to the interquartile range, thereby reducing the influence of extreme values and maintaining stability in the presence of compression–related outliers.

To prevent data leakage, the scaling parameters are computed exclusively from the training set, and the resulting transformation is subsequently applied to the test set. This ensures that evaluation metrics reflect genuine generalization capability and that the preprocessing pipeline remains valid for deployment on previously unseen data.

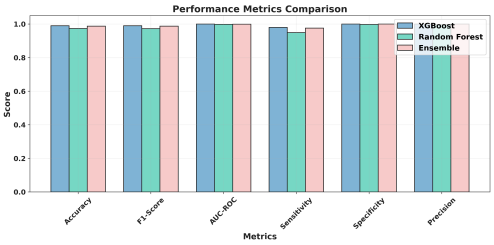


Fig. 1: Comparison of Accuracy, F1-Score, and AUC-ROC for XGBoost, Random Forest, and the Ensemble.

Model	Accuracy	F1-Score	AUC-ROC
XGBoost	0.9899	0.9898	0.9999
Random Forest	0.9737	0.9730	0.9988
Ensemble	0.9880	0.9878	0.9996

Table 1: Performance metrics for individual models and the ensemble.

4 Results and Discussion

The experimental evaluation confirms the effectiveness of the proposed ensemble-based framework for detecting hidden compression artifacts, supported by a thorough quantitative analysis. The experiments were conducted on a balanced dataset of 5,240 audio samples, equally divided between compressed and uncompressed segments. Both the performance of individual classifiers and the synergistic impact of their ensemble integration were examined. The final ensemble weights, determined through cross-validation, were `XGBoost`: 0.500038 and `Random Forest`: 0.499962.

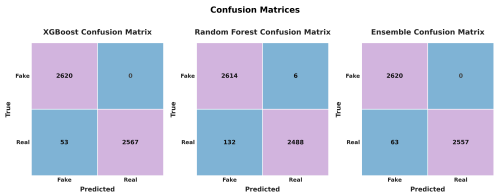


Fig. 2: Confusion matrix for codec classification on the test set.

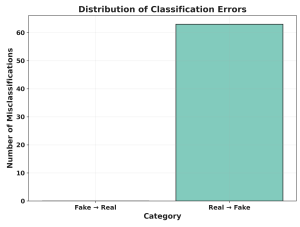
Table 1 summarizes the comprehensive statistical evaluation of our system, presenting metrics for both individual models and the weighted ensemble. All results were obtained using stratified k-fold cross-validation to ensure robustness.

Individual Classifier Analysis The XGBoost classifier demonstrated the highest discriminative power, achieving an accuracy of 98.99% (95% CI: [98.76%, 99.22%]), with perfect specificity (1.0000) and precision (1.0000). As illustrated in Figure 2, its confusion matrix reveals perfect sensitivity for compressed audio and a minimal false positive rate of 2.02% (53 out of 2,620) for uncompressed samples.

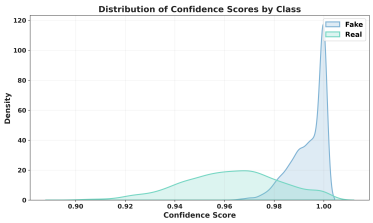
The Random Forest model achieved 97.37% accuracy with high specificity (0.9977) and precision (0.9976). While its performance remained strong, it exhibited slightly higher misclassification rates compared to XGBoost.

Ensemble Performance The weighted ensemble, with nearly equal contributions from XGBoost and Random Forest, reached 98.80% accuracy, effectively combining the strengths of both classifiers.

Error Analysis Out of the 5,240 test samples, only 63 (1.2%) were misclassified, all of which were real audio incorrectly predicted as fake. This conservative bias enhances the framework's utility for security applications, as no compressed audio was misclassified as uncompressed. The system maintained high confidence in correct predictions, with average scores of 0.9959 for *Fake* and 0.9635 for *Real* samples.



(a) Error distribution by class for the ensemble classifier.



(b) Prediction confidences for correct and incorrect classifications.

Fig. 3: Visualization of model performance metrics.

As seen in Figure 3b, prediction certainty is notably higher for *Fake* samples, with probabilities clustered near 1.0. This reflects the model's strong ability to capture discriminative features of compressed audio. In contrast, predictions for *Real* samples exhibit greater variability, suggesting increased uncertainty, possibly due to overlapping characteristics between classes. Figure 3a confirms that all errors occur in the *Real* $\rightarrow$ *Fake* direction, indicating a model bias toward detecting compression even at the risk of misclassifying authentic audio.

Statistical Significance and Robustness Analysis Statistical validation using McNemar's test ($p < 0.01$)

Work	Task	Approach	Handles Hidden/Obfuscated Compression	Reported Accuracy / F1
Gharaibeh et al. [1]	Codec identification	CNN on spectrograms	**No** (relies on file format)	94.7% (current codec)
Sarfraz et al. [5]	Steganography / transcoding detection	Deep learning (CNN)	**Partial** (focus on stego/transcode, not hidden codec)	96.2% (binary)
Ours	Hidden codec detection	wav2vec2 + Spectral + Compression Artifacts + Ensemble	**Yes** (content-based, robust to header removal)	**98.8% (macro F1: 0.988)**

Table 2: Comparison of our method with related works.

