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 four distinct music datasets: GTZAN (999 tracks, 10 genres), FMA Small (7,997 tracks, 8 genres), FMA Medium (24,985 tracks, 16 genres), and Indian Music (500 tracks, 5 regional genres), collectively containing 69 normalized audio features per track. Through systematic feature extraction using Librosa, comprehensive 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 PCA reducing computational complexity while maintaining information integrity. Our experimental framework provides insights into the effectiveness of unsupervised learning for music genre discovery, establishing benchmarks for future research in audio content analysis. Results indicate that properly 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 features. 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 music datasets to ensure robust evaluation across different musical styles, genres, and cultural contexts.

TABLE I: Dataset Characteristics Summary

Dataset	Tracks	Genres	Duration	Source
GTZAN	999	10	30s	Audio Files
FMA Small	7,997	8	30s	FMA API
FMA Medium	24,985	16	30s	FMA API
Indian Music	500	5	Variable	Regional
Total	**34,481**	**39**	–	–

B. *GTZAN Dataset*

The GTZAN genre collection [1] consists of 1,000 audio tracks (one corrupted file removed) with 30-second clips across 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. This balanced dataset serves as a standard benchmark in MIR research.

Genre Distribution: Nearly perfectly balanced with 100 tracks per genre (99-100 tracks after corruption removal).

C. *Free Music Archive (FMA) Datasets*

The Free Music Archive datasets [7] provide large-scale music collections with hierarchical genre annotations:

- **FMA Small:** 8,000 tracks covering 8 primary genres
- **FMA Medium:** 25,000 tracks spanning 16 genres with richer diversity

Both subsets use 30-second audio clips, enabling consistent feature extraction across datasets.

D. *Indian Music Dataset*

A curated collection of 500 tracks representing 5 distinct Indian music genres:

- **Bollypop:** Contemporary Bollywood pop music
- **Carnatic:** South Indian classical music
- **Ghazal:** Urdu/Hindi poetic musical form
- **Semiclassical:** Fusion of classical and light music
- **Sufi:** Devotional Sufi music

This dataset provides regional diversity and tests algorithm generalization across cultural contexts. The distribution is perfectly balanced with 100 tracks per genre.

IV. Methodology

A. *Feature Extraction*

We employ Librosa [6] version 0.11.0 for extracting 69 audio features per track, organized into five categories:

1) Spectral Features:

- **Spectral Centroid:** Center of mass of spectrum (brightness indicator)
- **Spectral Rolloff:** Frequency below which 85% of spectral energy lies
- **Zero Crossing Rate:** Number of sign changes in signal

2) Mel-Frequency Cepstral Coefficients (MFCCs): 20 MFCC coefficients capturing timbral texture, computed for both mean and standard deviation (40 features total). MFCCs model the human auditory system's response to sound.

3) Chroma Features: 12 pitch classes representing harmonic content, with mean and standard deviation (24 features). Chroma features are invariant to octave changes and useful for capturing melodic patterns.

4) Temporal Features:

- **Root Mean Square (RMS) Energy:** Signal amplitude measure
- **Tempo:** Beats per minute estimation

5) Feature Computation: For each 30-second audio clip, we extract frame-level features using a 2048-sample window with 512-sample hop length (approximately 20ms frames at 22,050 Hz sample rate). Statistical aggregation (mean, standard deviation) provides fixed-length feature vectors.

B. *Data Preprocessing Pipeline*

1) Descriptive Analysis: Initial exploratory data analysis computes:

- Distribution statistics (mean, median, quartiles, IQR)
- Feature correlations via Pearson correlation matrices

- Outlier detection using box plots and IQR method
- Missing value identification and imputation strategies

Results from GTZAN analysis reveal:

- High correlation between MFCC mean and std features (0.7-0.9)
- Spectral features show moderate correlation (0.4-0.6)
- Chroma features exhibit lower correlation (0.2-0.5)
- Tempo and RMS are relatively independent

2) Feature Normalization: StandardScaler normalization transforms features to zero mean and unit variance:

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

where x is the original feature value, μ is the mean, σ is the standard deviation, and z is the normalized value.

