Classification of CD- and LP-Derived Audio Using Hand-Crafted Acoustic Features and Interpretable Machine Learning

Sandhya Rani K. M.[1], S. D. Khamitkar[1], Parag U. Balachandra[1,*]

[1] School of Computational Sciences, Swami Ramanand Teerth Marathwada University (SRTM University), Nanded, Maharashtra, India

Email: sandhya.kiran.rao@gmail.com (S. R. K. M.); s.khamitkar@gmail.com (S. D. K.); srtmun.parag@gmail.com (P. U. B.)

ORCID: Sandhya Rani K. M. 0009-0009-2027-8513; Parag U. Balachandra 0000-0001-5422-9423

*Author for correspondence: srtmun.parag@gmail.com

Abstract

Compact discs (CDs) and long-playing records (LPs) reproduce audio through different mechanisms, creating measurable differences in digitised music signals. We classified CD- and LP-derived audio using hand-crafted acoustic features and interpretable machine learning models. A private dataset comprising 17,800 non-overlapping 10-s segments from 597 songs was evaluated using Random Forest, XGBoost, support vector machine (SVM) and k-nearest-neighbour (k-NN) classifiers. Because segments from the same song are correlated, we used song-grouped cross-validation, independent validation and a held-out test set to prevent data leakage. Random Forest and XGBoost achieved 99.96% segment-level accuracy and correctly classified all 90 held-out songs. Feature importance, SHAP and ablation analyses indicated that spectral-harmonic and psychoacoustic characteristics provided the strongest discrimination. An auxiliary format-conversion analysis showed that the pre-trained XGBoost model retained the LP label for 94.13% of LP segments converted from WAV to MP3 and the CD label for 93.90% of CD segments decoded from MP3 to WAV. These findings demonstrate a strong separation between the evaluated CD- and LP-derived audio classes. However, the classes contained different songs and were represented in different formats, so musical content, encoding, mastering and recording-chain differences may have influenced the results. Stricter content-matched and codec-controlled studies are needed to isolate the effects associated with the physical medium.

Keywords: audio provenance, compact disc, vinyl record, acoustic features, Random Forest, XGBoost, SHAP, machine learning

Introduction

Compact discs (CDs) and long-playing records (LPs) follow different playback and transfer paths. CD-derived audio is retrieved digitally, whereas LP playback involves stylus-groove interaction, equalisation, amplification and analogue-to-digital conversion. These processes may produce measurable differences in spectral balance, transient behaviour, low-frequency noise and inter-channel characteristics. Although such cues are not unique to either medium, their combined behaviour can provide useful information for audio-provenance assessment, archive organisation, metadata verification and restoration planning.

Previous work on analogue-disc restoration, music information retrieval and audio forensics shows that temporal, spectral, statistical and inter-channel descriptors can capture source- and acquisition-related characteristics [1]-[21]. However, CD-LP classification is difficult to interpret because commercial releases may also differ in repertoire, mastering, equalisation, compression, encoding and playback conditions [22]-[29]. In the present dataset the two classes contain different songs and source formats; high classification accuracy must therefore be interpreted as separability of the evaluated acquisition classes rather than as proof of a universal physical distinction between CD and LP media.

This study investigates whether CD- and LP-derived recordings can be separated on previously unseen songs using 89 hand-crafted acoustic features and conventional machine learning classifiers. The main contributions are: (1) song-disjoint training, validation, cross-validation and held-out testing of Random Forest, XGBoost, RBF-SVM and k-NN; (2) interpretable analysis using tree importance, SHAP and feature-family ablation; and (3) a supplementary format-control analysis that evaluates prediction retention after LP WAV-to-MP3 conversion and CD MP3-to-WAV decoding. The analysis therefore combines predictive evaluation with explicit examination of the acoustic and acquisition-related factors that may influence the classification.

Related Work

Vinyl Artefacts and Acoustic Descriptors

The degradation of analogue audio carriers introduces distinct, quantifiable artefacts. For instance, clicks in gramophone recordings manifest as random, finite-duration corruptions ranging from subtle crackle to severe scratches [1]. Because these transient crackle peaks typically last only 0.36 to 1.09 ms and spread across multiple frequency bands [2], they are best characterised by short-term metrics such as impulse density, transient energy and spectral flux. Furthermore, localising these brief disturbances requires bidirectional processing, as standard long-term averages fail to capture their highly impulsive nature [3].

Beyond high-frequency transients, the physical recording surface and playback mechanisms frequently introduce persistent, low-frequency disturbances. Modelling these media defects as strong discontinuities followed by low-frequency pulse tails highlights the diagnostic value of rumble-related descriptors [4]. However, such low-frequency energy can easily be conflated with musical bass or mechanical vibration. To disambiguate these sources, spatial measures such as inter-channel correlation and mid-side energy offer useful supplementary cues, as click-like disturbances are often correlated across stereo channels [5]. Although a carefully preserved LP transfer may exhibit very few artefacts, evidence from learned restoration systems [6], [7] and quality assessment models [8] demonstrates that vinyl degradation has a consistent, learnable structural footprint rather than behaving as purely random noise. Even computationally efficient auditory models have proven highly capable of detecting these structured, perceptible noises [9].

Hand-Crafted Representations and Audio Provenance

Foundational music information retrieval (MIR) frameworks have long relied on extracting timbral, rhythmic and pitch-related features to characterise audio [10]. However, the success of these representations depends heavily on the chosen classification algorithm, which can significantly alter performance even when the underlying feature space remains constant [11]. To parameterise this acoustic space, digital signal processing pipelines [12] commonly extract Mel-frequency cepstral coefficients (MFCCs) to capture spectral envelopes, alongside centroid, bandwidth and roll-off metrics to describe the spectral distribution. Temporal dynamics and amplitude behaviour are tracked through spectral flux, root-mean-square (RMS) energy and dynamic range variables.

Although the field has increasingly shifted from engineered features towards deep representation learning and raw-waveform modelling [13], [14], hand-crafted representations retain distinct advantages. They may discard fine-grained local spectrogram structure, but engineered features offer computational efficiency and physical interpretability, qualities that are particularly important when training data are limited.

In tasks concerning medium provenance, this physical interpretability is paramount. High accuracy alone cannot reveal whether a model is detecting true surface noise or merely overfitting to unrelated musical content. Evaluating multiple classifiers with fundamentally distinct decision boundaries therefore provides valuable insight. By comparing the decorrelated tree aggregation of Random Forests [15], the non-linear decision boundaries of RBF support vector machines [16], the scalable gradient boosting of XGBoost [17] and the simple distance-based logic of k-nearest neighbours, researchers can isolate algorithmic behaviour. Evaluating these diverse models under identical preprocessing conditions helps to determine whether successful classification stems from a specific learner's bias or reflects genuine, broad separability within the acoustic feature space.

