Spotify Prediction Using Audio Analysis and Clustering

Stephen Ace F. Sy
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
sysf@students.national-u.edu.ph

James Adrian Castro
College of Computing and Information Technologies
National University – Philippines
Manila, Philippines
castrojb@students.national-u.edu.ph

***Abstract*—This study applies unsupervised machine learning techniques within the field of Music Information Retrieval (MIR) to analyze and organize a large-scale audio dataset. Using 32,833 tracks sourced from the Spotify API, songs are clustered based on quantitative audio features—including danceability, energy, and acousticness—without relying on predefined genre labels. Principal Component Analysis (PCA) reduces the dimensionality of the high-dimensional feature space for visualization, while K-Means Clustering groups songs by acoustic similarity. A secondary DBSCAN pass identifies statistical outliers that do not conform to any dominant cluster pattern.**

The study investigates whether human listening contexts—such as gym, studying, upbeat, or melancholic atmospheres—can emerge naturally from data-driven methods alone. Results demonstrate that audio features encode sufficient information to reconstruct mood-based groupings without manual annotation, reflecting the underlying principles of modern recommendation systems such as Spotify's Discover Weekly.

***Index Terms*—music information retrieval, unsupervised learning, K-Means clustering, principal component analysis, DBSCAN, audio features, Spotify**

I. Introduction

The rapid expansion of music streaming platforms has generated large-scale datasets of quantitative audio features describing song characteristics in mathematical terms. This study uses a Kaggle dataset derived from the Spotify Web API containing approximately 32,833 tracks, each represented by numerical audio attributes such as danceability, energy, acousticness, valence, and tempo. These features provide a quantitative proxy for musical properties that listeners typically perceive subjectively.

Traditional genre labels—such as Pop, Rock, or EDM—are assigned by human curators and reflect commercial or historical categories rather than the functional experience of listening. A Pop track, for example, may be a high-energy workout anthem or a slow, melancholic ballad; genre alone does not distinguish between these contexts. This mismatch between label and function motivates the use of unsupervised learning to discover audio groupings that more closely reflect how people actually listen to music.

Rather than relying on predefined annotations, this study applies unsupervised machine learning to identify structure directly from the audio feature space. PCA reduces the dimensionality of the dataset while preserving the most significant variance, enabling visual inspection of relationships among songs. K-Means Clustering then partitions tracks into groups with similar acoustic profiles. The objective is to determine whether listening contexts—such as upbeat, melancholic, gym, or study atmospheres—emerge naturally from the data without any label supervision.

This approach is motivated by the methods underlying modern recommendation systems and contributes to the broader field of Music Information Retrieval by evaluating the degree to which quantitative audio features can substitute for subjective genre tags.

II. Review of Related Works

The application of data analytics and machine learning to music analysis—broadly termed Music Information Retrieval (MIR)—has grown substantially alongside the rise of digital streaming platforms. These platforms produce large volumes of audio metadata, enabling computational analysis of musical content at scale. While supervised learning has traditionally dominated genre classification and mood prediction tasks, unsupervised approaches for discovering latent musical structure remain an active area of research.

Early MIR research established that low-level audio features such as tempo, spectral characteristics, and timbre could achieve meaningful genre separation through statistical pattern recognition. Tzanetakis and Cook demonstrated that feature-based analysis of audio signals could produce robust genre classifiers, providing a foundational framework for the field [1]. Subsequent work expanded this approach to include higher-level perceptual attributes—rhythm, harmony, energy, and danceability—many of which are now standardized in the Spotify audio feature API.

More recent research has shifted toward unsupervised techniques to explore musical similarity without label supervision. Studies applying K-Means, DBSCAN, and hierarchical clustering to Spotify audio features consistently find that data-driven clusters do not align cleanly with traditional genre boundaries, suggesting that mood and listening context may be more salient organizing dimensions than genre alone [4]. These findings imply that genre labels impose categorical boundaries on what is actually a continuous acoustic space.