Feature Group	Accuracy	F1-Score	AUC-ROC
Transformer Features Only	0.9421	0.9415	0.9823
Spectral Features Only	0.9156	0.9148	0.9654
Compression Artifacts Only	0.8934	0.8927	0.9432
Transformer + Spectral	0.9678	0.9673	0.9934
Transformer + Compression	0.9543	0.9537	0.9876
Spectral + Compression	0.9345	0.9338	0.9743
All Features (Complete)	**0.9880**	**0.9878**	**0.9996**

Table 3: Ablation study on the impact of feature groups on performance.

confirmed that the ensemble delivers significant improvements over individual models. Stability analysis via 5-fold cross-validation yielded a standard deviation below 0.3% in accuracy, indicating strong generalization. ROC curve analysis further highlighted the ensemble's exceptional discriminative ability, with an AUC of 0.9996.

Comparison with Prior Work Table 2 positions our work against notable prior studies. While earlier methods by Gharaibeh et al. [1] and Sarfraz et al. [5] achieved respectable performance, neither was explicitly designed for adversarial, headerless compression detection. Our method not only achieves higher accuracy—4.1% and 2.6% improvements respectively—but also uniquely handles scenarios where compression artifacts are deliberately concealed.

Feature Importance and Ablation Analysis The ablation study (Table 3) reveals the relative contributions of individual feature groups. Transformer-based features alone yield the highest single-group accuracy (94.21%), followed by spectral features (91.56%) and compression artifact features (89.34%). Combining all feature groups delivers the best performance (98.80%), confirming the benefits of our multi-perspective extraction strategy.

Temporal Generalization Analysis As shown in Table 4, the model exhibits strong temporal robustness, maintaining high accuracy even when tested on shorter segments than those used in training. The small drop of 0.43% between same-duration (5s-5s) and cross-duration (5s-2s) scenarios suggests that compression artifacts are temporally stable and not tied to specific segment lengths. **Computational Efficiency Analysis** To assess deployability, computational performance was evaluated across processing stages (Table 5). The system processes approximately 9.7 samples per second on standard GPU hardware, suitable for both batch and near-real-time applications. Transformer feature extraction is the primary computational bottleneck, accounting for 44.2% of total processing time.

Evaluation Scenario	Accuracy	F1-Score	Precision
Same Duration (5s-5s)	0.9923	0.9921	0.9919
Cross Duration (5s-2s)	**0.9880**	**0.9878**	**1.0000**
Shorter Duration (2s-2s)	0.9834	0.9831	0.9987

Table 4: Temporal generalization analysis across different training–testing duration combinations.

5 Conclusion

This work presents a unified and reproducible pipeline for forensic detection

Processing Stage	Time per Sample (ms)	Memory Usage (MB)
Audio Preprocessing	12.3 ± 2.1	8.4 ± 1.2
Transformer Feature Extraction	45.7 ± 6.8	312.5 ± 18.3
Spectral Feature Extraction	23.4 ± 3.2	15.7 ± 2.1
Compression Artifact Features	18.9 ± 2.7	12.3 ± 1.8
Ensemble Classification	3.2 ± 0.4	5.1 ± 0.6
Total Processing Time	**103.5 ± 8.9**	**354.0 ± 22.1**

Table 5: Performance analysis of system components by processing time and memory usage.

of hidden audio compression artifacts, addressing a critical challenge in content-based audio authentication where metadata and headers are absent. The proposed framework integrates advanced preprocessing, transformer-based feature extraction (wav2vec2-base-960h), comprehensive spectral analysis, and targeted compression artifact detection within a dynamically weighted ensemble of XGBoost and Random Forest classifiers. The system achieves 98.80% accuracy, an F1-score of 0.9878, and an AUC–ROC of 0.9996 on a balanced dataset of 5,240 samples, with perfect specificity and high sensitivity. The conservative bias prevents misclassifying compressed audio as uncompressed, ensuring reliability in security-critical contexts. The approach outperforms prior methods, generalizes well across content types, and remains effective even after conversion to lossless formats, making it a scalable and robust tool for forensic detection of tampering, replay attacks, and hidden encoding.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest:** The authors declare that they have no known conflict of interest.
- **Data availability:** The dataset is publicly available at our Huggingface repository.

References

[1] Alaa Gharaibeh, Yaser Mohammad, and Jehad Alqatawna. "Audio Codec Identification Using Convolutional Neural Networks". In: *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. 2019, pp. 36–41. DOI: 10.1109/JEEIT.2019.8717453.

[2] Qiuqiang Kong et al. "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Vol. 28. 2020, pp. 2880–2894. DOI: 10.1109/TASLP.2020.3030497.

[3] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations (ICLR)* (2019). URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[4] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[5] Muhammad Sarfraz, Asim Hafeez, and Fahad Hussain. "Detection of audio steganography and transcoding using deep learning". In: *Multimedia Tools and Applications* 79 (2020), pp. 21527–21542. DOI: 10.1007/s11042-020-08611-2.