Audio-Forensic Provenance and Acquisition-Chain Identification

Audio forensic research provides a strong precedent for this type of classification, demonstrating that acquisition chains imprint persistent acoustic signatures on recorded media. Hardware source devices, for example, can be reliably identified using both spatiotemporal representation learning [18] and end-to-end transfer learning, even under data-constrained conditions. Furthermore, device-specific effects and environmental acoustics frequently coexist and can be disentangled through joint modelling techniques [19]. Although these forensic applications differ structurally from CD-LP identification, they establish that chain-dependent artefacts, such as frequency-response anomalies and compression traces, are inherently quantifiable.

To ensure that provenance models are robust, they must be evaluated independently of the underlying audio content. Testing under recompression, additive noise [20] and various format conversions [21] reveals the resilience of specific acoustic signatures, particularly MFCC derivatives. Notably, explainable AI techniques indicate that high-frequency log-Mel regions play a disproportionately large role in identifying source devices [22]. Because these high-frequency bands frequently encode both playback-chain characteristics and codec artefacts,

content-disjoint evaluation protocols are essential to prevent models from learning spurious, content-based correlations [23].

Leakage, Content Bias, Mastering and Domain Shift

The risk of spurious correlation is heavily exacerbated by the correlated nature of song segmentation. If segments from the same audio track appear in both training and test partitions, models can easily exploit song identity or mix characteristics, leading to artificially inflated accuracy [24]. Historical dataset audits have repeatedly exposed how replication, mislabelling and structural leakage can establish misleading performance benchmarks [25], [26]. Rigorous experimental design therefore requires all segments from an individual song to be grouped during cross-validation to eliminate direct dataset overlap.

Song grouping alone, however, is insufficient to address broader class-level confounding variables. For instance, systematic variations in dynamic range and RMS energy are closely tied to both musical genre [27] and historical mastering trends [28]. The well-documented "loudness war" clearly illustrates how evolving mastering practices alter audio profiles entirely independently of the physical carrier medium [29]. Such mastering-induced domain mismatches can severely degrade the performance of acoustic models [30], while varying spatial reproduction parameters can similarly skew classification boundaries [31].

Without strict controls, a model ostensibly intended for CD-LP classification may inadvertently learn to detect genres, release eras or mastering techniques instead of the storage medium. Synthesising these challenges yields three overarching principles for robust evaluation: first, vinyl-specific physical artefacts are objectively measurable; second, acquisition hardware can be classified using targeted representations; and third, high classification accuracy is meaningless if musical content, mastering or encoding is implicitly coupled with the target labels. Because no existing benchmark directly evaluates conventional feature-based models on strictly song-disjoint datasets of commercial CD and LP audio, the present study addresses this methodological gap while imposing strict limits on causal interpretation.

Materials and Methods

Dataset and Preprocessing

The dataset comprised commercially released music drawn from a private collection of CDs and LPs. The CD recordings were ripped and stored as 320-kbps MP3 files, whereas the LP recordings were digitised and stored as WAV files. The two collections contained different songs and were therefore neither paired nor content-matched. As the original sampling rates varied across files, all audio was standardised to 44.1 kHz during feature extraction. Silence was removed, and the complete tracks were divided into non-overlapping 10-s segments. Following preprocessing and feature-availability checks, the modelling corpus contained 17,800 segments from 597 songs: 12,404 segments from 424 CD-derived songs and 5,396 segments from 173 LP-derived songs.

To prevent direct leakage between segments from the same song, the dataset was partitioned at the song level. Song identifiers were reconstructed by removing the terminal segment suffix

from each filename. A stratified song-grouped split assigned approximately 70% of the songs to training and 15% each to validation and testing. All segments from a given song were retained in the same partition. The training partition contained 12,501 segments from 417 songs, the validation partition contained 2,728 segments from 90 songs, and the held-out test partition contained 2,571 segments from 90 songs. Set-intersection checks confirmed that no song identifier occurred in more than one partition.

Hand-Crafted Acoustic Features

Feature extraction was implemented in Python using Librosa and SciPy. Ninety candidate variables were calculated for each segment. They covered MFCC statistics, spectral centroid, bandwidth, roll-off, entropy, harmonic energy, frequency-band energy, noise floor, spectral flatness, high-frequency noise, low-frequency rumble, impulse density, signal-to-noise-related quantities, RMS, peaks, dynamic range, crest factor, compression and clipping measures, onset, transient, spectral flux, percussive energy, rhythmic descriptors, stereo correlation, imbalance, phase, mid-side ratio, spatial width, crosstalk estimates and psychoacoustic proxies, including loudness, sharpness, roughness, tonality, impulsiveness and perceived quality. The tempo_bpm variable was entirely missing in the training partition and was removed, leaving 89 predictor variables.

All feature columns were converted to numeric values. Missing values were imputed using median values estimated from the relevant fitting data. The imputer was placed inside each machine learning pipeline so that validation or test values could not influence the preprocessing. Random Forest and XGBoost used the imputed variables without scaling. RBF-SVM and k-nearest neighbours additionally used z-standardisation fitted only on the training data for each evaluation stage. No data-driven feature selection was performed before the main comparison.

Classifiers and Class Imbalance

Four conventional classifiers were evaluated. Random Forest used 100 trees, unrestricted depth, balanced class weights and seed 42. XGBoost used 100 trees, a maximum depth of 6, a learning rate of 0.3, log-loss evaluation, seed 42 and a positive class weight calculated from the CD-to-LP ratio in the fitting dataset. The RBF-SVM used C = 1.0, gamma = 'scale', probability estimation, balanced class weights and seed 42. The k-nearest neighbours classifier used k = 5 with Euclidean distance and did not support class weighting. Class weighting was restricted to the fitting process, and the validation and test distributions were not resampled during training. The complete workflow is shown in Figure 1.

