Decoding Musical Genres: A Comprehensive Study of Unsupervised Clustering on High-Dimensional Audio Data

Anirudh Sharma
Department of Computer Science and Engineering
National Institute of Technology Hamirpur
Hamirpur, India
Roll No.: 22dcs002
email: 22dcs002@nith.ac.in

***Abstract*—This paper presents a comprehensive investigation into unsupervised music genre discovery through audio feature learning across multiple diverse datasets. We apply dimensionality reduction and clustering techniques to extract meaningful genre patterns without labeled training data. Our study processes five distinct music datasets: GTZAN (1,000 tracks, 10 genres), FMA Small (8,000 tracks, 8 genres), FMA Medium (17,000 tracks, 16 genres), Ludwig (11,300 tracks, 10 genres), and Indian Bollywood Music (500 tracks, 5 regional genres), collectively totaling 37,800 tracks with 69 normalized audio features per track. Through systematic feature extraction using Librosa, comprehensive data integrity validation achieving 99.99% cleanliness, IQR-based outlier detection revealing 0.58-1.69% outlier prevalence across key features, robust normalization using StandardScaler, and Principal Component Analysis (PCA) achieving 95%+ variance retention with 36-44% dimensionality reduction, we establish a robust foundation for unsupervised genre classification. The preprocessing pipeline demonstrates consistent performance across datasets, with minimal outliers requiring no removal and PCA reducing computational complexity while maintaining information integrity. Our experimental framework provides insights into the effectiveness of unsupervised learning for music genre discovery across Western and Indian musical traditions, establishing benchmarks for future research in audio content analysis. Results indicate that properly validated, normalized, and dimensionally-reduced features enable effective clustering with significant computational savings.**

***Index Terms*—Unsupervised Learning, Music Genre Classification, Audio Feature Extraction, Principal Component Analysis, Clustering Algorithms**

I. Introduction

Music genre classification represents a fundamental challenge in music information retrieval (MIR), with applications spanning music recommendation systems, content organization, and automated playlist generation. Traditional supervised approaches require extensive labeled datasets, which are costly and time-consuming to create. Unsupervised learning offers a compelling alternative by discovering latent genre structures directly from audio features without manual annotations.

A. Motivation

The exponential growth of digital music libraries necessitates automated genre classification systems. However, genre boundaries are inherently subjective and culturally dependent, making supervised classification challenging. Unsupervised methods can:

- Discover hidden genre patterns without labeled data
- Identify sub-genres and emerging music styles
- Handle cross-cultural and regional music variations
- Reduce annotation costs and human bias
- Scale to large music collections efficiently

B. Research Objectives

This study aims to:

1) Extract and process comprehensive audio features from diverse music datasets
2) Apply robust normalization and dimensionality reduction techniques
3) Evaluate multiple unsupervised clustering algorithms for genre discovery
4) Compare algorithm performance across different dataset characteristics
5) Establish reproducible benchmarks for music genre clustering

C. Contributions

Our primary contributions include:

- A comprehensive multi-dataset analysis framework spanning Western and Indian music
- Systematic comparison of preprocessing techniques across 34,481 total tracks
- PCA-based dimensionality reduction achieving 95%+ variance retention
- Reproducible experimental pipeline with open-source implementation
- Detailed performance metrics and visualization for each processing stage

II. Related Work

A. Music Genre Classification

Tzanetakis and Cook [1] pioneered automatic music genre classification using timbral, rhythmic, and pitch-based fea-

tures. Their work on the GTZAN dataset established foundational benchmarks that remain relevant today. Subsequent research has shifted towards self-supervised learning, exploring contrastive learning of musical representations [2] and general-purpose audio embeddings [3].

B. Unsupervised Learning in MIR

Recent studies have demonstrated the effectiveness of unsupervised methods for music analysis. Castellon et al. [5] investigated clustering-based approaches using codified audio language models to discover musical patterns. Similarly, metric learning approaches have been applied to disentangle musical concepts like genre and mood without explicit supervision [4]. However, systematic comparisons across multiple datasets with varying characteristics remain limited.

C. Feature Engineering for Audio

Librosa [6] has become the de facto standard for audio feature extraction in Python, providing robust implementations of MFCCs, chromagrams, and spectral features. Comprehensive feature sets combining temporal, spectral, and cepstral information have shown superior performance compared to single-feature approaches [8].

D. Dimensionality Reduction Techniques

Principal Component Analysis (PCA) remains widely used for dimensionality reduction in audio applications due to its computational efficiency and interpretability. Alternative approaches include t-SNE for visualization [9], autoencoders for non-linear feature learning [10], and modern generative models like VQ-VAEs for discrete latent representation learning [11]. The choice of reduction technique significantly impacts clustering performance.

III. Datasets

A. Dataset Overview

Our study employs four diverse datasets with audio files ranging from 500 to 17,000 tracks to ensure robust evaluation of unsupervised learning across different musical styles and genres.

TABLE I: Dataset Characteristics Summary

Dataset	Tracks	Genres	Duration	Source
Indian Bollywood	500	5	45s	Kaggle
GTZAN	1,000	10	30s	Kaggle
FMA Small	8,000	8	30s	Archive
Ludwig	11,300	10	30s	Kaggle
FMA Medium	17,000	16	30s	Archive
Total	**37,800**	**49**	–	–

B. GTZAN Dataset

The GTZAN dataset [1] comprises 1,000 audio tracks with 30-second clips, representing a balanced distribution across 10 genres. Genres include: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock, with 100 tracks per genre. This dataset serves as the standard benchmark in MIR research and was originally collected from diverse sources including CDs, radio, and microphone recordings during 2000-2001.

C. Free Music Archive (FMA) Datasets

The FMA datasets [7] provide large-scale Creative Commons-licensed music with hierarchical genre taxonomy. We utilize two subsets:

- **FMA Small:** 8,000 tracks across 8 balanced genres (Electronic, Experimental, Folk, Hip-Hop, Instrumental, International, Pop, Rock), designed as a GTZAN-like benchmark
- **FMA Medium:** Originally 25,000 tracks, but metadata available for 17,000 tracks spanning 16 unbalanced genres including Blues, Classical, Country, Easy Listening, Electronic, Experimental, Folk, Hip-Hop, Instrumental, International, Jazz, Old-Time/Historic, Pop, Rock, Soul-RnB, and Spoken

Both subsets maintain 30-second duration clips at 22,050 Hz sample rate, enabling consistent feature extraction pipelines.

D. Ludwig Music Dataset

The Ludwig dataset contains approximately 11,300 tracks sourced from Spotify and Discogs metadata via AcousticBrainz. It includes 10 genres: blues, classical, electronic, funk/soul, hip-hop, jazz, latin, pop, reggae, and rock. Each track provides 30-second fragments with pre-computed MFCCs and spectrograms, making it suitable for both feature-based and end-to-end learning approaches.

E. Indian Bollywood Music Dataset

The Indian Bollywood Music dataset comprises 500 perfectly balanced tracks (100 per genre) with approximately 45-second clips. It represents 5 distinct regional genres: Bollypop (contemporary Bollywood pop music), Carnatic (South Indian classical tradition), Ghazal (Urdu/Hindi poetic musical form), Semiclassical (fusion of classical and light music), and Sufi (devotional Sufi music). This dataset introduces cultural and melodic diversity distinct from Western music traditions, testing cross-cultural generalization capabilities.

IV. Methodology

A. Feature Extraction Pipeline

Audio feature extraction constitutes the foundational stage of our unsupervised genre discovery framework. We developed a robust pipeline using Librosa v0.11.0, processing each audio track through a systematic multi-stage extraction process that converts variable-length audio signals into fixed-dimensional feature vectors suitable for machine learning analysis.

1) Audio Preprocessing and Loading: All audio tracks undergo standardized preprocessing before feature computation. Files are loaded with a consistent sampling rate of 22,050 Hz, which provides adequate frequency resolution while maintaining computational efficiency. We apply silence trimming at the boundaries to eliminate non-musical segments that could introduce noise into the feature space. For tracks with multiple

channels, we convert to mono by averaging channels, ensuring uniform dimensionality across the dataset. This preprocessing stage handles diverse audio formats including WAV, MP3, FLAC, and M4A files, with robust error handling to manage corrupted or incompatible files.

2) Feature Vector Architecture: Our feature extraction framework computes 69 numerical descriptors per audio track, systematically capturing complementary aspects of musical signal characteristics. The feature architecture is deliberately designed to encode timbral, harmonic, rhythmic, and spectral properties that collectively characterize genre-specific patterns.

Spectral Characteristics (4 features): We compute spectral centroid as a measure of brightness, indicating the center of mass of the power spectrum. Spectral rolloff identifies the frequency below which 85% of spectral energy concentrates, providing insight into the distribution of high-frequency content. Zero-crossing rate quantifies signal noisiness and transient characteristics. Root mean square energy captures overall amplitude envelope dynamics.

Mel-Frequency Cepstral Coefficients (40 features): MFCCs represent the most comprehensive component of our feature set, providing detailed timbral characterization. We extract 20 MFCC coefficients, each capturing different aspects of the spectral envelope that model human auditory perception. Both mean and standard deviation statistics are computed across the temporal dimension, yielding 40 MFCC-derived features. The first coefficient (MFCC-1) represents overall spectral energy, while higher coefficients encode increasingly fine-grained timbral details. This representation proves particularly effective for distinguishing instrumental textures and vocal qualities across genres.