Dimensionality reduction plays an essential role in this research context. PCA has been widely applied to high-dimensional Spotify datasets to project songs into lower-dimensional spaces that preserve the most informative variance. Visualizations of these projections frequently reveal continuous gradients rather than discrete genre boundaries, reinforcing the view that musical similarity is better captured by proximity in feature space than by categorical labels [6].

In the context of recommendation systems, Spotify combines collaborative filtering with content-based methods that rely on audio feature similarity. Research into these systems suggests that unsupervised clustering of audio features can surface connections between songs not captured by user behavior alone, particularly for playlist generation targeting specific listening atmospheres such as focus, exercise, or relaxation [3].

The reviewed literature establishes the effectiveness of unsupervised learning and dimensionality reduction for large-scale music analysis. While prior work has confirmed the utility of Spotify audio features for classification and recommendation tasks, gaps remain in understanding how well genre and atmosphere labels reflect the true structure of the acoustic feature space. This study builds on established MIR methodology by applying PCA and K-Means to a 32,833-track Spotify dataset, with the explicit goal of evaluating whether listening atmosphere labels emerge naturally from audio-only clustering.

III. Methodology

This study followed a standard unsupervised machine learning pipeline, from data acquisition through preprocessing, dimensionality reduction, clustering, and anomaly detection.

A. Data Collection

The dataset was obtained from a publicly available Kaggle repository derived from the Spotify Web API. It contains 32,833 tracks, each described by 23 raw attributes including track metadata (name, artist, album, playlist genre) and 12 numerical audio features computed by Spotify's audio analysis engine. The features used in this study are listed in Table I.

TABLE I
Spotify Audio Features Used in Clustering

Feature	Description	Range
Danceability	Suitability for dancing	0.0–1.0
Energy	Perceptual intensity and activity	0.0–1.0
Valence	Musical positiveness	0.0–1.0
Acousticness	Confidence the track is acoustic	0.0–1.0
Instrumentalness	Predicts absence of vocals	0.0–1.0
Liveness	Presence of a live audience	0.0–1.0
Speechiness	Presence of spoken words	0.0–1.0
Tempo	Estimated beats per minute	˜50–220
Loudness	Overall loudness in decibels	−60–0 dB
Key	Estimated musical key (0=C, 11=B)	0–11
Mode	Modality (1=major, 0=minor)	0 or 1
Duration (ms)	Track length in milliseconds	Continuous

B. Data Preprocessing

All non-numerical metadata columns—including track name, artist, album, playlist name, genre, and subgenre labels—were removed prior to analysis. Retaining genre labels would introduce supervision into what is intended as a strictly unsupervised pipeline. The resulting cleaned dataset contained 32,833 rows and 13 numerical columns with no missing values.

1) Initial Sanity Check: An initial quality check confirmed that all 13 retained features were complete and fell within expected ranges. No imputation was required. The dataset shape after cleaning—$(32{,}833 \times 13)$—was verified programmatically before proceeding.

2) Feature Scaling: StandardScaler (Z-score normalization) was applied to all features, transforming each to zero mean and unit variance. Scaling was essential because the 12 features occupy vastly different numerical ranges: Tempo spans approximately 50–220 BPM, while Speechiness and Instrumentalness are bounded between 0 and 1. Without normalization, high-magnitude features would dominate Euclidean distance calculations in K-Means, producing clusters driven by scale rather than genuine acoustic similarity.

C. Dimensionality Reduction (PCA)

Principal Component Analysis was applied to reduce the 12-dimensional scaled feature space to 2 principal components (PC1 and PC2). This projection retained approximately 28.92% of total variance. While this compression sacrifices some information, it enables interpretable visual inspection of the full 32,833-track dataset in a single scatter plot and reduces the computational cost of subsequent clustering. The first two components capture the dominant axes of variation in the feature space—broadly corresponding to production intensity and acoustic complexity.