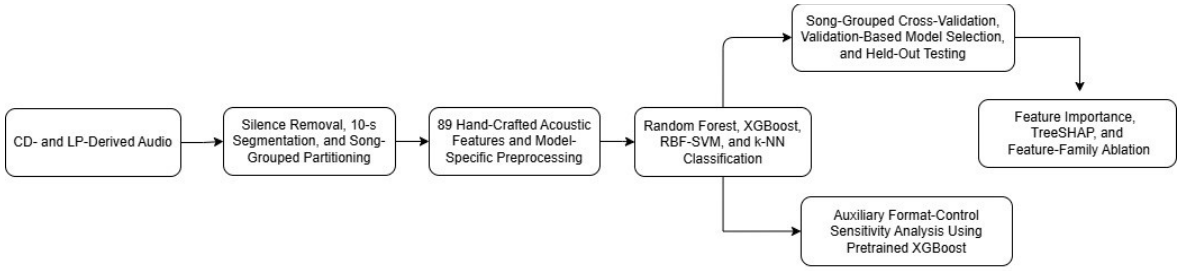

Figure 1. Machine learning workflow for CD-LP audio classification

Validation, Testing and Statistical Analysis

Five-fold stratified group cross-validation was performed within the training partition using the song identifier as the grouping variable. The pipeline repeated imputation, scaling, class-weight calculation and model fitting independently within each fold. Means and standard deviations were calculated across the five folds. Each candidate pipeline was then fitted to the complete training partition and evaluated on the independent validation partition. The primary classifier was selected using validation macro-F1, which gives equal weight to both classes.

After the final hyperparameter configurations were established, each pipeline was refitted using the combined training and validation partitions and evaluated once on the held-out test partition. Segment-level performance was evaluated using accuracy, balanced accuracy, macro-averaged precision, macro-averaged recall, macro-F1 score, area under the receiver operating characteristic curve (ROC-AUC), class-specific recall and confusion matrix analysis.

To evaluate track-level provenance, song-level LP probabilities were computed by aggregating the estimated segment probabilities within each test song. For song g containing N_g segments, the aggregated probability $\bar{p}_g$ is defined as:

$$\bar{p}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} p_{g,i} \tag{1}$$

where $p_{g,i}$ denotes the predicted LP probability for segment i of song g. The decision rule $\hat{y}_g$ assigning the song to either the LP or the CD class is:

$$\hat{y}_g = \begin{cases} \text{LP}, & \bar{p}_g \geq 0.5 \\ \text{CD}, & \bar{p}_g < 0.5 \end{cases} \tag{2}$$

Uncertainty was estimated by resampling the 90 held-out songs with replacement for 5,000 bootstrap replicates. The 95% confidence intervals were calculated for accuracy, balanced accuracy, macro-F1 score, ROC-AUC and class-specific recall. Resampling songs rather than individual segments preserves the recordings as the inferential units. Exact McNemar tests were used to compare song-level correctness for the validation-selected Random Forest against

XGBoost, SVM and k-NN. The Holm correction controlled the family-wise error rate for the three planned comparisons.

Learning curves were generated as a descriptive analysis using three stratified song-grouped folds. Within each fold, the classifiers were refitted using nested, class-stratified subsets containing 25%, 50%, 75% and 100% of the available training songs. Complete songs were retained in each subset, and macro-F1 was calculated for the fitting subset and the independent fold. The learning-curve results were not used for model selection.

Model Interpretation and Feature-Family Ablation

To determine which acoustic characteristics contributed most strongly to the classifications, feature importance was examined for both tree-based models. Impurity-based importance was calculated for Random Forest, and gain-based importance for XGBoost. The XGBoost predictions were further interpreted using TreeSHAP [32]. SHAP values were computed for 500 randomly selected test segments after applying the median imputer from the final pipeline fitted to the combined training and validation partitions. Positive SHAP values indicated a shift in the model output towards the LP class, whereas negative values indicated a shift towards the CD class. Global feature importance was determined from the mean absolute SHAP value. Because correlated features may share or redistribute their contributions, the rankings were treated as explanations of the fitted model rather than as causal or unique physical markers of either medium.

Following the model-level interpretation, a single-family ablation analysis was conducted to assess the discriminative contribution of each feature group. The validation-selected Random Forest was trained separately using spectral-harmonic, psychoacoustic, noise, MFCC, stereo, transient, dynamic/compression and temporal/rhythmic descriptors. All experiments followed the song-grouped partitions and preprocessing procedure used in the primary analysis. Feature families were defined before evaluation, and held-out performance was reported descriptively without influencing model selection. The complete assignment of the 89 predictors to the eight feature families is provided in Supplementary Table S1.

Format-Control Sensitivity Analysis

Because the CD-derived recordings were stored as MP3 files and the LP-derived recordings as WAV files, a separate sensitivity analysis was conducted to examine whether the XGBoost classifier relied primarily on format- or codec-related cues. The analysis used the 2,670-segment test partition from the original format-control workflow, comprising 1,852 CD-derived and 818 LP-derived segments. The LP WAV files were encoded as 320-kbps MP3 files, and the CD MP3 files were decoded to 44.1-kHz WAV files. The same feature-extraction procedure and training-derived preprocessing parameters were applied to each condition, and the pre-trained XGBoost model was evaluated without retraining.

Overall accuracy was calculated for the original mixed-format test set because it contained both classes. Each converted condition, however, contained only one class, so the converted results were reported as class-label retention rates: LP recall following WAV-to-MP3 conversion and CD recall following MP3-to-WAV decoding. These values do not represent complete binary-

classification accuracies and are not directly comparable with the results from the primary 2,571-segment song-grouped test set. In addition, decoding MP3 audio to WAV does not remove compression artefacts already introduced during MP3 encoding, while converting LP WAV files to MP3 introduces a new lossy-compression stage. The analysis was therefore used to determine whether the presented file format was the sole basis of classification, rather than to establish complete codec independence.

Results

Song-Grouped Cross-Validation and Validation Selection

The tree ensembles produced the strongest song-grouped cross-validation results (Table 1). Random Forest achieved a mean accuracy of 0.9883 and a macro-F1 of 0.9870, whereas XGBoost achieved 0.9864 and 0.9846, respectively. SVM remained strong, with a macro-F1 of 0.9667, whereas k-NN achieved 0.9156. ROC-AUC was high for every classifier, but LP recall varied more across folds than CD recall. This variation indicates that performance was more sensitive to the excluded minority-class songs.