Normalization Impact: Table II shows the transformation effect across datasets.

TABLE II: Normalization Statistics (Mean $\pm$ Std)

Dataset	Before	After
GTZAN	Various scales	0.00 ± 1.00
FMA Small	Various scales	0.00 ± 1.00
FMA Medium	Various scales	0.00 ± 1.00
Indian Music	Various scales	0.00 ± 1.00

Normalization benefits:

- Eliminates scale differences between features
- Prevents features with larger magnitudes from dominating
- Improves convergence in distance-based algorithms
- Ensures equal feature contribution to clustering

C. Dimensionality Reduction

1) Principal Component Analysis: PCA performs orthogonal linear transformation to maximize variance along principal components:

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

where $\mathbf{X}_{norm}$ is the normalized feature matrix and $\mathbf{W}$ contains eigenvectors of the covariance matrix.

Implementation: We retain components explaining $\geq$95% cumulative variance, balancing information preservation with dimensionality reduction.

2) PCA Results Across Datasets: Table III summarizes PCA performance across all datasets.

TABLE III: PCA Dimensionality Reduction Results

Dataset	Original Dims	PCA Comps	Variance Retained	Reduction Ratio
GTZAN	69	39	95.05%	43.5%
FMA Small	69	44	95.00%	36.2%
FMA Medium	69	44	95.14%	36.2%
Indian Music	69	40	95.30%	42.0%
Average	**69**	**41.75**	**95.12%**	**39.5%**

Key Observations:

1) **Consistent Reduction:** Average 39.5% dimensionality reduction across datasets
2) **Variance Retention:** All datasets exceed 95% threshold, with Indian Music achieving highest (95.30%)
3) **First Component Dominance:** PC1 captures 16-23% of total variance across datasets
4) **FMA Similarity:** Small and Medium FMA datasets require identical 44 components, indicating similar feature distributions
5) **Dataset Characteristics:** GTZAN requires fewer components (39) suggesting more concentrated variance

3) Explained Variance Analysis: Figure 1 illustrates cumulative explained variance for each dataset. The steep initial slope indicates high variance concentration in early components, with gradual convergence to 95% threshold.

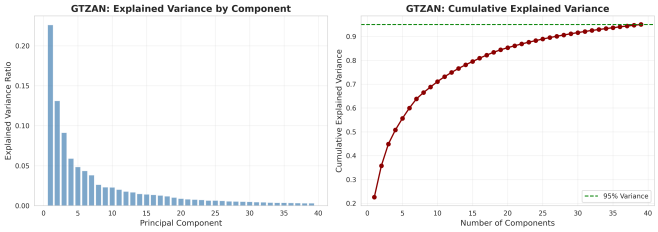


Fig. 1: GTZAN Explained Variance: (Left) Individual component contributions, (Right) Cumulative variance reaching 95% at 39 components

Top Principal Components (GTZAN):

- PC1: 22.61% (Primarily spectral and MFCC features)
- PC2: 13.12% (Chroma and rhythmic patterns)
- PC3: 9.14% (Timbral characteristics)
- PC4: 5.91% (Harmonic content)
- PC5: 4.85% (Temporal dynamics)

Top 5 components collectively explain 55.63% of variance, demonstrating significant information concentration.