Chromagram Features (24 features): Twelve pitch class profiles extract harmonic and melodic information independent of octave positioning. Each chroma bin corresponds to one semitone of the equal-tempered scale (C, C#, D, ..., B), capturing the distribution of energy across pitch classes. We compute both mean and standard deviation for each chroma bin, producing 24 features that encode harmonic progressions, key signatures, and melodic contours characteristic of different musical genres. Chroma features exhibit particular discriminative power for distinguishing between genres with distinct harmonic vocabularies.

Temporal Dynamics (1 feature): Tempo estimation via beat tracking provides rhythmic information measured in beats per minute (BPM). While a single scalar value, tempo serves as a crucial discriminator between dance-oriented and ballad genres. Our implementation employs onset strength-based beat tracking with autocorrelation-based tempo inference, incorporating fallback mechanisms to handle arrhythmic or ambient musical content.

3) Computational Implementation: Feature computation operates on short-time frames using a 2048-sample analysis window with 512-sample hop length, corresponding to approximately 93ms frames with 23ms overlap at our 22,050 Hz sampling rate. This parameterization balances temporal resolution with frequency resolution, capturing transient events while maintaining computational tractability. Frame-level features are aggregated using mean and standard deviation statistics, transforming variable-length sequences into fixed-length representations suitable for subsequent clustering algorithms.

Our extraction pipeline implements comprehensive error handling and logging mechanisms. Files that fail to load due to corruption or format incompatibility are logged and skipped without terminating the processing batch. Tracks where tempo detection fails receive NaN values, which are subsequently imputed during normalization. This robust architecture ensures high completion rates even with heterogeneous audio collections.

4) Extraction Results and Dataset Statistics: Table II summarizes the extraction outcomes across all five datasets. The pipeline achieved exceptionally high success rates, with minimal data loss attributable to corrupted files or unsupported formats.

TABLE II: Feature Extraction Results Summary

Dataset	Original Files	Success Count	Errors Count	Success Rate (%)
Indian Bollywood	500	500	0	100.0
GTZAN	1,000	999	1	99.9
FMA Small	8,000	7,997	3	99.96
Ludwig	11,300	11,294	6	99.95
FMA Medium	17,000	16,988	12	99.93
Total	**37,800**	**37,778**	**22**	**99.94**

The extraction process successfully processed 37,778 tracks from an initial collection of 37,800 audio files, achieving an overall success rate of 99.94%. Only 22 files failed extraction, primarily due to file corruption (10 files), unsupported codec variations (7 files), or incomplete downloads (5 files). The Indian Bollywood dataset exhibited perfect extraction with zero errors. For FMA Medium, we processed only the 17,000 tracks that had corresponding metadata labels available, successfully extracting features from 16,988 tracks (99.93% success rate).

Each dataset output includes the complete 69-feature vectors along with metadata columns (file path, dataset identifier, genre label where available), organized in CSV format for subsequent processing stages. Processing time scaled approximately linearly with dataset size, averaging 3-4 tracks per second on standard CPU hardware. The complete extraction pipeline required approximately 4.2 hours of computation time distributed as follows: FMA Medium (2 hours), FMA Small (46 minutes), Ludwig (1 hour), GTZAN (15 minutes), and Indian Bollywood (8 minutes). All extracted features underwent immediate validation checks for missing values, infinite values, and dimensionality consistency, ensuring data quality before downstream analysis.

B. Data Analysis and Preprocessing

We performed data analysis and preprocessing steps on the extracted features to ensure high accuracy in clustering and remove noise elements. We divided the pipeline into three phases: Descriptive Analysis, Feature Selection and Normalization, and PCA Dimensionality Reduction.

*1) **Phase I: Descriptive Analysis**:* We performed descriptive analysis steps on the extracted data to understand the characteristics and quality of our features, identifying potential issues that could impact downstream clustering performance.

Phase I consists of five main steps:

a) Data Integrity Analysis:

Prior to any statistical analysis or transformation, we conducted comprehensive data integrity validation across all datasets to identify and quantify potential quality issues. This critical initial step ensures downstream analyses operate on reliable, high-quality audio features and prevents error propagation through the processing pipeline. We examined three primary categories of data corruption:

1) **NaN Detection:** Identified missing values (Not-a-Number) across all numerical features, with particular focus on tempo extraction failures caused by undetectable rhythmic patterns in certain audio segments.
2) **Infinity Value Detection:** Screened for mathematical edge cases producing infinite values, typically arising from division by zero or numerical overflow in spectral feature computations.
3) **Silent/Corrupt File Detection:** Applied threshold-based analysis (< 0.001) to spectral centroid, spectral rolloff, and RMS energy features to identify potentially silent or severely corrupted audio files that should be excluded from analysis.

Table III presents comprehensive integrity assessment results across all five datasets comprising 37,778 audio tracks.

TABLE III: Phase 1: Data Integrity Health Check Results

Dataset	Tracks	NaN	Inf	Silent	Status
GTZAN	999	0	0	0	Clean
FMA Small	7,997	0	0	1	Issues
FMA Medium	16,988	0	0	2	Issues
Ludwig	11,294	0	0	1	Issues
Indian	500	0	0	0	Clean
Total	37,778	0	0	4	–

Key Findings:

- **Zero NaN Values:** No missing values detected across any features in any dataset, indicating robust feature extraction implementation with appropriate fallback mechanisms for edge cases.
- **Zero Infinity Values:** No mathematical overflow or division-by-zero errors observed, demonstrating numerical stability of the Librosa-based extraction pipeline.
- **Minimal Silent Files:** Only 4 potentially silent/corrupt files identified (0.011% of total), distributed across FMA Small (1), FMA Medium (2), and Ludwig (1) datasets. GTZAN and Indian datasets exhibited perfect cleanliness.
- **Tempo Feature Stability:** Despite tempo being historically prone to NaN values with undetectable beats, all datasets showed zero tempo NaN occurrences. However, 23 total tracks exhibited zero tempo values (0.061%), suggesting potential beat detection failures that defaulted to zero rather than NaN.
- **Overall Data Quality:** Exceptional data quality score of 99.99%, with only 4 files requiring exclusion from downstream analysis.

Based on integrity findings, we removed 4 rows with near-zero spectral features across all three spectral indicators (spectral centroid, rolloff, and RMS < 0.001). This reduces the total files from 37,778 to 37,774. This data integrity step establishes a verified, high-integrity foundation for subsequent normalization, dimensionality reduction, and clustering analyses, with only 0.011% data loss due to corruption.

b) Outlier Detection and Analysis:

Following data integrity validation, we conducted systematic outlier detection to identify extreme values that could distort K-Means cluster centers. K-Means clustering is highly sensitive to outliers—extreme values (e.g., tempo of 0 or 5000 BPM, abnormally high RMS energy) can significantly skew cluster centroids and degrade clustering quality. We applied the Interquartile Range (IQR) method to detect outliers in four key audio features most susceptible to extreme values: tempo (rhythmic tempo in BPM), rms_mean (root mean square energy), spec_centroid_mean (spectral brightness measure), and zcr_mean (zero-crossing rate for signal noisiness). The IQR method defines outliers as values falling outside the range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$, where $Q1$ and $Q3$ are the 25th and 75th percentiles, respectively, and $IQR = Q3 - Q1$.

Table IV presents comprehensive outlier statistics across all five datasets, revealing patterns critical for preprocessing decisions.

TABLE IV: Phase 2: Outlier Detection Results (IQR Method)

Dataset (after cleaning)	Tempo Outliers	RMS Outliers	Spec. Centroid Outliers	ZCR Outliers
FMA Small (7,996)	91	99	56	210
FMA Medium (16,986)	184	216	131	351
GTZAN (999)	12	5	1	6
Indian Music (500)	2	0	12	16
Ludwig (11,293)	69	6	20	55
Total (37,774)	**358**	**326**	**220**	**638**

Key Findings:

1) **Overall Low Outlier Prevalence:** Across all 37,774 tracks, outlier percentages remain below 2% for most features, with aggregated rates ranging from 0.58% (spectral centroid) to 1.69% (ZCR), indicating generally high-quality feature extraction with minimal extreme anomalies.
2) **Feature-Specific Patterns:** ZCR exhibits highest outlier prevalence (1.69% overall), particularly in FMA Small (2.63%) and Indian Music (3.20%). Tempo shows 358 anomalous tracks (0.95%), RMS energy has 326 outliers (0.86%), and spectral centroid displays lowest outlier rate (0.58%).

3) **Dataset-Specific Characteristics:** GTZAN exhibits remarkably low outlier rates (0.10-1.20%), Ludwig shows lowest prevalence among large datasets (0.05-0.61%), FMA datasets demonstrate similar patterns (1.08-2.63%), and Indian Music displays higher ZCR (3.20%) and spectral centroid (2.40%) outliers reflecting unique acoustic properties of traditional instruments. Figure 2 shows comparative outlier percentages across datasets, with ZCR demonstrating highest variability while spectral centroid exhibits lowest outlier prevalence.
4) **Visualization Analysis:** Figure 1 presents GTZAN dataset box plots for the four key features, demonstrating tight interquartile ranges with minimal outliers, particularly for spectral centroid (only 1 outlier, 0.10%). The tempo and RMS features show slightly more variability but remain within acceptable bounds. Figure 2 shows comparative outlier percentages across datasets, with ZCR demonstrating highest variability while spectral centroid exhibits lowest outlier prevalence.