Table 1. Five-fold stratified song-grouped cross-validation (mean ± standard deviation)

Model	Accuracy	Balanced acc.	Macro-F1	ROC-AUC	CD recall	LP recall
Random Forest	0.9883 ± 0.0238	0.9844 ± 0.0322	0.9870 ± 0.0264	0.99998 ± 0.00003	0.9996 ± 0.0006	0.9691 ± 0.0647
XGBoost	0.9864 ± 0.0235	0.9831 ± 0.0324	0.9846 ± 0.0260	0.99989 ± 0.00018	0.9966 ± 0.0055	0.9696 ± 0.0665
RBF-SVM	0.9718 ± 0.0227	0.9632 ± 0.0314	0.9667 ± 0.0241	0.99597 ± 0.00504	0.9915 ± 0.0056	0.9349 ± 0.0658
k-NN	0.9311 ± 0.0213	0.9177 ± 0.0302	0.9156 ± 0.0214	0.96513 ± 0.01911	0.9591 ± 0.0222	0.8763 ± 0.0665

On the independent validation partition, Random Forest achieved the highest macro-F1 score (0.9792) and was selected as the primary classifier. The validation macro-F1 was 0.9762 for XGBoost, 0.9384 for SVM and 0.8909 for k-NN. Random Forest validation accuracy was 0.9820, balanced accuracy 0.9724, ROC-AUC 0.9999, CD recall 1.0000 and LP recall 0.9448.

Held-Out Segment-Level Performance

After all four pipelines were refitted using the combined training and validation partitions, Random Forest and XGBoost each misclassified one of the 2,571 test segments. In both cases one CD-derived segment was labelled as LP, while all 849 LP-derived segments were classified correctly. Both models achieved an accuracy and macro-F1 of 0.9996. SVM and k-NN achieved macro-F1 scores of 0.9757 and 0.9355, respectively, and both showed lower recall for LP-derived than for CD-derived segments (Table 2). The confusion matrices and ROC curves are shown in Figures 2 and 3.

Table 2. Segment-level performance on the held-out test partition

Model	Accuracy	Balanced acc.	Macro precision	Macro-F1	ROC-AUC	CD recall	LP recall
Random Forest	0.9996	0.9997	0.9994	0.9996	1.0000	0.9994	1.0000
XGBoost	0.9996	0.9997	0.9994	0.9996	1.0000	0.9994	1.0000
RBF-SVM	0.9786	0.9736	0.9779	0.9757	0.9982	0.9884	0.9588
k-NN	0.9428	0.9364	0.9346	0.9355	0.9691	0.9553	0.9176

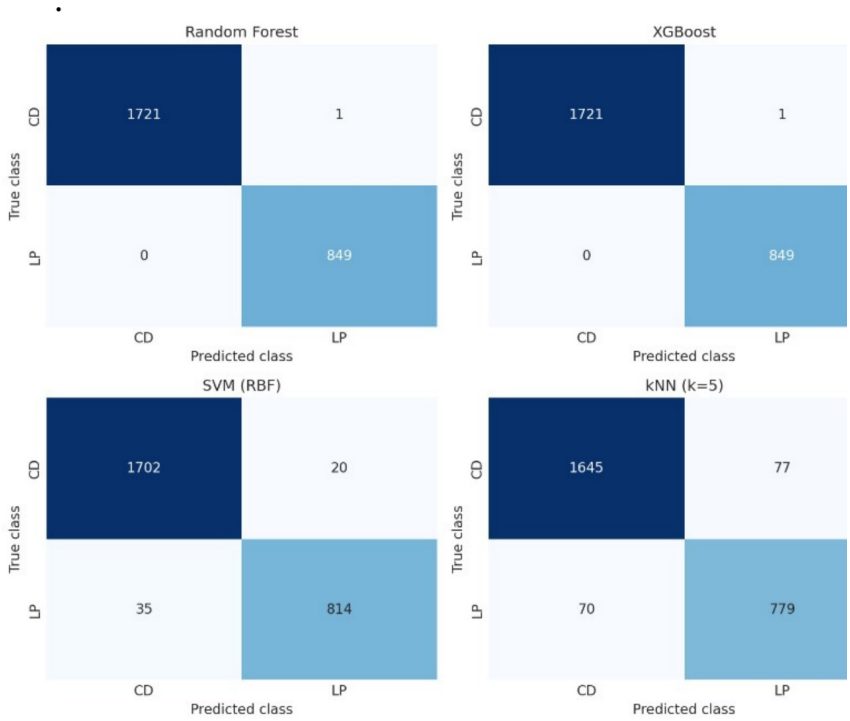

Figure 2. Segment-level confusion matrices for the song-grouped held-out test partition

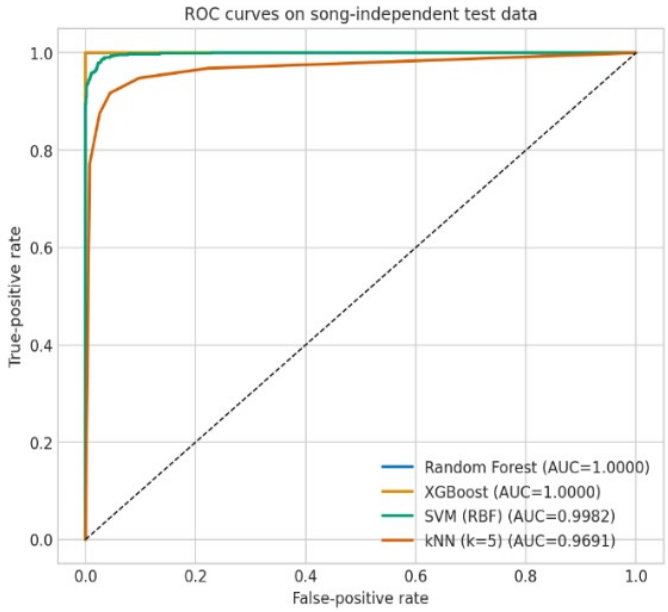

Figure 3. Receiver operating characteristic curves for the song-grouped held-out test partition

Song-Level Performance, Confidence Intervals and Paired Tests

Aggregating segment probabilities at the song level produced correct classifications for all 90 held-out songs with Random Forest and XGBoost. SVM misclassified two songs, whereas k-NN misclassified four. The bootstrap intervals of [1.000, 1.000] for the two tree ensembles resulted mechanically from observing no errors in the fixed test set and should not be interpreted as evidence of error-free performance on new collections (Table 3).