D. Clustering

1) K-Means: K-Means clustering was applied to the scaled 12-dimensional data (not the PCA-reduced version) to ensure clustering was informed by all features. The optimal cluster count was determined using the Elbow Method, which plots within-cluster sum of squares (WCSS) against k from 1 to 10. A distinct elbow at $k = 5$ indicated a meaningful reduction in cluster tightness without over-segmenting the data. A secondary analysis with $k = 8$ was performed to explore finer-grained groupings.

2) DBSCAN (Anomaly Detection): DBSCAN was applied to the 2D PCA projection to identify statistical outliers—songs that are acoustically unusual and do not belong to any dominant cluster. Parameters were set to $\varepsilon = 0.5$ and $min\_samples = 5$. Points labeled as noise (cluster ID $= -1$) were treated as anomalies. DBSCAN identified 38 anomalous tracks in the dataset.

E. Cluster Interpretation

Cluster centroids—the mean values of each audio feature per cluster—were computed to support interpretive labeling.

Each cluster was assigned a descriptive name based on its dominant feature profile. This labeling step is qualitative and serves to communicate findings rather than to introduce supervised annotations.

F. Prototype Application

A functional web application was built using Gradio to demonstrate the clustering system as a song recommendation engine. Given a track name, the application identifies its cluster and returns five songs from the same cluster as playlist recommendations.

IV. Algorithm

A. Principal Component Analysis (PCA)

PCA transforms a high-dimensional dataset into a lower-dimensional space by computing orthogonal linear combinations of the original features, ordered by the amount of variance they explain. The first principal component (PC1) captures the greatest variance in the data; each subsequent component captures the most remaining variance while remaining orthogonal to all prior components. Applying PCA before clustering reduces noise from low-variance features, mitigates the curse of dimensionality, and produces low-dimensional representations suitable for visualization.

B. K-Means Clustering

K-Means partitions n data points into k clusters by minimizing intra-cluster variance. The algorithm initializes k centroids and iterates between: (1) assigning each point to its nearest centroid using Euclidean distance, and (2) recomputing each centroid as the mean of its assigned points. Iteration continues until centroid positions converge. The result is a set of compact, approximately spherical clusters. The number of clusters k is determined empirically using the Elbow Method.

C. DBSCAN

DBSCAN identifies clusters as regions of high point density separated by low-density boundaries. Two parameters govern its behavior: ε defines the radius of a point's neighborhood, and *min_samples* defines the minimum number of neighbors required for a point to be classified as a core point. Points neither classified as core points nor reachable from any core point are labeled as noise. Unlike K-Means, DBSCAN does not require a predefined cluster count and handles outliers natively. In this study, DBSCAN is used exclusively for anomaly detection.

V. Results and Discussion

A. Exploratory Data Analysis

Exploratory analysis revealed clear structural patterns consistent with the application of unsupervised clustering. Feature distributions showed substantial variability across tracks, indicating the presence of acoustically distinct subpopulations rather than a homogeneous distribution.

Several features demonstrated strong inter-feature relationships. Energy and loudness were positively correlated (r = 0.68), consistent with the intuition that louder tracks tend to be perceived as more intense. Conversely, acousticness was negatively correlated with both energy (r = −0.54) and loudness (r = −0.36), reflecting a natural separation between acoustic and electronically produced content. Features such as danceability and energy were concentrated around moderate-to-high values, while instrumentalness and acousticness exhibited right-skewed distributions. These extremes represent niche styles likely to form distinct clusters.

The full correlation structure of the scaled feature space is shown in Fig. 1.