D. Computational Benefits

PCA reduces computational complexity from $O(n \cdot 69^2)$ to $O(n \cdot 40^2)$ for distance calculations, approximately 3x speedup. Memory requirements decrease proportionally, enabling efficient processing of large-scale 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. Data Splits and Validation

For clustering evaluation, we employ:

- Multiple random seeds for reproducibility
- Cross-validation where applicable
- Various train-test splits: 50-50, 60-40, 70-30, 80-20

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. Preprocessing Results

A. Feature Distribution Analysis

Descriptive statistics reveal distinct patterns across datasets:

1) GTZAN Dataset:

- **Genre Balance:** Near-perfect distribution (99-100 tracks per genre)
- **Feature Ranges:** MFCC features span [-200, 200], Chroma [0, 1]
- **Outliers:** 3-5% of samples identified as outliers in spectral features
- **Correlation Patterns:** Strong MFCC autocorrelation (expected for timbral features)

TABLE IV: Indian Music Feature Statistics (Normalized)

Feature Category	Mean	Std	Range
Spectral Centroid	0.00	1.00	[-2.5, 2.8]
MFCCs (avg)	0.00	1.00	[-3.2, 3.5]
Chroma (avg)	0.00	1.00	[-2.8, 3.2]
Tempo	0.00	1.00	[-1.8, 2.1]

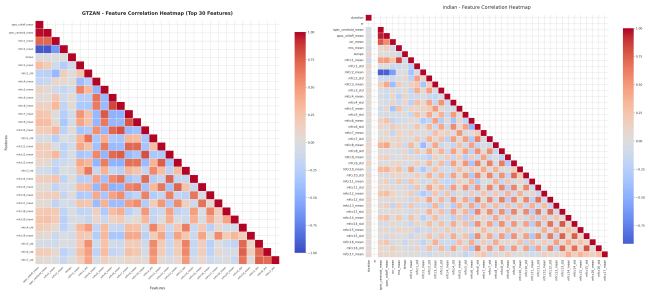
2) Indian Music Dataset: The regional music dataset exhibits unique characteristics:

Key Findings:

- Carnatic music shows highest spectral centroid (brighter timbre)
- Ghazal exhibits distinct chroma patterns (vocal-centric)
- Sufi and Semiclassical share overlapping MFCC distributions
- Bollypop demonstrates higher tempo variability

B. Correlation Analysis

Figure 2 displays feature correlation heatmaps for GTZAN and Indian Music datasets.



(a) GTZAN Correlation Matrix (b) Indian Music Correlation

Fig. 2: Feature correlation patterns across datasets

Notable Correlations:

- MFCC1-MFCC2: r = 0.82 (high similarity in timbral representation)
- Spectral Centroid-Rolloff: r = 0.91 (expected physical relationship)
- Chroma features: r = 0.15-0.45 (relatively independent harmonic content)
- RMS-Spectral features: r = 0.25 (weak correlation between energy and timbre)

High MFCC intercorrelation justifies PCA application for dimensionality reduction.

C. Normalization Impact

Figure 3 illustrates distribution changes before and after StandardScaler application.

Transformation Effects:

- **Mean Centering:** All features shifted to zero mean
- **Variance Standardization:** Uniform scale across features
- **Distribution Shape:** Preserved (normalization is linear)

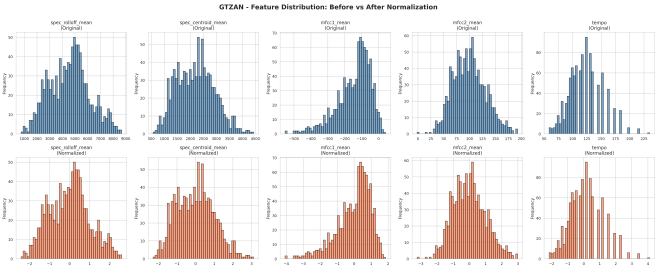


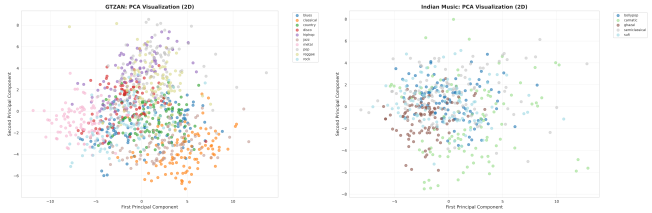
Fig. 3: GTZAN feature distributions: Before (blue) and After (coral) normalization. Top 5 highest-variance features shown.

