Automated Music Playlist Generator

Aahan Piplani
IIITD
aahan22001@iiitd.ac.in

Abhishek Bansal
IIITD
abhishek22021@iiitd.ac.in

Akshat Kothari
IIITD
akshat22053@iiitd.ac.in

Kartikeya
IIITD
kartikeya22241@iiitd.ac.in

Vinayak Agrawal
IIITD
vinayak22574@iiitd.ac.in

Abstract

With the explosive growth of music streaming platforms such as Spotify, Apple Music, and Wynk, the global number of songs has surged exponentially. Curating personalized playlists for users has become a daunting task, requiring extensive manual effort to listen to and classify songs based on their unique audio characteristics. This project aims to automate the categorization of songs with similar musical features into playlists, leveraging modern machine learning techniques and visualization tools.

Our approach seeks to enhance playlist generation accuracy by incorporating lyrical analysis, an aspect often overlooked in related studies. By applying topic modeling techniques on song lyrics, we aim to extract thematic features, which will be integrated alongside audio-based features to create robust models for playlist generation. This innovative combination of lyrical and audio analysis promises to refine the organization of millions of songs into tailored playlists that resonate with user preferences.

Source code: GitHub

1 Introduction

The rapid expansion of music streaming platforms like Spotify, Apple Music, and Wynk has revolutionized how users access and enjoy music. With millions of songs now available at their fingertips, these platforms face the complex challenge of curating personalized playlists that align with individual user preferences. Traditional approaches to playlist generation often rely on manual classification of songs based on audio characteristics, which is both time-consuming and limited in scalability.

To address this challenge, our project explores the automation of song categorization by leveraging advanced machine learning techniques and visualization tools. Unlike many existing studies that focus solely on audio-based features, our approach introduces an innovative aspect—lyrical analysis. By applying topic modeling techniques to song lyrics, we aim to extract thematic features that capture the underlying essence of a song. These lyrical features will be integrated with audio-based attributes, such as MFCC, Chroma, and Spectral Contrast, to develop robust models for playlist generation.

By combining lyrical and audio analysis, this approach not only enhances the accuracy of playlist recommendations but also offers a more holistic understanding of a song's characteristics. This comprehensive strategy has the potential to transform playlist curation, enabling

the organization of vast song libraries into meaningful and personalized listening experiences for users.

2 Literature Review

Music genre classification has always been a primary focus in the nexus of machine learning and music. This literature review brings together important contributions and methodologies as collated in the aforementioned studies.

2.1 Aggregate Features and ADABOOST for Music Classification

A new approach was defined by Bergstra et al. (2006) based on aggregate features and ADABOOST. Relevant features that capture spectral, temporal, and timbral properties of an audio signal were emphasized as critical in the process of classification. The ability to combine weak classifiers into strong ensemble ADABOOST thus proved effective in coping with the high variability of music data. Their results exposed the strong potentiality of ensemble-based approaches toward increasing classification accuracy without exhausting computational resources.

2.2 Song-Level Features and Support Vector Machines for Music Classification

Mandel and Ellis (2005) focused on rhythmic patterns, harmonic contributions, and timbral textures at the level of song for the efforts made toward classification. The method used for that purpose was Support Vector Machines (SVM), which is one of the most used traditional machine learning methods. The authors have shown that engineering of features has given competitive accuracy with an SVM in technical configuration for its robust classification ability. The study revealed the need for selecting discriminative features, outlining SVMs as a strong baseline in the genre classification task.

2.3 Music Genre Classification with the Million Song Dataset

Liang and his colleagues (2011) utilized the Million Song Dataset as an evaluation metric to consider a broad range of machine learning algorithms including the scalability of their models. They studied traditional methods such as k-Nearest Neighbor (k-NN) and much more advanced ones like matrix factorization. It was well found that traditional models do their jobs well when proper feature selection and dimensionality reduction have been applied. However, as the dataset grew, there were concerns with scalability and performance problems, making the trade-off towards slightly less computationally intensive techniques.

2.4 Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches

Ndou et al. (2021) comprehensively reviewed deep learning and traditional machine learning techniques in an effort to look at the similarity and differences between the two. It dived into traditional machine learning models such as Decision Trees, Random Forests, Naïve Bayes, and k-NN while using less data and reducing demands on computation. For instance, Decision Trees and Random Forests were found to be interpretable models that possessed the ability to handle non-linear relationships efficiently. The discussion also covered the idea that Naïve Bayes, while simple, was effective with feature independence assumptions. Traditional models, such as k-NN, take a step further and praise their straightforward implementation but have their performance heavily dependent on correct feature normalization and distance metrics. Although deep learning models showed much better performance on the accuracy scale when

compared to traditional models, the review stressed that traditional models are still relevant in scenarios with limited computation or data availability.