Table 3. Song-level performance on 90 held-out songs; percentile intervals used 5,000 song-level bootstrap replicates

Model	Accuracy (95% CI)	Balanced acc.	Macro-F1 (95% CI)	ROC-AUC	CD recall	LP recall
Random Forest	1.0000 (1.000-1.000)	1.0000	1.0000 (1.000-1.000)	1.0000	1.0000	1.0000
XGBoost	1.0000 (1.000-1.000)	1.0000	1.0000 (1.000-1.000)	1.0000	1.0000	1.0000
RBF-SVM	0.9778 (0.944-1.000)	0.9730	0.9730 (0.928-1.000)	0.9994	0.9844	0.9615
k-NN	0.9556 (0.911-0.989)	0.9459	0.9459 (0.888-0.988)	0.9862	0.9688	0.9231

Exact McNemar comparisons did not reveal significant song-level differences after Holm correction. Random Forest and XGBoost produced identical song predictions (adjusted $p = 1.000$). The adjusted p-value was 1.000 for Random Forest versus SVM and 0.375 for Random Forest versus k-NN. The numerical ranking should therefore not be taken as evidence that the Random Forest model statistically outperformed the alternative models.

Model Interpretation

Both tree ensembles gave the greatest importance to spectral and high-frequency variables. Random Forest ranked ultra_high_freq_energy highest (importance = 0.232), followed by high_freq_noise (0.146), spectral_flatness_mean (0.140), tonality (0.101) and perceived_quality (0.093). XGBoost placed substantially greater importance on ultra_high_freq_energy (0.669), followed by mfcc_8_mean (0.056), spectral_rolloff_85_mean (0.051), spectral_flatness_mean (0.050) and high_freq_energy_12k (0.028).

TreeSHAP confirmed the dominant contribution of ultra_high_freq_energy but gave a somewhat different ordering of the remaining variables (Figures 4 and 5). Mean absolute SHAP magnitude was next highest for high_freq_noise and high_freq_energy_12k, followed by mfcc_5_mean and spectral_flatness_mean. The SHAP distribution also showed that the direction of a contribution depended on the observed feature value. These attributions explain how the fitted model used the variables; they cannot determine whether a pattern originated from vinyl playback, MP3 encoding, mastering, musical content or interactions among these factors.

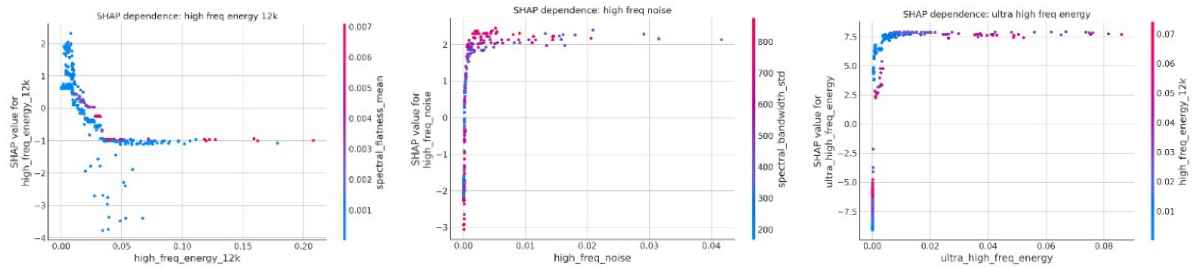

Figure 4. XGBoost SHAP dependence plots for high_freq_energy_12k, high_freq_noise and ultra_high_freq_energy. Colour denotes the interacting feature selected by the SHAP dependence analysis

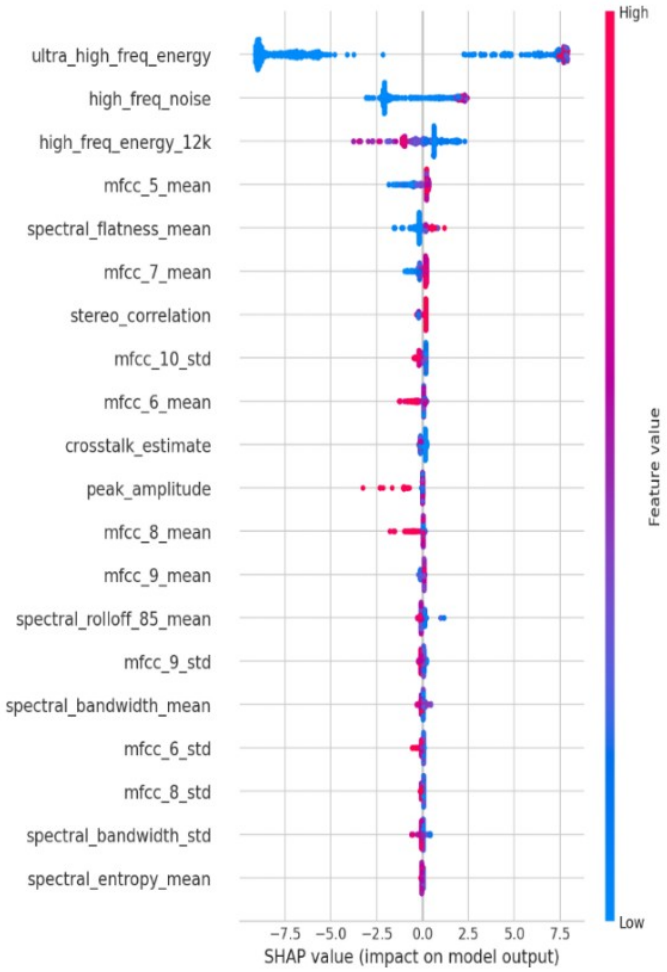

Figure 5. XGBoost SHAP summary for 500 randomly sampled held-out test segments. Positive values move the model output towards LP, whereas negative values move it towards CD

Feature-Family Ablation

Single-family ablation showed that the spectral-harmonic descriptors nearly reproduced the complete model, with an accuracy of 0.9981, macro-F1 of 0.9978 and ROC-AUC of 0.99997 (Table 4 and Figure 6). Psychoacoustic variables also showed strong discrimination (macro-F1 = 0.9840), followed by noise-related variables (0.9371) and MFCCs (0.8946). Stereo, transient, dynamic/compression and temporal/rhythmic features yielded progressively lower

performance. Their lower LP recall indicates a tendency to assign LP-derived segments to the majority CD class when stronger spectral and noise cues are unavailable.