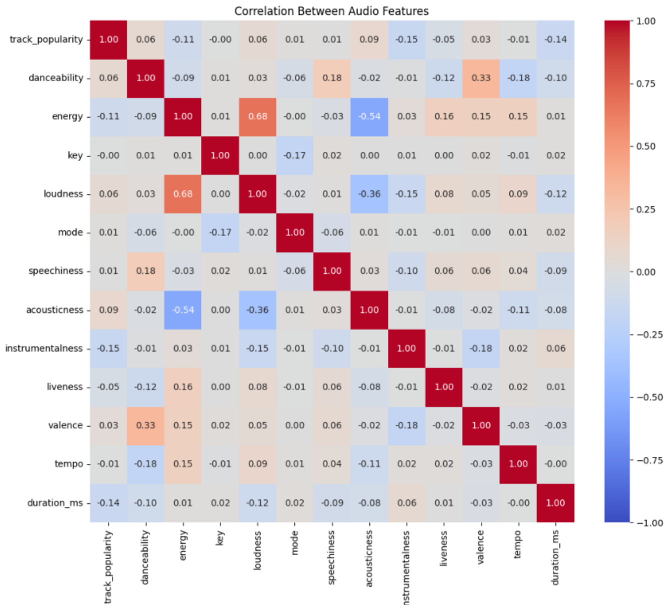


Fig. 1. Correlation heatmap of standardized audio features. Energy–loudness show the strongest positive correlation (r = 0.68); acousticness is strongly negatively correlated with energy (r = −0.54) and loudness (r = −0.36).

PCA results confirmed the presence of latent structure. The first two principal components captured 28.92% of total variance. The two-dimensional scatter plot of all 32,833 tracks (Fig. 2) reveals a dense core of acoustically similar tracks with a visible tail extending toward the lower-left, corresponding to acoustic and low-energy outlier tracks.

B. Elbow Method and Cluster Selection

The Elbow Method was applied to K-Means inertia values for $k = 1$ through $k = 10$. The resulting curve (Fig. 3) shows that inertia decreases steeply from $k = 1$ to $k = 5$, after which the rate of decrease slows substantially, indicating that five clusters provide the optimal trade-off between cohesion and separation. A secondary analysis with $k = 8$ was conducted to explore finer-grained segmentation.

C. K-Means Results ($k = 5$)

With $k = 5$, the model produced the cluster profiles summarized in Table II and visualized in Fig. 4.

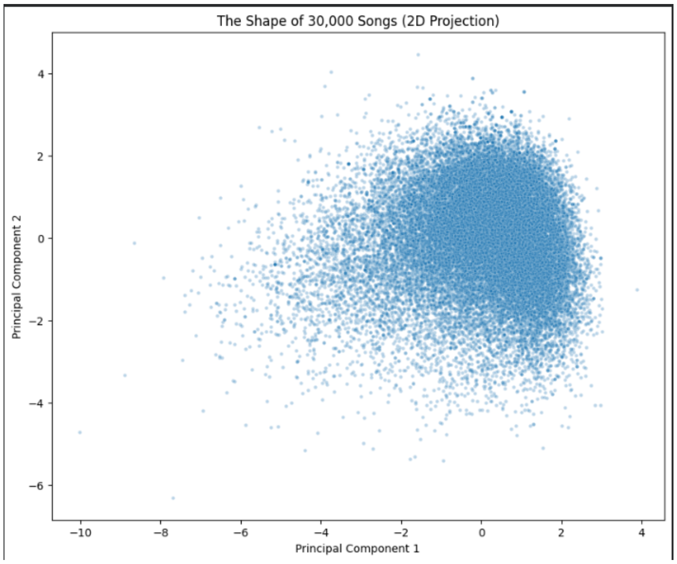


Fig. 2. Scatter plot of all 32,833 tracks projected onto PC1 and PC2 (28.92% variance retained). The distribution reveals a dominant dense region and a trailing low-energy population, confirming latent structure in the audio feature space.

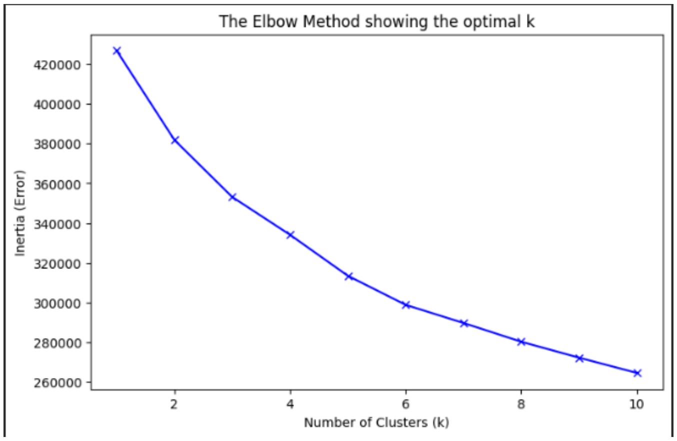


Fig. 3. Within-cluster sum of squares (inertia) plotted against k from 1 to 10. The curve flattens after $k = 5$, indicating the optimal number of clusters for this dataset.