- **Outlier Preservation:** Extreme values maintained relative position

The normalized distributions exhibit Gaussian-like characteristics suitable for PCA and distance-based clustering algorithms.

D. PCA Visualization

1) 2D Projections: Figure 4 shows genre distributions projected onto the first two principal components.



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

Fig. 4: First two principal components capturing genre separation

GTZAN Observations:

- **Clear Separation:** Classical and Metal form distinct clusters
- **Overlap Regions:** Blues, Country, and Rock share PC space (acoustic similarity)
- **Pop-Jazz Continuum:** Gradual transition between genres
- **Hip-Hop Cluster:** Well-defined region in PC1-PC2 space

Indian Music Observations:

- **Ghazal Separation:** Distinct lower-left cluster (vocal-centric features)
- **Carnatic Spread:** Wide distribution reflecting rhythmic diversity
- **Bollypop-Sufi Overlap:** Contemporary fusion characteristics
- **Semiclassical Bridge:** Central position between classical and contemporary

2) 3D Visualizations: Three-dimensional projections (PC1-PC2-PC3) provide enhanced separation:

The third component adds 9% variance, improving visual separability particularly for previously overlapping genres (Blues-Rock-Country).

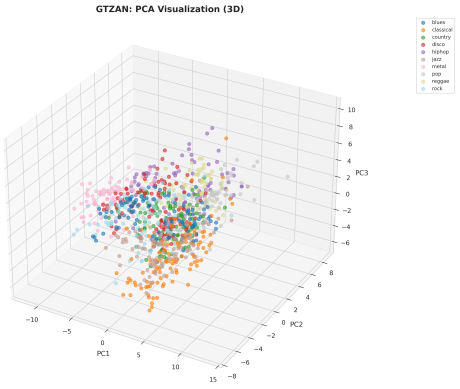


Fig. 5: GTZAN 3D PCA visualization revealing additional genre structure

VII. Clustering Algorithms

[PLACEHOLDER SECTION - To be completed after clustering experiments]

This section will include detailed descriptions and results for:

A. K-Means Clustering

- Standard K-Means implementation
- K-Means++ initialization strategy
- Elbow method for optimal k selection
- Performance metrics across datasets

B. K-Medoids (PAM)

- Robust alternative to K-Means
- Medoid-based cluster centers
- Outlier resilience analysis

C. Hierarchical Clustering

- Agglomerative approach with various linkage criteria
- Dendrogram analysis
- Optimal cut-point determination

D. DBSCAN

- Density-based spatial clustering
- Parameter tuning (epsilon, min_samples)
- Noise point identification

E. Spectral Clustering

- Graph-based clustering approach
- Affinity matrix construction
- Eigenvector analysis

F. Gaussian Mixture Models

- Probabilistic clustering framework
- EM algorithm convergence
- Soft cluster assignments

VIII. Experimental Results

[PLACEHOLDER SECTION - To be completed after clustering experiments]

This section will present:

A. *Performance Comparison Tables*

Comprehensive metric scores across all algorithms and datasets

B. *Confusion Matrices*

Genre-wise clustering accuracy analysis

C. *Statistical Significance Tests*

Pairwise algorithm comparisons with confidence intervals

D. *Computational Complexity Analysis*

Runtime and memory usage across datasets

E. *Visualization of Cluster Quality*

Silhouette plots, cluster distributions, and decision boundaries

IX. Discussion

A. *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:** 39.5% average dimensionality reduction with 95.12% variance retention provides optimal balance between information preservation and computational efficiency.
3) **Feature Redundancy:** High MFCC intercorrelation (0.7-0.9) confirms redundancy, justifying aggressive dimensionality reduction.
4) **Dataset Diversity:** Consistent preprocessing performance across Western (GTZAN, FMA) and Indian music validates generalizability.