Table 4. Held-out segment-level results for Random Forest single-family feature ablation

Feature family	No. of features	Accuracy	Balanced acc.	Macro-F1	ROC-AUC	CD recall	LP recall
Spectral-harmonic	14	0.9981	0.9983	0.9978	1.0000	0.9977	0.9988
Psychoacoustic	6	0.9860	0.9803	0.9840	0.9997	0.9971	0.9635
Noise	8	0.9452	0.9307	0.9371	0.9852	0.9733	0.8881
MFCC	26	0.9102	0.8813	0.8946	0.9686	0.9663	0.7962
Stereo	6	0.8051	0.7530	0.7659	0.8578	0.9065	0.5995
Transient	10	0.7406	0.6690	0.6790	0.7837	0.8798	0.4582
Dynamic/ compression	16	0.7106	0.6317	0.6384	0.7456	0.8641	0.3993
Temporal/rhythmic	3	0.6324	0.5274	0.5172	0.5659	0.8368	0.2179

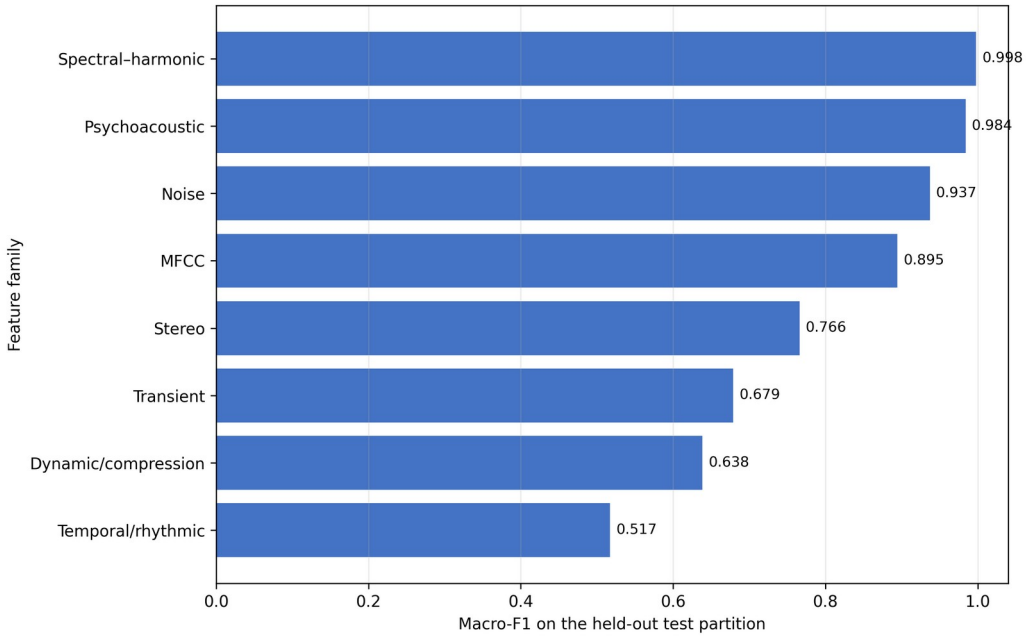

Figure 6. Macro-F1 for Random Forest models trained using one hand-crafted feature family at a time. The results are descriptive and were not used to select the primary model

Song-Grouped Learning Curves

Random Forest and XGBoost maintained high grouped-validation performance across the training fractions (Figure 7). With 25% of the available training songs, their mean validation macro-F1 scores were 0.9779 and 0.9712, respectively; with all available fold-training songs, the corresponding values were 0.9835 and 0.9815. SVM increased from 0.9304 to 0.9691, while k-NN increased from 0.8616 to 0.9125. Random Forest and XGBoost achieved a training macro-F1 of 1.0000 at every fraction.

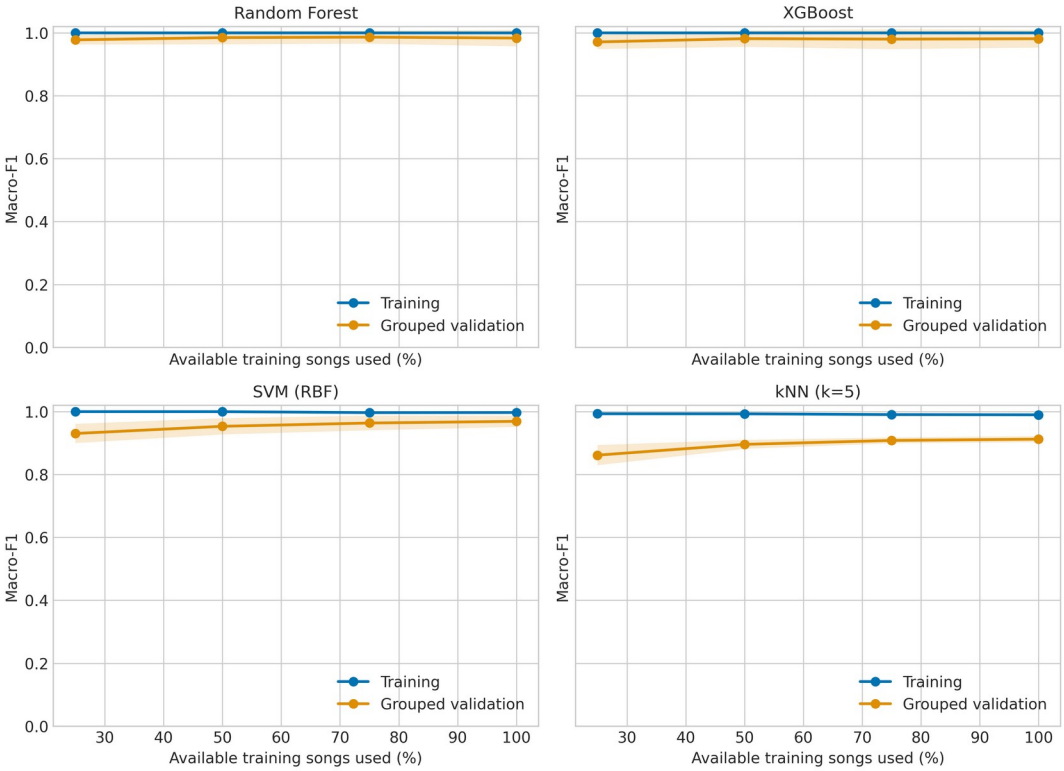

Figure 7. Song-grouped learning curves for the four classifiers. Lines show mean macro-F1 across three folds and shaded regions show ±1 standard deviation

Format-Control Sensitivity Analysis

On the original mixed-format auxiliary test set, XGBoost achieved an overall accuracy of 0.9993. The LP-label retention rate was 0.9413 after WAV-to-MP3 conversion, and the CD-label retention rate was 0.9390 after MP3-to-WAV decoding (Table 5).