TABLE II
K-MEANS CLUSTER PROFILES ($k = 5$) — MEAN FEATURE VALUES

Label	**Dance.**	**Energy**	**Acoust.**	**Valence**	n
Party / Dance	0.74	0.72	0.14	0.66	11,819
Gym / High-Intensity	0.54	0.80	0.06	0.40	8,923
Focus / Study	0.60	0.43	0.51	0.40	4,880
Chill / Roadtrip	0.72	0.67	0.18	0.55	4,606
Deep / Instrumental	0.66	0.79	0.07	0.39	2,605

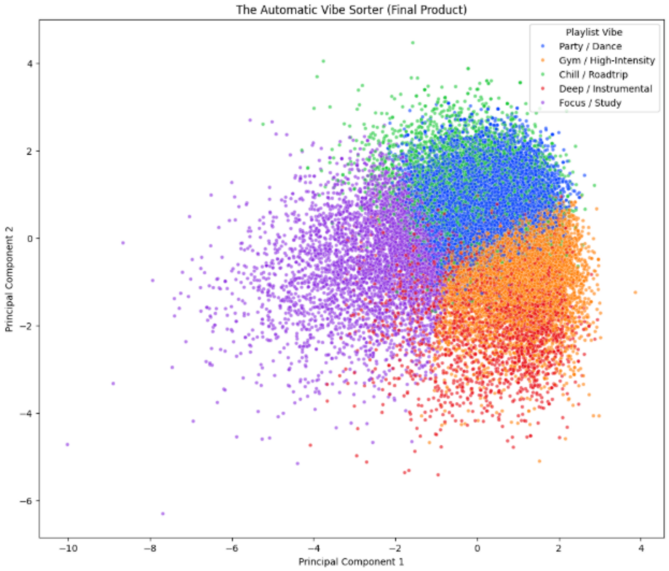


Fig. 4. PCA scatter plot with five K-Means clusters color-coded by listening context label. Clusters occupy broadly distinct regions, with the Focus/Study cluster (high acousticness, low energy) separated toward the lower-left and the Gym/High-Intensity cluster concentrated in the high-energy region.

D. Extended Analysis ($k = 8$)

Re-running K-Means with $k = 8$ produced finer-grained segmentation, subdividing dominant clusters into more specific listening contexts (Table III and Fig. 5). The eight-cluster model retained the primary acoustic distinctions identified at $k = 5$ while further differentiating sub-populations within the high-energy and dance-oriented segments.

TABLE III
K-MEANS CLUSTER PROFILES ($k = 8$) — MEAN FEATURE VALUES

Label	**Dance.**	**Energy**	**Acoust.**	**Instrum.**
Gym Phonk / Sigma Mode	0.55	0.79	0.08	0.02
Y2K Nostalgia (Happy)	0.71	0.62	0.14	0.03
Sad Girl / Depresso Espresso	0.59	0.40	0.61	0.12
Glitchcore / Gaming	0.66	0.79	0.07	0.76
Rave / Speed Garage	0.72	0.66	0.19	0.01
Villain Arc (Angst)	0.62	0.77	0.09	0.02
NPC / Background Vibes	0.61	0.78	0.12	0.06
Main Character Energy	0.76	0.73	0.17	0.01

E. Anomaly Detection

DBSCAN identified 38 tracks as statistical anomalies—songs that did not conform to any dominant density region in the PCA projection (Fig. 6). These outliers are concentrated in the sparse lower-left region of the feature space, corresponding to tracks with extreme acousticness, very low energy, or unusual tempo profiles. They likely represent experimental compositions, field recordings, or data artifacts in the source dataset.