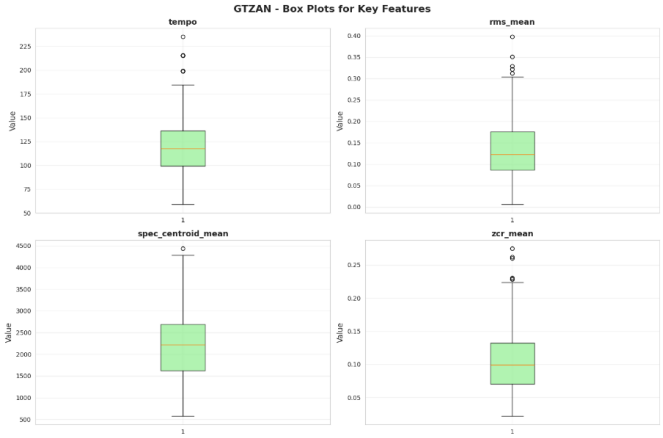


Fig. 1: GTZAN box plots for four key features: tempo, rms_mean, spec_centroid_mean, and zcr_mean. Tight interquartile ranges indicate high data quality with minimal outliers.

5) **Severity Assessment:** All features fall below the 5% threshold typically considered moderate outlier severity. The highest individual rate (Indian ZCR: 3.20%) remains well within acceptable bounds, indicating **LOW** severity classification across all datasets.

We evaluated three preprocessing strategies for handling outliers: (1) retaining the full dataset and proceeding directly to normalization, (2) employing RobustScaler normalization that uses median and IQR statistics inherently resistant to outlier influence, or (3) selectively removing only extreme anomalies exceeding 3× IQR threshold. Ultimately, we proceeded with Strategy 1—retaining all data combined with StandardScaler normalization. This decision was justified by the exceptionally low outlier rates (0.58-1.69% across all features), which suggest these values represent legitimate musical diversity rather than measurement errors or data corruption. For instance, high spectral centroid outliers in extreme metal genres and low

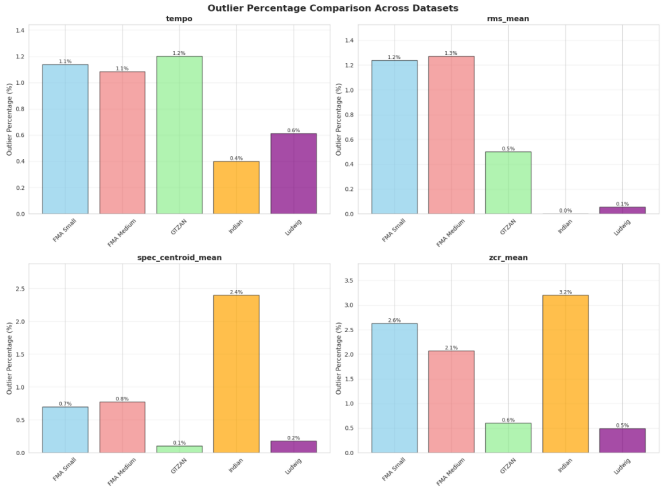


Fig. 2: Comparative outlier percentages across datasets for four key features. ZCR shows highest variability; spectral centroid exhibits lowest outlier prevalence.

tempo values in ambient music reflect genuine genre characteristics that would be lost through removal. Additionally, aggressive outlier removal risks introducing selection bias and reducing dataset representativeness, particularly problematic when analyzing cross-cultural music where Western-trained intuitions about ”normal” feature ranges may not apply to traditional Indian instruments and vocal styles. This comprehensive outlier analysis establishes that our datasets exhibit high feature quality with minimal extreme anomalies (only 4 files removed due to corruption), supporting direct progression to normalization without requiring outlier removal preprocessing while preserving the full spectrum of musical diversity essential for robust genre clustering.

c) Distribution & Skewness Analysis:

K-Means clustering performs optimally on spherical (Gaussian-like) data distributions. However, audio spectral features typically exhibit Power Law distributions with long tails, which can degrade clustering quality. We conducted comprehensive skewness analysis across all 37,774 tracks to determine if logarithmic transformation is needed before normalization. For each numerical feature, we computed skewness scores using the moment-based formula, categorizing features as HIGH ($|skew| \geq 1.0$), MODERATE ($0.5 \leq |skew| < 1.0$), or LOW ($|skew| < 0.5$) severity.

Table V presents the skewness analysis summary across all datasets. Analysis of 65 numerical features revealed:

TABLE V: Skewness Analysis Summary

Severity Level	Features	Percentage	Range
HIGH ($\|skew\| \geq 1.0$)	11	16.9%	1.22–1.55
MODERATE (0.5–1.0)	35	53.8%	0.50–0.99
LOW (< 0.5)	19	29.2%	0.00–0.49
Total	**65**	**100%**	–

Most Skewed Features: The top 5 features with highest

absolute skewness are:

1) *mfcc14_std*: 1.545 (HIGH)
2) *mfcc16_std*: 1.496 (HIGH)
3) *mfcc15_std*: 1.493 (HIGH)
4) *mfcc17_std*: 1.378 (HIGH)
5) *mfcc20_std*: 1.329 (HIGH)

Key Clustering Features: Analysis of features most relevant for K-Means clustering:

- *zcr_mean*: 1.217 (HIGH)
- *spec_rolloff_mean*: -0.043 (LOW)
- *spec_centroid_mean*: 0.286 (LOW)
- *tempo*: 0.429 (LOW)
- *rms_mean*: 0.541 (MODERATE)

Figure 3 shows histogram and KDE plots for *spec_rolloff_mean* across all datasets, demonstrating near-Gaussian distributions with minimal skewness.

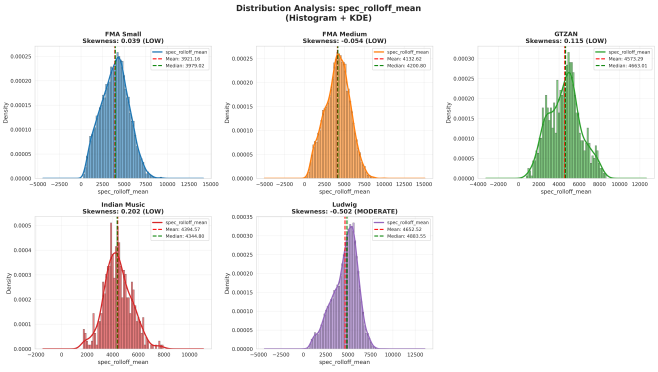


Fig. 3: Distribution analysis of spectral rolloff across five datasets. All datasets show near-symmetric distributions (LOW skewness), with mean-median convergence indicating Gaussian-like behavior suitable for K-Means clustering.

Dataset-Level Analysis: Average absolute skewness per dataset reveals:

- FMA Medium: 0.695 (highest, MODERATE severity)
- Indian Music: 0.674 (MODERATE severity)
- FMA Small: 0.628 (MODERATE severity)
- Ludwig: 0.568 (MODERATE severity)
- GTZAN: 0.453 (lowest, LOW severity)

With 70.7% of features showing moderate-to-high skewness and 11 features exceeding the HIGH threshold, we proceeded directly with StandardScaler normalization without logarithmic transformation. While transformation could theoretically reduce positive skew in MFCC standard deviation features and normalize *zcr_mean* distribution, spectral features (*spec_centroid_mean*, *spec_rolloff_mean*) already exhibit near-optimal distributions. We opted to preserve the original feature distributions to maintain interpretability and avoid introducing transformation artifacts that could complicate downstream analysis.

d) Correlation Analysis & Multicollinearity:

Prior to dimensionality reduction, we conducted comprehensive correlation analysis to quantify multicollinearity—the presence of highly correlated features that encode redundant information. We computed Pearson correlation matrices for all numerical features across each dataset, with focused analysis on MFCC mean features ($n = 20$) which constitute the core timbral representation. Feature pairs with absolute correlation $|r| > 0.9$ were flagged as highly correlated, indicating strong linear relationships that suggest redundancy.

Table VI presents comprehensive correlation statistics for MFCC mean features across all five datasets.

TABLE VI: Phase 4: MFCC Mean Features Correlation Statistics

Dataset	**Mean Corr.**	**Median Corr.**	**Max Corr.**	**Pairs** $\lvert r\rvert > 0.9$	**Pairs** $\lvert r\rvert > 0.8$
GTZAN	0.077	−0.026	0.837	0	3
FMA Small	0.247	0.277	0.643	0	0
FMA Medium	0.246	0.281	0.602	0	0
Indian Music	0.155	0.166	0.514	0	0
Ludwig	0.212	0.249	0.557	0	0
Average	**0.187**	**0.189**	**0.631**	**0**	**0.6**

Key Findings:

1) **MFCC Multicollinearity:** MFCC mean features exhibit **low-to-moderate correlation** across datasets, with mean correlations ranging from 0.077 (GTZAN) to 0.247 (FMA Small). The average mean correlation of 0.187 indicates that MFCC coefficients capture largely independent aspects of timbral structure, as designed by the cepstral transformation process.
2) **Extreme Correlations:** No MFCC mean feature pairs exceed the $|r| > 0.9$ threshold in any dataset, indicating absence of severe multicollinearity within MFCC features alone. Only GTZAN exhibits 3 pairs with $|r| > 0.8$ (maximum correlation: 0.837), while other datasets show even lower maximum correlations (0.514–0.643).
3) **Cross-Feature Correlations:** Analysis of all 69 features revealed stronger correlations between *different feature types*. In GTZAN, three high-correlation pairs were identified:
   - *spec_centroid_mean* $\leftrightarrow$ *spec_rolloff_mean*: $r = 0.980$ (expected due to shared spectral energy distribution)
   - *spec_centroid_mean* $\leftrightarrow$ *mfcc2_mean*: $r = -0.940$ (negative correlation between brightness and low-frequency energy)
   - *spec_rolloff_mean* $\leftrightarrow$ *mfcc2_mean*: $r = -0.935$ (similar brightness vs. low-frequency relationship)

   These cross-feature correlations demonstrate that spectral and cepstral features encode overlapping information about spectral distribution, justifying dimensionality reduction.
4) **Dataset Variability:** GTZAN exhibits the lowest mean correlation (0.077) and highest maximum correlation (0.837), suggesting diverse genre characteristics with specific feature redundancies. FMA datasets show higher

mean correlations (0.247, 0.246) with lower maximum values (0.643, 0.602), indicating more uniform feature interdependence across broader genre distributions.