Table 5. Auxiliary XGBoost format-control sensitivity analysis

Condition	Class composition	Segments	Reported statistic	Rate
Original files	CD MP3 + LP WAV	2,670	Overall accuracy	0.9993
LP WAV → MP3 (320 kbps)	LP only	818	LP-label retention (LP recall)	0.9413
CD MP3 → WAV (44.1 kHz)	CD only	1,852	CD-label retention (CD recall)	0.9390

Most converted segments retained predictions associated with their original acquisition class, suggesting that the converted signal format presented at inference was not the sole determinant of classification. Because each converted condition contained only one class, these values should be interpreted as label-retention rates rather than complete classification accuracies.

Discussion

Classification Performance

The hand-crafted acoustic features clearly separated the two audio collections examined in this study. Random Forest and XGBoost each misclassified only one test segment, and both models correctly classified all 90 held-out songs. Because all segments belonging to a song remained within the same partition, these results were not caused by the same song appearing in both the training and test data. The similar performance obtained through song-grouped cross-validation and held-out testing further indicates that the models could classify songs not used during training.

Random Forest and XGBoost produced nearly identical results, showing that the observed performance was not restricted to one tree-ensemble method. SVM also performed well, although it made more errors on LP-derived recordings. Among the four models, k-NN produced the lowest accuracy and macro-F1 score. At the song level, however, the differences between the models were small: Random Forest and XGBoost classified all songs correctly, while SVM and k-NN misclassified only two and four songs, respectively. Because so few song-level predictions differed, the McNemar tests had limited power to distinguish the models statistically. The findings therefore support the use of tree ensembles for this dataset but do not establish that one ensemble was superior to the other.

The learning curves followed the same comparative pattern. Grouped-validation macro-F1 was already high for Random Forest and XGBoost when 25% of the available training songs were used and changed only modestly as the training subset increased. SVM and k-NN improved more steadily with additional songs. The tree ensembles achieved perfect training macro-F1 at every fraction, showing that they could fit the development subsets completely; their lower grouped-validation scores remain the more relevant estimate of performance on unseen songs within this collection.

Interpretation of the Acoustic Features

The feature-importance, SHAP and ablation analyses identified spectral-harmonic characteristics as the main source of discrimination. Ultra-high-frequency energy, high-frequency noise, spectral flatness, spectral roll-off and MFCC statistics contributed strongly to the model predictions. The spectral-harmonic family also retained almost the entire performance of the full feature set, reaching a macro-F1 score of 0.9978 when evaluated on its own.

These features may represent differences in bandwidth, spectral balance, background-noise texture, equalisation, analogue playback response or encoding. Previous studies have shown that clicks, crackles, scratches and other disturbances in disc recordings produce measurable spectral and temporal patterns [1]-[5]. The present results are compatible with this evidence, although the important features cannot be attributed exclusively to vinyl artefacts.

Psychoacoustic and noise-related descriptors also provided strong discrimination. Their contribution may partly overlap with that of the spectral-harmonic features, because measures

such as tonality, perceived quality, spectral flatness and high-frequency noise describe related aspects of the signal. The performance of these groups therefore suggests that differences in spectral texture and bandwidth were important to the classifier. It does not demonstrate that every descriptor represented a separate physical property of LP reproduction.

Stereo, transient, dynamic/compression and temporal/rhythmic features were less effective when evaluated on their own, although they may still have contributed to the complete model through interactions with other variables. The poor LP recall obtained from the dynamic/compression and temporal/rhythmic groups shows that differences in signal level and musical rhythm alone were insufficient for reliable classification. Rhythmic descriptors are also more likely to reflect differences in musical content or genre than differences in the acquisition medium. Stereo features showed moderate discrimination, which may have arisen from channel correlation, playback configuration or mastering practices [29].

Limitations and Scope of the Findings

The findings should be interpreted within the design of the dataset. CD-derived recordings were available as 320-kbps MP3 files, whereas LP-derived recordings were available as WAV files. Some of the observed differences in high-frequency energy and spectral characteristics may therefore reflect encoding as well as medium- and playback-related effects.

The format-control analysis helped to examine this issue. Most LP recordings converted from WAV to MP3 and most CD recordings decoded from MP3 to WAV retained their original class predictions, indicating that the file format presented at inference was not the sole basis for classification. However, these conversions did not remove the signal changes introduced by the original codec. The results should therefore be interpreted as sensitivity evidence rather than as proof of complete codec independence.

The two classes also contained different songs. Song-grouped splitting prevented direct overlap between the development and test partitions, but factors such as genre, mastering, release period and production style could not be completely controlled [25]-[28]. These factors may have contributed to the observed discrimination together with acquisition-related characteristics. The strong song-level performance therefore demonstrates clear separability within the evaluated collection, while remaining specific to the experimental conditions considered in this study.

Implications and Future Validation

Despite these limitations, the two source classes were clearly separable within the present dataset. The feature-family analysis and SHAP results also showed that this separation was mainly associated with spectral-harmonic, psychoacoustic and noise-related characteristics. This adds interpretive value to the classification results by showing which acoustic properties contributed most strongly to the model decisions.

Such findings may be useful in practical settings such as archive organisation, source verification and provenance assessment, particularly when the recording and transfer conditions are similar to those represented in the training data. At the same time, broader

validation is necessary before the observed patterns can be treated as medium-specific characteristics.

Future studies should therefore examine CD- and LP-derived recordings under more closely matched processing conditions and, where possible, use the same musical performances and mastering sources. Evaluation across different playback systems, recording chains, musical styles and independent collections would help to establish how consistently the identified acoustic characteristics generalise beyond the present dataset.

Conclusion

This study evaluated four conventional machine learning classifiers for distinguishing CD-derived from LP-derived music using 89 hand-crafted acoustic descriptors and song-grouped evaluation. Random Forest and XGBoost achieved 99.96% segment-level accuracy and classified all 90 held-out songs correctly, while SVM and k-NN also performed strongly. Exact McNemar testing did not reveal significant song-level differences between Random Forest and the alternative classifiers. XGBoost SHAP values, tree-importance measures and feature-family ablation consistently identified spectral-harmonic, psychoacoustic and noise-related properties as the principal sources of discrimination. In the auxiliary format-control analysis, most converted LP and CD segments retained their original class predictions, suggesting that the converted signal format presented during inference was not the sole basis for classification; however, this analysis did not eliminate residual codec effects. The findings demonstrate the practical separability of the evaluated acquisition classes and the value of interpretable feature-based modelling. They do not establish a universal intrinsic difference between CD and LP audio, because source format, repertoire, mastering and recording chain were not independently controlled. Format-matched, content-controlled and externally validated experiments are required before medium-specific claims can be made.