F. Prototype Application

The Gradio-based recommendation application successfully demonstrated the practical utility of the clustering pipeline

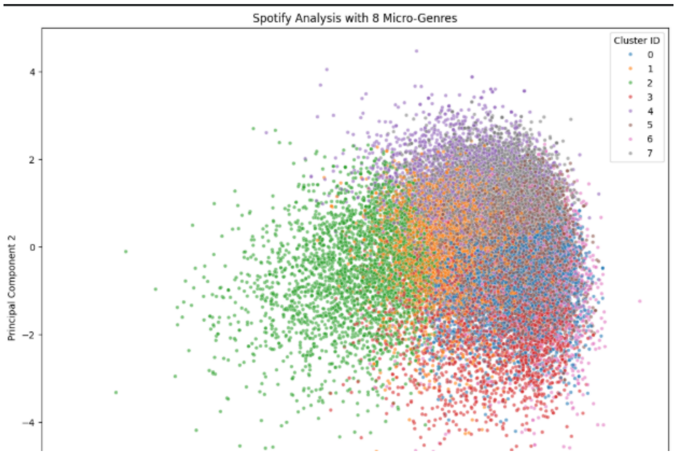


Fig. 5. Eight-cluster PCA visualization. The Sad Girl/Depresso Espresso cluster (high acousticness, low energy) occupies the lower-left region, while Glitchcore/Gaming (high instrumentalness) and Main Character Energy (high valence and danceability) form distinct sub-populations within the dense core.

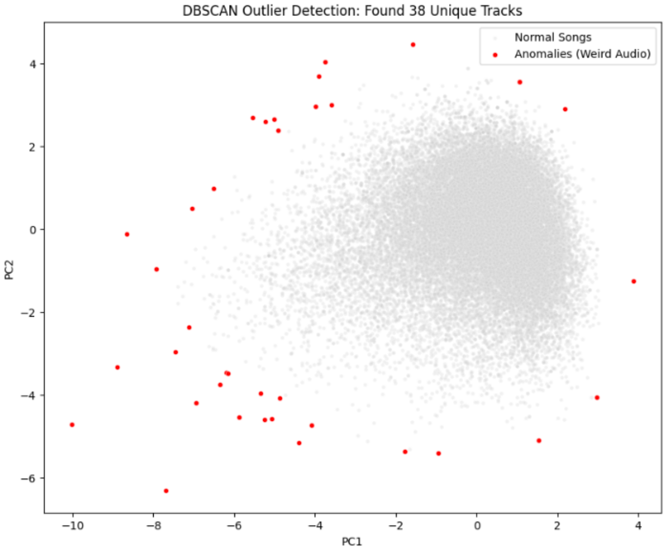
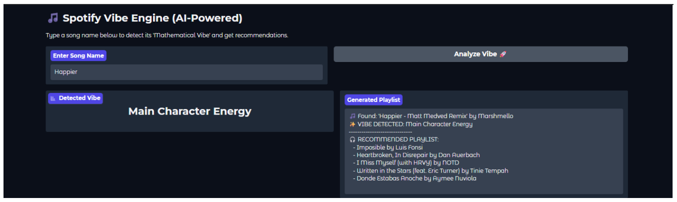


Fig. 6. DBSCAN anomaly detection on the 2D PCA projection. Red points ($n = 38$) mark acoustically unique tracks classified as noise. Their concentration in the sparse lower-left region indicates extreme deviations in acousticness or energy from the dominant track population.

(Fig. 7). When a user inputs a song name, the system locates the track in the dataset, retrieves its cluster assignment, and returns five songs from the same cluster as a generated playlist. Validation tests confirmed that inputting "Happier" returned a playlist labeled "Main Character Energy," and inputting "Talk Dirty" returned songs from the "Y2K Nostalgia" cluster—both consistent with reasonable listener expectations.