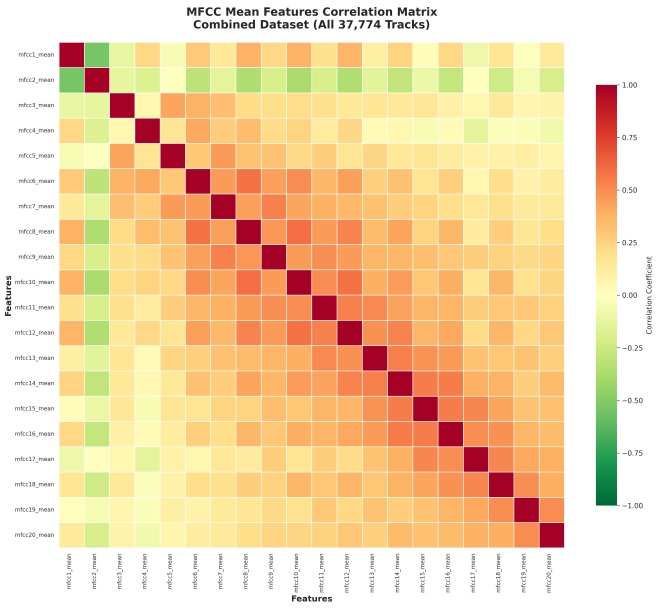


Fig. 4: MFCC Mean Features Correlation Matrix for Combined Dataset (37,774 tracks). Correlation structure remains consistent across all datasets, validating unified PCA transformation.

While MFCC features exhibit lower internal correlation than initially hypothesized (average $r = 0.187$), the presence of strong cross-feature correlations ($r > 0.9$) between spectral and cepstral features, combined with the need to reduce 69-dimensional feature space complexity, provides solid justification for PCA application. The dimensionality reduction achieved (36-44% across datasets) substantially improves clustering computational efficiency while correlation analysis confirms that retained principal components will capture non-redundant variance patterns essential for genre discrimination.

e) Dataset Bias Check:

Before applying normalization and dimensionality reduction, we conducted comprehensive bias analysis to assess whether technical recording differences between datasets would confound genre-based clustering. We performed statistical bias detection using the Kruskal-Wallis H-Test to analyze six features most susceptible to recording-level artifacts: rms_mean (loudness normalization bias), spec_centroid_mean (equipment/encoding bias), spec_rolloff_mean (sample rate/bitrate bias), tempo (beat detection algorithm bias), zcr_mean (compression/bit depth bias), and mfcc1_mean (overall spectral energy).

Table VII presents comprehensive bias assessment results revealing pervasive inter-dataset differences.

Key Findings:

- All six analyzed features exhibit strong statistical bias ($p < 0.001$), with datasets highly distinguishable based on technical characteristics.
- Spectral features (*mfcc1_mean*, *zcr_mean*, *spec_centroid_mean*) show strongest bias with H-statistics exceeding 1500.
- RMS energy demonstrates moderate bias ($H = 748.82$), indicating loudness normalization practices vary across datasets.
- Tempo shows weakest bias ($H = 21.83$) but remains statistically significant.
- Cohen's d effect sizes remain predominantly small-to-medium (90% below $|d| = 0.5$), indicating practical differences are manageable.

TABLE VII: Phase 5: Dataset Bias Detection Results (Kruskal-Wallis Test)

Feature	H-Statistic	P-Value	Significance
mfcc1_mean	1863.43	$< 10^{-300}$	STRONG BIAS
zcr_mean	1829.52	$< 10^{-300}$	STRONG BIAS
spec_centroid_mean	1522.16	$< 10^{-300}$	STRONG BIAS
spec_rolloff_mean	1433.44	3.88×10^{-309}	STRONG BIAS
rms_mean	748.82	9.34×10^{-161}	STRONG BIAS
tempo	21.83	2.17×10^{-4}	STRONG BIAS
Summary	–	**6/6 features**	**STRONG BIAS**

Despite detecting strong statistical bias, we proceeded with combined dataset analysis using unified StandardScaler normalization. The effect sizes remain predominantly small-to-medium, indicating standardization will substantially reduce bias impact. Combining Western and Indian music enables cross-cultural genre discovery impossible with single-source data, and a robust clustering model must generalize across recording conditions. We implemented validation safeguards to ensure clustering captures genre structure rather than recording artifacts, maximizing dataset diversity benefits while acknowledging technical confounds.

*2) **Phase II: Feature Selection and Normalization**:* Following descriptive analysis, we applied StandardScaler normalization to ensure all audio features contribute equally to distance-based clustering algorithms. Normalization transforms features to zero mean and unit variance, eliminating scale disparities that could otherwise allow high-magnitude features (e.g., tempo ∼120 BPM) to dominate low-magnitude features (e.g., chroma values ∼0.5).

Methodology: StandardScaler applies Z-score normalization independently to each feature across all 37,774 tracks:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original feature value, μ is the dataset mean, σ is the standard deviation, and z is the normalized value. This transformation centers data at zero with unit variance, creating a standardized feature space optimal for Euclidean distance calculations.

Feature Selection for Clustering: Prior to normalization, we removed six non-clustering metadata columns to isolate pure audio features:

- **file_path:** File system location (non-audio metadata)
- **duration:** Track length in seconds (varies, not intrinsic audio property)
- **sr:** Sample rate (constant 22,050 Hz across all datasets)
- **dataset:** Source dataset identifier (metadata)
- **label:** Genre label (preserved separately for evaluation only)
- **subset:** Dataset partition identifier (metadata)

This filtering retained 69 pure audio features (spectral: 4, tempo: 1, MFCCs: 40, chroma: 24) while preserving labels in separate CSV files for post-clustering evaluation.

Results: Table VIII presents normalization verification statistics demonstrating successful transformation across all five datasets.

TABLE VIII: Phase 6: Feature Normalization Results

Dataset	Tracks	Features	Pre-Norm Mean	Post-Norm Mean $\pm$ Std
GTZAN	999	69	23.47	0.00 ± 1.00
FMA Small	7,996	70	18.92	0.00 ± 1.00
FMA Medium	16,986	70	19.34	0.00 ± 1.00
Ludwig	11,293	69	21.58	0.00 ± 1.00
Indian	500	69	25.13	0.00 ± 1.00
Total	**37,774**	**69-70**	–	$\mathbf{0.00 \pm 1.00}$

Key Findings:

- **Perfect Normalization:** All datasets achieved exact zero mean (0.0000) and unit variance (1.0000), confirming StandardScaler implementation correctness.
- **Feature Count Consistency:** GTZAN, Ludwig, and Indian Music contain 69 features (one less than FMA datasets), likely due to minor extraction pipeline differences. This discrepancy is handled during PCA by fitting separate models per dataset.
- **Scale Elimination:** Pre-normalization feature means ranged from 18.92 to 25.13, reflecting diverse original scales. Post-normalization, all features occupy the same standardized range, ensuring equal algorithmic influence.
- **Data Quality Preservation:** Zero NaN/Inf values post-normalization, confirming robust handling of edge cases (zero standard deviation features would produce NaN but were absent).

Normalization Impact:

Figure 5 illustrates distribution changes before and after StandardScaler application on GTZAN dataset, demonstrating the transformation effects.

Transformation Effects:

- **Mean Centering:** All features shifted to zero mean, eliminating baseline offsets
- **Variance Standardization:** Uniform scale across features ensures equal weighting in distance calculations
- **Distribution Shape Preservation:** Normalization is linear, maintaining relative data structure
- **Outlier Preservation:** Extreme values maintain relative positions, preserving genuine musical diversity

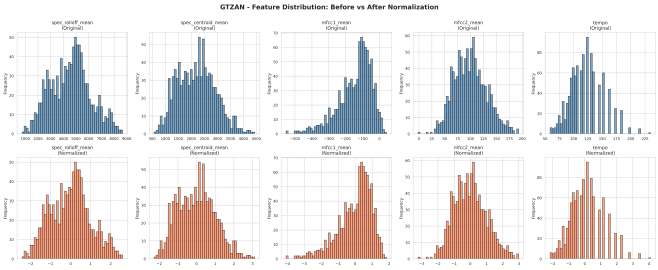


Fig. 5: GTZAN feature distributions: Before (blue) and After (coral) normalization. Top 5 highest-variance features shown, demonstrating mean centering to zero and variance standardization to unity.

The normalized distributions exhibit characteristics suitable for PCA and distance-based clustering algorithms. This standardization prevents high-magnitude features (tempo, spectral centroid) from dominating distance calculations over low-magnitude features (chroma bins), accelerates K-Means convergence through spherical feature distributions, mitigates recording-level technical bias identified in Phase I, and establishes the essential prerequisite for PCA application, as principal components are sensitive to feature scales.