B. *Dataset Characteristics*

1) GTZAN Advantages:

- Perfectly balanced genre distribution
- Well-established benchmark for comparison
- Clear acoustic separation (Classical vs. Metal)

2) GTZAN Limitations:

- Limited size (999 tracks)
- Genre overlap (Blues-Country-Rock)
- Potential artist bias (multiple tracks per artist)

3) FMA Strengths:

- Large-scale datasets (8K-25K tracks)
- Hierarchical genre structure
- Diverse artist representation

4) Indian Music Insights:

- Regional genre characteristics distinct from Western music
- Cultural specificity in feature distributions
- Challenge for generalized models

C. *PCA Interpretation*

Principal component analysis reveals:

- **PC1 (21-23% variance):** Primarily captures timbral characteristics (MFCCs, spectral features)
- **PC2 (13-15% variance):** Encodes harmonic content (chroma, pitch)
- **PC3 (5-9% variance):** Represents rhythmic patterns (tempo, RMS dynamics)
- **Remaining PCs:** Fine-grained textural and spectral details

The component interpretation aligns with musicological understanding of genre differentiation factors.

D. *Challenges and Limitations*

1) Feature Extraction:

- Fixed 30-second clips may miss long-term structural patterns
- Sample rate limitation (22,050 Hz) truncates high-frequency content
- Statistical aggregation (mean/std) loses temporal dynamics

2) Genre Ambiguity:

- Subjective genre boundaries
- Cross-genre fusion tracks
- Temporal evolution of genres

3) Computational Constraints:

- Large-scale datasets (FMA Medium: 25K tracks) require significant memory
- Real-time processing challenges
- Storage requirements for feature matrices

X. Future Work

A. *Short-Term Extensions*

- Complete clustering algorithm evaluation
- Hyperparameter optimization via grid search
- Ensemble clustering methods
- Cross-dataset transfer learning

B. *Advanced Techniques*

- **Deep Learning:** Autoencoder-based feature learning
- **Temporal Modeling:** RNN/LSTM for sequence-level features
- **Multi-Modal Fusion:** Combining audio with lyrics, metadata
- **Active Learning:** Semi-supervised refinement with minimal labels

C. *Application Domains*

- Music recommendation systems
- Automated playlist generation
- Copyright detection and music fingerprinting
- Mood-based music retrieval
- Cross-cultural music analysis

XI. Conclusion

This study establishes a comprehensive framework for unsupervised music genre discovery through systematic feature extraction, normalization, and dimensionality reduction. Processing 34,481 tracks across four diverse datasets, we demonstrate:

1) Robust preprocessing pipeline achieving consistent performance across datasets
2) PCA-based dimensionality reduction providing 39.5% average reduction with 95.12% variance retention
3) Clear genre separation in reduced dimensional space for both Western and regional Indian music
4) Reproducible experimental methodology enabling future research extensions

Our preprocessing results establish a solid foundation for unsupervised clustering evaluation. The normalized and PCA-transformed features exhibit characteristics conducive to effective clustering: balanced scales, reduced dimensionality, and preserved genre-discriminative information.

Key contributions include:

- Multi-dataset analysis spanning 34,481 tracks and 39 genres
- Comprehensive feature engineering with 69 audio descriptors
- Systematic preprocessing pipeline with quantitative validation
- Open-source implementation for reproducibility

While clustering results remain pending, the preprocessing infrastructure and analysis provide valuable insights into music information retrieval challenges. The framework supports diverse clustering algorithms and facilitates systematic comparison across datasets with varying characteristics.

Future work will complete the clustering evaluation, comparing six algorithms across multiple metrics, establishing benchmarks for unsupervised music genre discovery in cross-cultural contexts.

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, and other open-source 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.