G. Discussion

The results confirm that audio features alone carry sufficient information to reconstruct human listening contexts without label supervision. Three clusters are particularly notable as evidence of this finding:



Spotify Vibe Engine (AI-Powered)

Type a song name below to detect its 'Mathematical Vibe' and get recommendations.

Happier

Main Character Energy

Found: 'Happier - Matt Medved Remix' by Marshmello
VIBE DETECTED: Main Character Energy

RECOMMENDED PLAYLIST:
- Imposible by Luis Fonsi
- Heartbroken, In Disrepair by Dan Auerbach
- I Miss Myself (with HRVY) by NOTD
- Written in the Stars (feat. Eric Turner) by Tinie Tempah
- Donde Estabas Anoche by Aymee Nuviola

Fig. 7. Gradio prototype application interface. The user inputs a song name; the system detects its vibe cluster and generates a five-song playlist recommendation from the same cluster.

- **Sad Girl / Depresso Espresso** (Cluster 2) recorded the highest acousticness (0.61) and lowest energy (0.40) of any cluster. This combination—acoustic instrumentation and low intensity—isolates melancholic tracks purely from numerical data, with no access to mood tags, lyrics, or listener behavior.
- **Glitchcore / Gaming** (Cluster 3) showed a markedly elevated instrumentalness score (0.76), identifying a subset of predominantly vocal-free tracks corresponding to instrumental or electronic music used for focused activities.
- **Main Character Energy** (Cluster 7) recorded the highest valence (0.69) and danceability (0.76), capturing high-positivity tracks associated with confident, upbeat listening contexts.

These findings are consistent with the broader MIR literature: data-driven clusters align more closely with mood and listening atmosphere than with traditional genre categories. The overlap between genre and cluster boundaries—for example, Pop tracks appearing in both the "Y2K Nostalgia" and "Main Character Energy" clusters—supports the conclusion that genre labels are an imprecise proxy for functional listening context.

VI. Conclusion

This study demonstrated that unsupervised machine learning techniques can effectively uncover meaningful musical structure from quantitative audio features alone. By applying PCA for dimensionality reduction and K-Means clustering for segmentation to a 32,833-track Spotify dataset, the analysis produced interpretable listening context groupings—including gym, study, melancholic, and upbeat atmospheres—without relying on any predefined genre labels or mood annotations. DBSCAN further identified 38 statistically anomalous tracks that fall outside dominant acoustic patterns. A functional Gradio prototype validated the pipeline as a practical song recommendation engine.

Several limitations warrant acknowledgment. First, the study relied on a single dataset that may not fully represent musical diversity across cultures, eras, or niche genres. Second, clustering outcomes were sensitive to hyperparameter selection—particularly the choice of k in K-Means and the ε and *min_samples* parameters in DBSCAN. Third, audio features alone cannot capture lyrical content, cultural associations, or listener demographics, constraining the interpretability of the discovered clusters.

Future work should expand the dataset to include wider musical traditions and time periods. Integrating lyrical features through natural language processing would provide richer, multimodal representations. Exploring hierarchical clustering, Gaussian mixture models, and deep learning-based representations may reveal more complex structural patterns. Validating cluster quality against user behavior and recommendation performance would further establish practical relevance.

Overall, this study establishes that quantitative audio features encode sufficient information to recover human listening contexts through unsupervised learning, and demonstrates a viable pipeline for context-aware music recommendation.

References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.

[2] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. ISMIR*, 2000.

[3] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. ISMIR*, 2011.

[4] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Found. Trends Inf. Retr.*, vol. 8, no. 2–3, pp. 127–261, 2014.

[5] R. van den Berg and M. Holzapfel, "Unsupervised clustering of Spotify audio features for playlist generation," in *Proc. ISMIR*, 2020.

[6] P. Knees and M. Schedl, *Music Similarity and Retrieval: An Introduction to Audio- and Web-Based Strategies*. Springer, 2016.

[7] Spotify, "Audio features object," *Spotify for Developers Web API Reference*, 2023. [Online]. Available: https://developer.spotify.com/documentation/web-api