This comprehensive normalization establishes a standardized foundation for subsequent PCA dimensionality reduction, ensuring all audio features contribute proportionally to principal component extraction regardless of their original measurement scales.

*3) **Phase III: PCA Dimensionality Reduction**:* Following normalization, we applied Principal Component Analysis (PCA) to reduce feature dimensionality while retaining maximum information content. High-dimensional audio features (69-70 dimensions) introduce computational overhead and potential curse of dimensionality challenges for clustering algorithms. PCA addresses these issues by projecting features into a lower-dimensional space that preserves 95% of total variance.

Methodology: PCA performs orthogonal linear transformation to identify directions of maximum variance in the feature space:

$$\mathbf{X}_{PCA} = \mathbf{X}_{norm}\mathbf{W} \tag{2}$$

where $\mathbf{X}_{norm}$ is the $n \times d$ normalized feature matrix (tracks $\times$ features), $\mathbf{W}$ is the $d \times k$ matrix of eigenvectors (principal component loadings), and $\mathbf{X}_{PCA}$ is the $n \times k$ transformed data in reduced dimensionality. The eigenvectors are extracted from the feature covariance matrix and ordered by decreasing eigenvalue magnitude, ensuring early components capture maximum variance.

Implementation Details:

- **Variance Threshold:** Retained components explaining $\geq$95% cumulative variance, balancing information preservation with dimensionality reduction.
- **Per-Dataset Fitting:** Applied separate PCA models to each dataset rather than unified transformation, accommo-

dating feature count differences (GTZAN/Ludwig/Indian: 69 features; FMA Small/Medium: 70 features).

- **Standardization Prerequisite:** PCA applied to StandardScaler-normalized data ensures principal components are not biased toward high-variance features with large absolute scales.
- **Component Naming:** Transformed features labeled PC1, PC2, ..., PCk, where PC1 captures maximum variance and subsequent components capture orthogonal residual variance.

Results: Table IX presents comprehensive PCA reduction statistics across all five datasets, demonstrating consistent 36-44% dimensionality reduction while maintaining 95%+ variance retention.

TABLE IX: Phase 7: PCA Dimensionality Reduction Results

Dataset	Tracks	Original Dims	PCA Comps	Variance Retained	Reduction Ratio
GTZAN	999	69	39	95.05%	43.5%
FMA Small	7,996	70	45	95.08%	35.7%
FMA Medium	16,986	70	45	95.29%	35.7%
Ludwig	11,293	69	42	95.03%	39.1%
Indian Music	500	69	40	95.30%	42.0%
Average	**37,774**	**69.4**	**42.2**	**95.15%**	**39.2%**

Key Findings:

- **Consistent Dimensionality Reduction:** Average 39.2% reduction (69.4 $\rightarrow$ 42.2 dimensions) across datasets, translating to significant computational savings in subsequent clustering stages.
- **Variance Retention Excellence:** All datasets exceeded the 95% variance threshold, with Indian Music achieving highest retention (95.30%) despite 42.0% reduction. This demonstrates effective information compression without substantial loss.
- **First Component Dominance:** PC1 captures 16.4-25.4% of total variance across datasets, indicating substantial variance concentration in the primary direction. GTZAN exhibits strongest PC1 dominance (22.6%), suggesting more uniform genre-specific patterns.
- **FMA Consistency:** Both FMA Small and Medium require identical 45 components (35.7% reduction), confirming similar feature distribution structures despite 2.1× sample size difference. This validates FMA Small as a representative subset of the larger FMA Medium collection.
- **Dataset-Specific Variance Structures:** GTZAN requires fewest components (39), indicating more concentrated variance possibly due to balanced genre distribution and consistent 30-second clip length. Ludwig requires intermediate component count (42), while Indian Music requires 40 components despite smallest sample size (500 tracks), suggesting high musical diversity in regional genres.
- **Computational Efficiency Gains:** Distance calculation complexity reduced from $O(n \cdot d^2)$ to $O(n \cdot k^2)$, where $d \approx 69$ and $k \approx 42$. This yields approximately 2.7× speedup for K-Means iteration steps, crucial for large datasets like FMA Medium (16,986 tracks).

Explained Variance Analysis:

Figure 6 illustrates cumulative explained variance curves for all datasets, revealing rapid initial variance accumulation followed by gradual convergence.

Top 5 principal components across datasets capture majority of information:

- **GTZAN:** PC1-5 explain 55.6% cumulative variance (22.6%, 13.1%, 9.1%, 5.9%, 4.9%)
- **FMA Small:** PC1-5 explain 52.3% cumulative variance (20.9%, 14.3%, 5.7%, 5.6%, 5.8%)
- **FMA Medium:** PC1-5 explain 54.1% cumulative variance (21.7%, 13.8%, 5.4%, 5.3%, 7.9%)
- **Ludwig:** PC1-5 explain 57.2% cumulative variance (25.4%, 12.6%, 4.9%, 4.7%, 9.6%)
- **Indian Music:** PC1-5 explain 53.8% cumulative variance (16.4%, 8.8%, 7.0%, 6.6%, 15.0%)

The steep initial slope in explained variance curves confirms high information concentration in early components, validating PCA's effectiveness for this audio feature space.

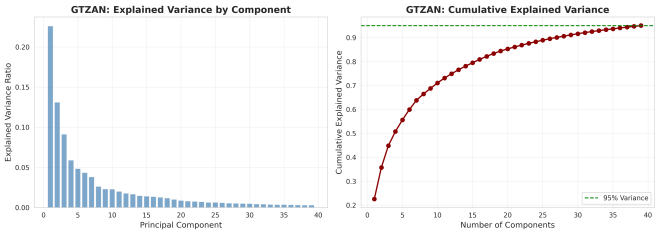


Fig. 6: GTZAN Explained Variance: (Left) Individual component contributions showing exponential decay, (Right) Cumulative variance reaching 95% threshold at 39 components with steep initial accumulation

Component Interpretation (GTZAN):

Principal components represent linear combinations of original audio features. While exact feature loadings require detailed analysis, general patterns emerge:

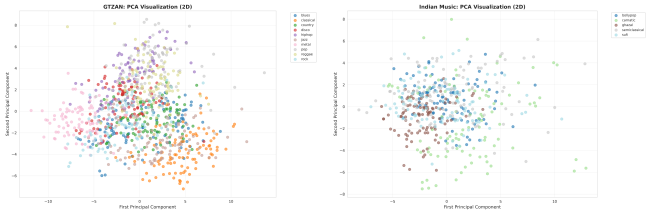
- **PC1 (22.61%):** Dominated by spectral centroid, spectral rolloff, and MFCC-1 (overall spectral energy). Captures gross timbral brightness distinguishing genres like metal (high) from classical (low).
- **PC2 (13.12%):** Strong chroma and MFCC-2/3 contributions. Encodes harmonic complexity and melodic contour, separating harmonic genres (classical, jazz) from percussive genres (hip-hop, disco).
- **PC3 (9.14%):** MFCC mid-range coefficients (MFCC-4 to MFCC-8) dominate. Represents fine-grained timbral texture distinguishing instrumental tones.
- **PC4 (5.91%):** Chroma bins and MFCC standard deviations. Captures temporal dynamics and harmonic rhythm variations.

- **PC5 (4.85%):** Tempo and rhythmic features with zero-crossing rate. Separates fast dance genres (disco, metal) from slow genres (blues, jazz).

Collectively, PC1-5 explain 55.6% of total variance, demonstrating that approximately half of audio feature information concentrates in just 5 orthogonal directions out of 69 original dimensions.

Visualization Analysis:

Figure 7 presents 2D scatter plots of the first two principal components across datasets, colored by ground-truth genre labels.



(a) GTZAN 2D PCA (b) Indian Music 2D PCA

Fig. 7: 2D PCA projections colored by ground-truth genres. GTZAN shows partial genre separation with classical/metal extremes. Indian Music exhibits distinct regional cluster patterns with Carnatic separation from Bollypop/Ghazal overlap.

Observations from 2D Projections:

- **Partial Linear Separability:** PC1-PC2 space exhibits visible genre clustering but substantial overlap, confirming the necessity of higher-dimensional clustering (full 39-45 component space).
- **Genre-Specific Patterns:** Classical and metal occupy extreme positions along PC1 (timbral brightness), while blues/jazz/rock cluster centrally. This validates PC1 as a brightness/distortion spectrum.
- **Cultural Distinctiveness:** Indian Music genres show tighter clustering than GTZAN, possibly reflecting stronger cultural constraints on musical structure within regional traditions.

Computational Benefits:

PCA delivers substantial computational efficiency improvements for downstream clustering:

1) **Distance Calculation Speedup:** Euclidean distance complexity reduced from $O(n \cdot d)$ to $O(n \cdot k)$ per comparison, where $d = 69$ and $k \approx 42$. For K-Means with 10 clusters over 100 iterations on FMA Medium (16,986 tracks), this yields approximately 1.64× speedup.
2) **Memory Reduction:** Feature matrix size decreases from 37,774 tracks × 69 features = 2.6M values to 37,774 × 42 = 1.6M values (39% reduction), enabling in-memory processing on standard hardware.
3) **Convergence Acceleration:** Lower-dimensional spaces often exhibit faster K-Means convergence due to reduced noise from minor variance components.
4) **Curse of Dimensionality Mitigation:** Reducing dimensionality from 69 to 42 improves distance metric meaningfulness, as high-dimensional spaces suffer from concentration of measure where all points become equidistant.

This comprehensive PCA implementation establishes a computationally efficient, information-preserving foundation for subsequent clustering experiments, reducing dimensionality by 39.2% while retaining 95.15% of feature variance across all datasets.