2.5 A Machine Learning Approach to Automatic Music Genre Classification

According to Silla Jr. et al. (2008), the ability and preprocessing of features were indicated important concerning conducting a proper analysis of machine-controlled genre musical classification. It had demonstrated the traditional power of classifiers like Random Forests and ensemble methods integrating low-level descriptors like MFCCs, with musicological features. The construction was notably proficient, especially for overfitting and noisy data, with excellent performance in conjunction with PCA techniques for dimension reductions.

2.6 Comparative Analysis and Key Insight

In almost all these studies, traditional machine learning approaches are emphasized due to interpretability, computational efficiency, or adaptability to small dataset scenarios. Important insights include:

Feature Dependence: Traditional methods rest heavily upon feature engineering; well-designed features—in this case, MFCCs; rhythmic patterns; and timbral textures—are crucial to models such as SVMs, Random Forests, and k-NN.

Model Performance: Ensemble approaches such as ADABOOST and Random Forests were found especially useful in enhancing classification accuracy through reducing overfitting problems. Perhaps most interestingly, high-dimensional data generally offered a strong baseline for such support vector machines.

Limitations: Most of the time, traditional techniques also face scalability issues, and as a result suffer from poorer quality as the complexity of the datasets increases or more levels of feature hierarchy have to be deepened.

3 Dataset Features

We are utilizing the GTZAN Dataset for Music Genre Classification, which contains 10 distinct genres, each represented by 100 audio samples. These samples are provided in a 30-second waveform format. All audio samples in the dataset have been preprocessed to ensure they are clean and noise-free.

Genre	Number of samples
blues	100
classical	100
country	100
disco	100
hiphop	100
jazz	100
metal	100
pop	100
reggae	100
rock	100

Table 1: Table 1.Genres of music

The class distribution of playlists is shown in Figure 1:

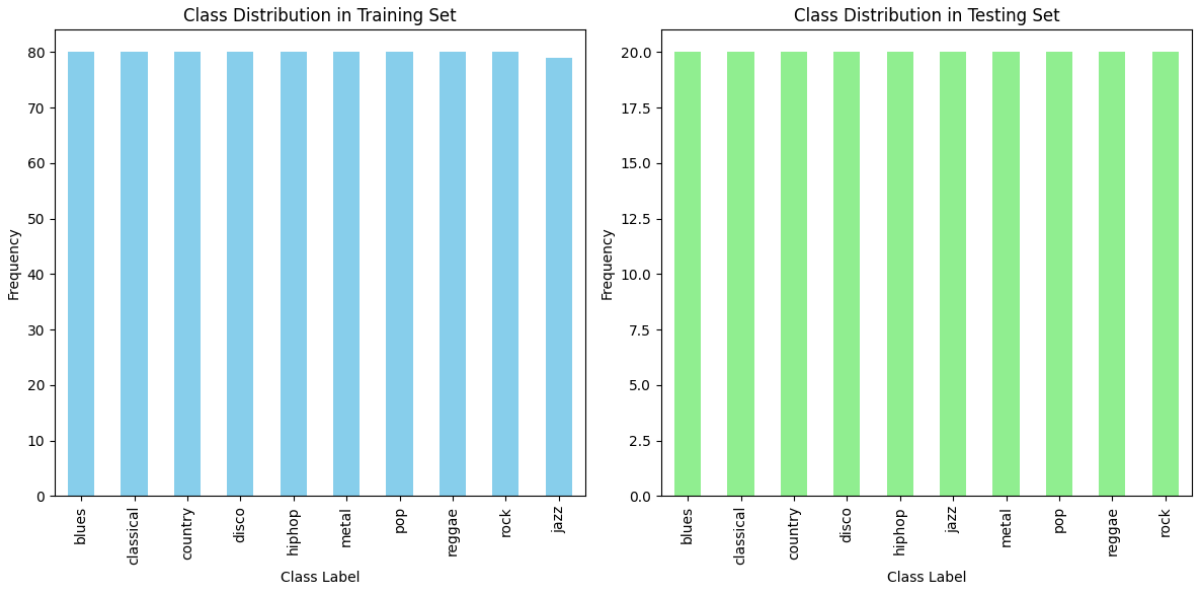


Figure 1: Class Distribution of Playlists.

3.1 Preprocessing, Feature Extraction

The GTZAN Music Genre Classification dataset contains 30-second compressed waveform audio samples, with each of the ten genres represented by 100 audio files in waveform format. During feature extraction, we ensured that only audio files were processed to maintain consistency and avoid discrepancies in the results. Preprocessing involved standardizing audio sample rates, trimming or padding files to uniform lengths, and applying noise reduction to improve the quality of extracted features.