Ethics Statement

Copyrighted commercial recordings were used solely for computational analysis and were not redistributed. Generative artificial intelligence tools were used only for grammar and language editing.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

[1] P. J. W. Rayner and S. J. Godsill, "The detection and correction of artefacts in degraded gramophone recordings," in *Final Program and Paper Summaries, 1991 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA: IEEE, 1991, pp. 0_151-0_152, doi: 10.1109/ASPAA.1991.634139.

[2] P. Rajmic and J. Klimek, "Removing crackle from an LP record via wavelet analysis," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx-04)*, Naples, Italy, Oct. 2004.

[3] M. Niedzwiecki and M. Ciolek, "Elimination of impulsive disturbances from archive audio signals using bidirectional processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1046-1059, May 2013, doi: 10.1109/TASL.2013.2244090.

[4] H. T. De Carvalho, F. R. Avila and L. W. P. Biscainho, "Bayesian restoration of audio degraded by low-frequency pulses modeled via Gaussian process," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 90-103, Jan. 2021, doi: 10.1109/JSTSP.2020.3033410.

[5] Y. Liu and S. Godsill, "Multi-channel audio statistical restoration," in *2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Macau, China: IEEE, Aug. 2020, pp. 1-5, doi: 10.1109/ICSPCC50002.2020.9259487.

[6] C. F. Stallmann and A. P. Engelbrecht, "Gramophone noise detection and reconstruction using time delay artificial neural networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 6, pp. 893-905, Jun. 2017, doi: 10.1109/TSMC.2016.2523927.

[7] E. Moliner and V. Välimäki, "A two-stage U-Net for high-fidelity denoising of historical recordings," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore: IEEE, May 2022, pp. 841-845, doi: 10.1109/ICASSP43922.2022.9746977.

[8] A. Ragano, E. Benetos and A. Hines, "Audio quality assessment of vinyl music collections using self-supervised learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096274.

[9] A. Özdoğru, F. Rund and K. Fliegel, "Performance evaluation of perceptible impulsive noise detection methods based on auditory models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 1, p. 2, Jan. 2025, doi: 10.1186/s13636-024-00389-9.

[10] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, Jul. 2002, doi: 10.1109/TSA.2002.800560.

[11] T. Li and G. Tzanetakis, "Factors in automatic musical genre classification of audio signals," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA: IEEE, 2003, pp. 143-146, doi: 10.1109/ASPAA.2003.1285840.

[12] B. McFee *et al.*, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, Austin, TX, USA, 2015, pp. 18-24, doi: 10.25080/Majora-7b98e3ed-003.

[13] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206-219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[14] K. Zaman, M. Sah, C. Direkoglu and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, vol. 11, pp. 106620-106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, Sep. 1995, doi: 10.1007/BF00994018.

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA: ACM, Aug. 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[18] C. Zeng, S. Feng, D. Zhu and Z. Wang, "Source acquisition device identification from recorded audio based on spatiotemporal representation learning with multi-attention mechanisms," *Entropy*, vol. 25, no. 4, p. 626, Apr. 2023, doi: 10.3390/e25040626.

[19] M. A. Qamhan, H. Altaheri, A. H. Meftah, G. Muhammad and Y. A. Alotaibi, "Digital audio forensics: Microphone and environment classification using deep learning," *IEEE Access*, vol. 9, pp. 62719-62733, 2021, doi: 10.1109/ACCESS.2021.3073786.

[20] V. Verma and N. Khanna, "Speaker-independent source cell-phone identification for re-compressed and noisy audio recordings," *Multimedia Tools and Applications*, vol. 80, no. 15, pp. 23581-23603, Jun. 2021, doi: 10.1007/s11042-020-10205-z.

[21] V. A. Hadoltikar, V. Ratnaparkhe and R. Kumar, "Effect of format conversion on source identification from audio recordings: A study for forensic purposes," *SN Computer Science*, vol. 5, no. 1, p. 19, Nov. 2023, doi: 10.1007/s42979-023-02379-8.

[22] C. Korgialas, G. Tzolopoulos and C. Kotropoulos, "On explainable closed-set source device identification using log-Mel spectrograms from videos' audio: A Grad-CAM approach," *IEEE Access*, vol. 12, pp. 121822-121836, 2024, doi: 10.1109/ACCESS.2024.3453119.

[23] C. Korgialas and C. Kotropoulos, "Attention source device identification using audio content from videos and Grad-CAM explanations," *IEEE Open Journal of Signal Processing*, vol. 6, pp. 1124-1138, 2025, doi: 10.1109/OJSP.2025.3620713.

[24] R. Puvanendran, A. P. Sayakkara, T. Jeyarathna and R. G. Ragel, "Benchmarking data leakage and generalization in audio classification: An empirical analysis," in *2025 IEEE 19th International Conference on Industrial and Information Systems (ICIIS)*, Peradeniya, Sri Lanka: IEEE, Jan. 2026, pp. 174-179, doi: 10.1109/ICIIS69028.2026.11450559.

[25] B. L. Sturm, "An analysis of the GTZAN music genre dataset," in *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, Nara, Japan: ACM, Nov. 2012, pp. 7-12, doi: 10.1145/2390848.2390851.

[26] A. Jerzak, "An accidental benchmark: The history, contingent power, and lasting traces of the GTZAN dataset," *Digital Society*, vol. 4, no. 2, p. 34, Aug. 2025, doi: 10.1007/s44206-025-00191-w.

[27] M. Kirchberger and F. A. Russo, "Dynamic range across music genres and the perception of dynamic compression in hearing-impaired listeners," *Trends in Hearing*, vol. 20, p. 2331216516630549, Jan. 2016, doi: 10.1177/2331216516630549.

[28] M. J. Hove, P. Vuust and J. Stupacher, "Increased levels of bass in popular music recordings 1955-2016 and their relation to loudness," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2247-2253, Apr. 2019, doi: 10.1121/1.5097587.

[29] E. Vickers, "The loudness war: Background, speculation, and recommendations," in *Audio Engineering Society Convention 129*, San Francisco, CA, USA: Audio Engineering Society, Nov. 2010.

[30] C.-B. Jeon and K. Lee, "Towards robust music source separation on loud commercial music," Zenodo, Dec. 2022, doi: 10.5281/ZENODO.7342777.

[31] J. Bergner, D. Schössow, S. Preihs and J. Peissig, "Identification of discriminative acoustic dimensions in stereo, surround and 3D music reproduction," *Journal of the Audio Engineering Society*, vol. 71, no. 7/8, pp. 420-430, Jul. 2023, doi: 10.17743/jaes.2022.0071.

[32] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56-67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.