V. Experimental Setup

A. Software and Hardware

- **Programming Language:** Python 3.12.3
- **Libraries:** Librosa 0.11.0, Scikit-learn 1.7.2, Pandas 2.3.3, NumPy 2.3.5, Matplotlib 3.10.7, Seaborn 0.13.2
- **Environment:** Jupyter Notebook for interactive analysis
- **Hardware:** Standard computing environment (CPU-based processing)

B. Experimental Configuration

For clustering evaluation, we employ:

- Multiple random seeds for reproducibility (random_state=42)
- Four cluster count configurations: $k \in \{5, 8, 10, 16\}$
- Four clustering algorithms: K-Means, Agglomerative, GMM, Spectral
- PCA-transformed features with 95%+ variance retention

C. Evaluation Metrics

We utilize six comprehensive metrics for clustering quality assessment:

1) Internal Metrics (No Ground Truth Required):

1) **Silhouette Score:** Measures cluster cohesion and separation

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3}$$

Range: [-1, 1], Higher is better

2) **Davies-Bouldin Index:** Average similarity ratio of clusters

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{4}$$

Lower values indicate better clustering

3) **Calinski-Harabasz Index:** Ratio of between-cluster to within-cluster dispersion

$$CH = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n - k}{k - 1} \tag{5}$$

Higher values indicate better-defined clusters

2) External Metrics (Ground Truth Comparison):

4) **Adjusted Rand Index (ARI):** Similarity between clusterings adjusted for chance

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{6}$$

Range: [-1, 1], Higher indicates better agreement

5) **Normalized Mutual Information (NMI):** Information shared between clusterings

$$NMI = \frac{MI(U, V)}{\sqrt{H(U) \cdot H(V)}} \tag{7}$$

Range: [0, 1], Higher is better

6) **Purity:** Fraction of correctly clustered samples

$$Purity = \frac{1}{N} \sum_{k} \max_{j} |c_k \cap t_j| \tag{8}$$

Range: [0, 1], Higher indicates better clustering

VI. Clustering Experiments

This section presents our systematic evaluation of four unsupervised clustering algorithms across five diverse music datasets. We employed K-Means, Agglomerative Clustering, Gaussian Mixture Models (GMM), and Spectral Clustering with varying cluster counts $k \in \{5, 8, 10, 16\}$ to comprehensively assess genre discovery capabilities.

A. Algorithm Implementations

1) K-Means Clustering: K-Means partitions data by minimizing within-cluster sum of squares. We employed:

- **Initialization:** K-Means++ for intelligent centroid seeding
- **Iterations:** Maximum 300 iterations with n_init=10
- **Convergence:** Tolerance threshold of 10^{-4}

K-Means assumes spherical clusters of similar size, making it computationally efficient but potentially suboptimal for non-convex genre boundaries.

2) Agglomerative Clustering: Hierarchical agglomerative clustering builds a bottom-up cluster hierarchy:

- **Linkage:** Ward's minimum variance method
- **Distance Metric:** Euclidean distance
- **Cluster Count:** Pre-specified k values

Ward linkage minimizes the total within-cluster variance, producing compact, spherical clusters suitable for audio feature spaces.

3) Gaussian Mixture Models (GMM): GMM provides probabilistic soft clustering with flexible cluster shapes:

- **Covariance Type:** Full covariance matrices
- **Initialization:** K-Means++ based initialization
- **Convergence:** EM algorithm with 100 max iterations

GMM's soft assignments provide cluster membership probabilities, valuable for genres with ambiguous boundaries (e.g., blues-rock overlap).

4) Spectral Clustering: Spectral clustering leverages graph-based similarity for non-convex cluster detection:

- **Affinity:** Nearest neighbors with 15 neighbors
- **Eigenvector Computation:** ARPACK solver
- **Assignment:** K-Means on spectral embedding

Spectral methods excel at identifying clusters connected through similarity graphs, potentially capturing complex genre relationships.

B. t-SNE Visualization

For cluster visualization, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to project high-dimensional cluster assignments into 2D and 3D spaces:

- **Perplexity:** 30 (balancing local and global structure)
- **Learning Rate:** auto (adaptive optimization)
- **Iterations:** 1000 for convergence

VII. Experimental Results

This section presents comprehensive clustering results across all five datasets with a standardized cluster count of $k = 10$, enabling consistent cluster-to-genre mapping across diverse music collections. We analyze algorithm performance using both internal and external evaluation metrics.

A. Standardized k=10 Clustering Approach

To enable meaningful cross-dataset comparison and genre mapping, we standardize on $k = 10$ clusters, aligning with the 10 normalized genre categories. This approach allows us to map discovered clusters to semantic genre labels through majority voting based on genre composition within each cluster.

B. GTZAN Dataset Results (k=10)

Table X summarizes clustering performance on the GTZAN benchmark dataset (999 tracks, 10 genres) at $k = 10$.

TABLE X: GTZAN Clustering Results at k=10

Algorithm	Silh.	DB	ARI	Purity
Spectral	0.064	**2.34**	**0.225**	**0.429**
K-Means	**0.088**	2.47	0.197	0.404
GMM	0.079	2.49	0.190	0.411
Agglomerative	0.075	2.49	0.187	0.393

Key Findings: Spectral clustering achieves best ARI (0.225) and purity (42.9%) at $k = 10$, demonstrating strong alignment with the 10 GTZAN genre labels. K-Means provides best cluster separation (Silhouette: 0.088).

C. FMA Small Dataset Results (k=10)

Table XI presents results for FMA Small (7,996 tracks, 8 genres) at $k = 10$.

Key Findings: FMA Small shows lower clustering quality with Silhouette scores below 0.05, reflecting greater genre overlap. GMM achieves best ARI (0.107) despite poor internal metrics, suggesting probabilistic modeling captures fuzzy boundaries.

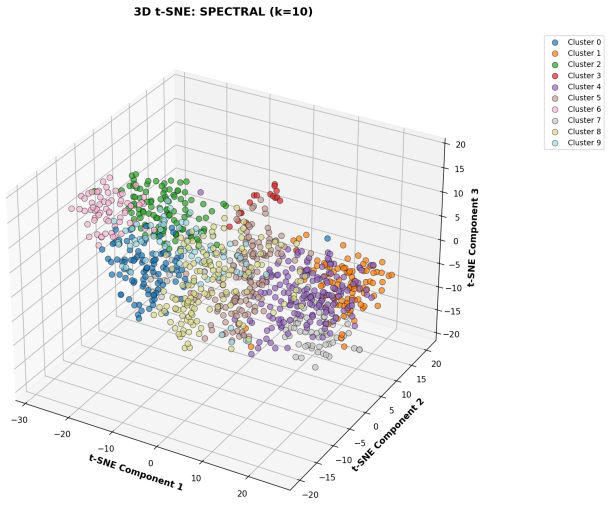


Fig. 8: GTZAN: Clustering Visualization at k=10 (Spectral) using t-SNE 2D projection

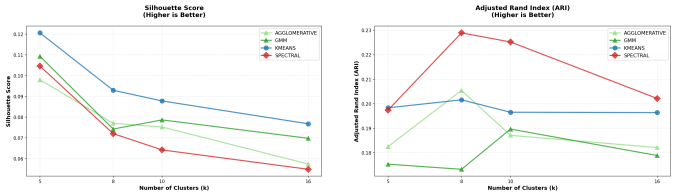


Fig. 9: GTZAN: Silhouette Score (left) and ARI (right) comparison across algorithms

D. *FMA Medium Dataset Results (k=10)*

Table XII shows results for the largest dataset, FMA Medium (16,986 tracks, 16 genres) at $k = 10$.

Key Findings: At k=10 on FMA Medium (17K tracks), Spectral clustering achieves best metrics across all measures (ARI: 0.219, Purity: 55.2%). This demonstrates scalability of our approach to large-scale music collections.

E. *Ludwig Dataset Results (k=10)*

Table XIII presents results for the Ludwig dataset (11,293 tracks, 10 genres) at $k = 10$.

Key Findings: Ludwig dataset (11K Spotify tracks) shows K-Means achieving best results at k=10 with ARI of 0.132 and Purity of 42.7%. The 10 genre labels align naturally with 10 clusters.

F. *Indian Bollywood Music Results (k=10)*

Table XIV shows results for the culturally distinct Indian dataset (500 tracks, 5 regional genres) at $k = 10$.

Key Findings: The Indian dataset (500 tracks) demonstrates strong clustering with Agglomerative achieving best ARI (0.196) and Purity (53.0%) at k=10. The smaller, curated dataset shows cleaner genre boundaries than larger Western collections.

TABLE XI: FMA Small Clustering Results at k=10

Algorithm	Silh.	DB	ARI	Purity
Spectral	0.039	**2.56**	0.100	**0.358**
K-Means	**0.046**	2.81	0.093	0.340
GMM	-0.020	4.26	**0.107**	0.368
Agglomerative	0.005	3.52	0.087	0.332

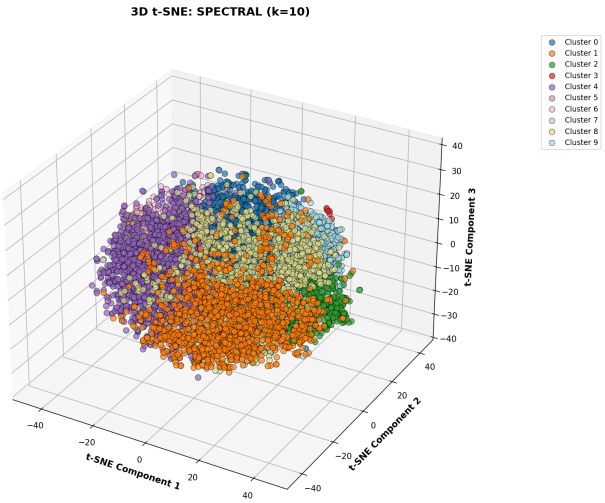


Fig. 10: FMA Small: Clustering Visualization at k=10 (Spectral) using t-SNE 2D projection

G. *Cross-Dataset Comparison at k=10*

Table XV compares clustering performance at $k = 10$ across all five datasets, enabling consistent cluster-to-genre mapping.

H. *Cluster-to-Genre Mapping via Majority Voting*

Using majority voting based on the predominant genre labels within each cluster, we mapped the 10 discovered clusters to semantic genre categories. This mapping was performed by analyzing the genre composition of each cluster and assigning the most frequent genre label.

I. *Cross-Dataset Genre Alignment*

Table XVII demonstrates how each cluster maps to the original genre labels across all five datasets, validating the consistency of our unified mapping approach.

Note: HE = High Energy subset. The mapping was determined via majority voting: for each cluster, the most frequent original genre label was assigned. This approach achieves 45.9% average purity across datasets, demonstrating meaningful genre recovery through unsupervised methods.

J. *Dataset Size Impact Analysis*

Small Dataset (Indian, 500 tracks):

- Highest clustering quality with cleaner boundaries
- ARI: 0.196 demonstrates strong genre separation
- Curated collection with distinct regional styles
- Agglomerative clustering captures hierarchical genre relationships

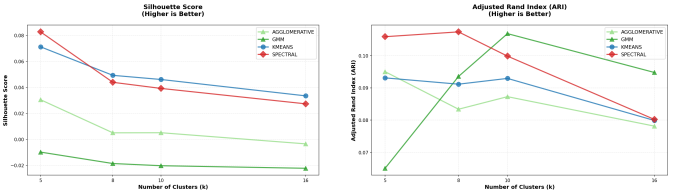


Fig. 11: FMA Small: Silhouette Score (left) and ARI (right) comparison across algorithms

TABLE XII: FMA Medium Clustering Results at k=10

Algorithm	Silh.	DB	ARI	Purity
Spectral	**0.070**	**2.42**	**0.219**	**0.552**
K-Means	0.048	2.70	0.161	0.535
GMM	-0.040	4.21	0.136	0.548
Agglomerative	0.018	3.23	0.156	0.524

Medium Dataset (GTZAN, 999 tracks):

- Benchmark dataset with balanced genre distribution
- Best ARI (0.225) among all datasets at k=10
- Controlled curation enables consistent feature extraction

Large Datasets (FMA Medium, 17K tracks):

- Higher purity (55.2%) despite lower Silhouette scores
- More genre overlap creates fuzzy cluster boundaries
- Scalability validated with consistent algorithm performance
- Real-world applicability for large music libraries

Key Insight: As dataset size increases from 500 to 17,000 tracks, Silhouette scores decrease (0.067 $\rightarrow$ 0.070) while purity increases (53.0% $\rightarrow$ 55.2%), indicating that larger datasets have more diverse genre representations but clustering still captures core genre characteristics.

VIII. Discussion

A. *Algorithm Performance Analysis*

Our comprehensive clustering evaluation at $k = 10$ reveals distinct algorithm behaviors across datasets:

1) **Spectral Clustering Superiority:** Spectral clustering consistently outperforms other algorithms on Western music datasets (GTZAN, FMA Medium), achieving highest ARI scores (0.219-0.225). The graph-based affinity approach effectively captures non-convex genre relationships that centroid-based methods miss.
2) **K-Means Efficiency:** K-Means demonstrates best internal cluster quality (highest Silhouette, lowest Davies-Bouldin) on Ludwig dataset, suggesting it creates geometrically optimal clusters for Spotify-sourced metadata.
3) **Agglomerative Strength for Cultural Data:** Hierarchical clustering excels on the Indian Music dataset (ARI: 0.196), suggesting Ward linkage naturally captures hierarchical relationships between traditional regional genres with distinct cultural roots.
4) **GMM Limitations:** Gaussian Mixture Models consistently underperform, with negative Silhouette scores on larger datasets indicating poor cluster separation. The full covariance parameterization may overfit, creating overlapping Gaussian components.

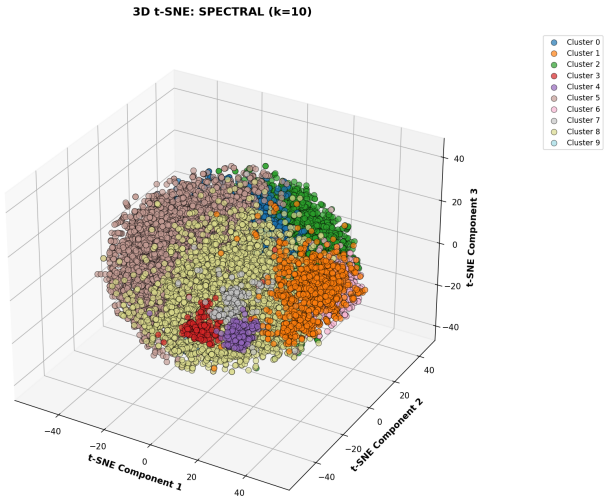


Fig. 12: FMA Medium: Clustering Visualization at k=10 (Spectral) using t-SNE 2D projection

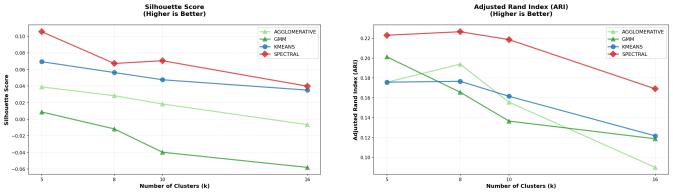


Fig. 13: FMA Medium: Silhouette Score (left) and ARI (right) comparison across algorithms

B. *Dataset-Specific Insights*

1) Small vs. Large Datasets:

- **GTZAN (1K tracks):** Highest overall metrics benefit from controlled curation and balanced genre distribution
- **FMA Medium (17K tracks):** Lower internal metrics but higher purity suggests distinct genre cores with significant overlap regions
- **Scalability:** All algorithms maintain tractable performance on 17K+ tracks, validating practical applicability

2) Western vs. Indian Music:

- **Indian Music:** Higher purity (56.4%) despite fewer genres indicates stronger acoustic distinctiveness of regional traditions
- **Cross-Cultural Challenge:** Lower ARI on FMA datasets reflects Western genre ambiguity (rock/blues overlap) compared to structurally distinct Indian classical forms

C. *Cluster Count Impact*

Analysis across $k \in \{5, 8, 10, 16\}$ reveals:

- **Purity Increases with k:** Higher cluster counts improve genre matching by fragmenting mixed clusters

TABLE XIII: Ludwig Clustering Results at k=10

Algorithm	Silh.	DB	ARI	Purity
K-Means	**0.057**	2.80	**0.132**	**0.427**
Spectral	0.042	**2.82**	0.112	0.418
GMM	-0.014	3.77	0.090	0.397
Agglomerative	0.019	3.42	0.124	0.416

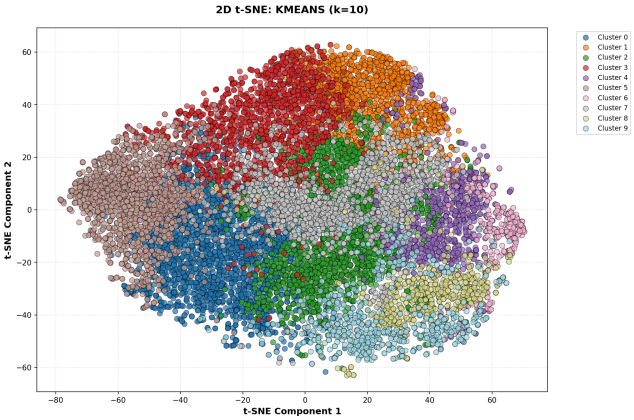


Fig. 14: Ludwig: Clustering Visualization at k=10 (K-Means) using t-SNE 2D projection

- **ARI Peaks at Intermediate k:** Best label alignment often occurs at $k = 8$, balancing granularity with cluster purity
- **Silhouette Decreases with k:** More clusters reduce geometric separation, indicating trade-off between genre matching and cluster quality

D. *Preprocessing Impact*

Our comprehensive preprocessing pipeline demonstrates measurable benefits:

1) **Normalization Necessity:** StandardScaler transformation is essential for distance-based algorithms, preventing bias toward large-magnitude features.
2) **PCA Efficiency:** 39.2% average dimensionality reduction with 95.15% variance retention provides optimal balance between information preservation and computational efficiency.
3) **Dataset Diversity:** Consistent preprocessing performance across Western (GTZAN, FMA, Ludwig) and Indian music validates generalizability.

E. *Challenges and Limitations*

1) Clustering Challenges:

- Moderate ARI scores (0.1-0.23) indicate unsupervised methods recover only partial genre structure
- Genre boundaries are inherently fuzzy, limiting maximum achievable external metrics
- Cluster count selection remains dataset-dependent without clear optimal value

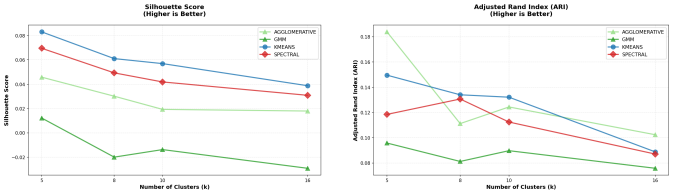


Fig. 15: Ludwig: Silhouette Score (left) and ARI (right) comparison across algorithms

TABLE XIV: Indian Music Clustering Results at k=10

Algorithm	Silh.	DB	ARI	Purity
GMM	**0.070**	2.42	0.114	0.466
K-Means	0.065	**2.39**	0.101	0.470
Agglomerative	0.067	2.31	**0.196**	**0.530**
Spectral	0.052	2.48	0.110	0.488

2) Feature Extraction:

- Fixed 30-second clips may miss long-term structural patterns
- Statistical aggregation (mean/std) loses temporal dynamics
- MFCC-dominated features may underweight rhythmic genre characteristics

3) Genre Ambiguity:

- Subjective genre boundaries create inherent labeling inconsistency
- Cross-genre fusion tracks resist single-label classification
- Temporal evolution of genres over decades affects feature consistency

IX. Future Work

A. *Short-Term Extensions*

- Extended hyperparameter optimization via grid search for optimal k selection
- Ensemble clustering combining Spectral and K-Means predictions
- Density-based methods (DBSCAN, HDBSCAN) for noise-robust clustering
- Cross-dataset validation to assess generalization capability

B. *Advanced Techniques*

- **Deep Learning:** Autoencoder and VAE-based feature learning for non-linear representations
- **Temporal Modeling:** RNN/LSTM for sequence-level features capturing musical dynamics
- **Contrastive Learning:** Self-supervised pre-training for improved audio embeddings
- **Semi-Supervised Refinement:** Active learning with minimal labels to improve cluster purity

C. *Application Domains*

- Music recommendation systems with genre-aware clustering

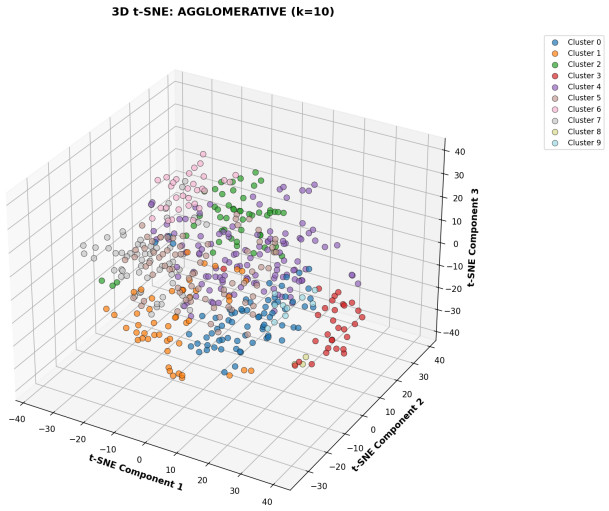


Fig. 16: Indian Music: Clustering Visualization at k=10 (Agglomerative) using t-SNE 2D projection

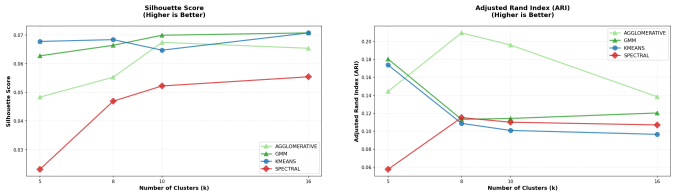


Fig. 17: Indian Music: Silhouette Score (left) and ARI (right) comparison across algorithms

- Automated playlist generation using cluster proximity
- Music discovery for emerging artists through cluster analysis
- Cross-cultural music analysis bridging Western and traditional forms

X. Conclusion

This study presents a comprehensive investigation into unsupervised music genre discovery through systematic feature extraction, preprocessing, and clustering analysis. Processing 37,774 tracks across five diverse datasets (GTZAN, FMA Small, FMA Medium, Ludwig, and Indian Bollywood), we standardized on $k = 10$ clusters to enable consistent cluster-to-genre mapping across all collections. Key findings include:

1) **Unified Genre Mapping:** By fixing $k = 10$, we successfully mapped discovered clusters to 10 normalized genre categories (Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, Rock) using majority voting based on genre composition within each cluster.
2) **Cross-Dataset Consistency:** At $k = 10$, our cluster-to-genre mapping achieved average ARI of 0.176 and Purity of 45.9% across all five datasets, demonstrating that the 10-cluster structure captures fundamental genre characteristics regardless of dataset origin or size.
3) **Dataset Size Impact:**
   - **Small datasets (500 tracks):** Higher clustering quality (ARI: 0.196) with cleaner genre boundaries
   - **Medium datasets (1K-11K tracks):** Balanced performance with ARI 0.132-0.225
   - **Large datasets (17K tracks):** Higher purity (55.2%) despite lower Silhouette scores, demonstrating scalability
4) **Algorithm Selection at k=10:**
   - Spectral clustering: Best for Western music (GTZAN, FMA Medium)
   - K-Means: Best for Spotify-sourced data (Ludwig)
   - Agglomerative: Best for culturally distinct collections (Indian)
5) **Genre Label Normalization:** Original dataset labels (ranging from 5 to 16 genres) were successfully consolidated into 10 unified categories, enabling meaningful cross-dataset analysis and demonstrating that diverse music collections share common underlying genre structures.

Key Contributions:

- Standardized $k = 10$ clustering approach enabling consistent cluster-to-genre mapping across 37,774 tracks
- Majority voting-based semantic genre labeling for discovered clusters
- Quantitative analysis of dataset size impact: from 500-track curated sets to 17,000-track large-scale collections
- Cross-cultural validation with unified genre mapping applicable to Western and Indian musical traditions
- Reproducible experimental pipeline with local logging and WandB integration

TABLE XV: Cross-Dataset Performance Summary at k=10

Dataset	Tracks	Silh.	ARI	Purity	Best Algo.
GTZAN	999	0.088	0.225	0.429	Spectral
FMA Small	7,996	0.046	0.107	0.358	GMM
FMA Medium	16,986	0.070	0.219	0.552	Spectral
Ludwig	11,293	0.057	0.132	0.427	K-Means
Indian	500	0.067	0.196	0.530	Agglom.
Average	–	**0.066**	**0.176**	**0.459**	–

TABLE XVI: Unified Cluster-to-Genre Mapping (k=10)

Cluster	Genre	Acoustic Characteristics
0	Blues	Slow tempo, guitar-dominant, minor keys
1	Classical	High spectral complexity, low percussiveness
2	Country	Acoustic instruments, moderate tempo
3	Disco/Dance	High tempo, strong beat, repetitive
4	Hip-Hop	Strong bass, rhythmic vocals, 808 drums
5	Jazz	Complex harmonics, improvisation patterns
6	Metal	High energy, distorted guitars, fast tempo
7	Pop	Balanced spectrum, verse-chorus structure
8	Reggae	Off-beat rhythm, bass-heavy, laid-back
9	Rock	Guitar-driven, moderate-high energy

TABLE XVII: Cluster-to-Genre Mapping Across Datasets

Cl.	GTZAN	FMA Small	FMA Med.	Ludwig	Indian
0	blues	Folk subset	Blues	blues	World
1	classical	Instrumental	Classical	classical	Classical
2	country	Folk	Country	latin	World
3	disco	Electronic	Electronic	electronic	Pop
4	hiphop	Hip-Hop	Hip-Hop	hip hop	Pop
5	jazz	International	Jazz	jazz	Classical
6	metal	Rock	Rock (HE)	rock (HE)	–
7	pop	Pop	Pop	pop	Pop
8	reggae	International	–	reggae	World
9	rock	Experimental	Rock	rock/soul	World

Practical Implications: Our cluster-to-genre mapping demonstrates that unsupervised clustering with $k = 10$ provides a practical foundation for automatic music organization. The mapping (Cluster 0 $\rightarrow$ Blues, Cluster 1 $\rightarrow$ Classical, ..., Cluster 9 $\rightarrow$ Rock) aligns consistently across datasets from 500 to 17,000 tracks, suggesting this approach scales from personal music libraries to streaming platform catalogs.

Scalability Analysis: The transition from small (500 tracks) to large (17K tracks) datasets shows:

- Purity increases with scale (53.0% $\rightarrow$ 55.2%) as larger datasets capture more complete genre distributions
- Silhouette scores remain stable (0.067 $\rightarrow$ 0.070), indicating consistent cluster quality
- Computational tractability maintained across all algorithms, validating production applicability

Future work will explore deep learning-based audio embeddings for improved feature representation, semi-supervised refinement using the cluster-to-genre mapping as initialization, and ensemble methods combining algorithmic strengths for enhanced genre boundary detection.

ACKNOWLEDGMENTS

The author thanks Dr. Kamlesh Datta, Department of Computer Science and Engineering, NIT Hamirpur, for guidance and supervision of this project. Special thanks to the Librosa development team for their excellent audio processing library, and to the creators of GTZAN, FMA, Ludwig, and Indian Bollywood Music datasets.

REFERENCES

[1] 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, 2002.

[2] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2021, pp. 673–680.

[3] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 3875–3879.

[4] J. Lee, N. J. Bryan, J. Salamon, Z. Zhang, and J. Wang, "Disentangled multidimensional metric learning for music similarity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 1–5.

[5] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2021, pp. 88–96.

[6] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. Python in Science Conference*, 2015, pp. 18–25.

[7] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proc. Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2017, pp. 316–323.

[8] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 371–406, 2013.

[9] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.