3.1.1 Extracted Features

We preprocessed the data and extracted key audio features from waveform-format samples. Based on an analysis of various research papers, we identified important features for classifying audio samples. The extracted features are as follows:

1. Chroma features: Chroma features represent the intensity of the 12 different pitch classes (e.g., C, C, D, etc.) in an audio signal, capturing harmonic and tonal content.
2. Root Mean square: The RMS (Root Mean Square) value represents the average power or energy of an audio signal, providing a measure of its loudness.
3. Spectral centroid: The spectral centroid indicates the "center of mass" of a sound's spectrum, often perceived as the brightness or sharpness of the audio.
4. Spectral bandwidth: Spectral bandwidth measures the width of the range of frequencies present in a sound, indicating how spread out the frequencies are around the spectral centroid.
5. Spectral rolloff: Spectral rolloff is the frequency below which a specified percentage (typically 85-90%) of the total spectral energy is concentrated, indicating the point where high frequencies taper off.
6. Harmony: Harmony refers to the degree of harmonic content in an audio signal, indicating the presence of tonal, musical components.

7. Percussive: Percussive refers to the transient or sharp, rhythmic elements in an audio signal, often used to identify drum-like or percussive sounds.

8. Tempo: Tempo refers to the speed or pace of a musical piece, typically measured in beats per minute (BPM), indicating how fast or slow the rhythm is.

3.1.2 Dimensionality Reduction using Principal Axis Component (PCA)

Principal Component Analysis is a method of reducing dimensions of a dataset by transforming a large set of variables into a smaller set of variables that still contains most of the information in the larger data set. The various steps in PCA include standardization, computation of covariance matrix and eigenvectors to identify principle components. PCA can be thought of as fitting a "p-dimensional ellipsoid" to the data; every axis of the ellipsoid means a single principal component. For our project, we are working with 57 principle components.The PCA plot of the audio features indicates significant overlap between the data points when reduced to two dimensions. This suggests that reducing the feature set to only two components does not sufficiently capture the essential characteristics of the data, thereby emphasizing the importance of utilizing all 57 features for effective classification. Each feature likely contributes critical information that aids in differentiating between the various categories, making them indispensable for accurate analysis. Figure 2 shows the plot of the PCA.

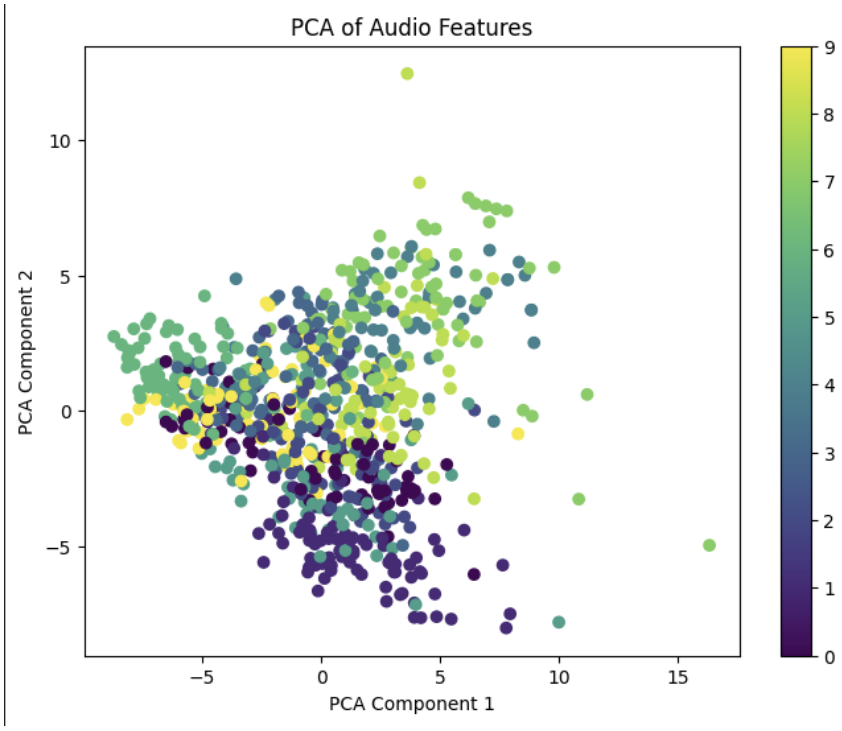


Figure 2: PCA plot

3.1.3 Visualizing High dimensional data using t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for exploring and visualizing high-dimensional data. Unlike PCA, t-SNE is not a linear projection; it captures non-linear structures by preserving local relationships between data points. t-SNE creates a probability distribution using a Gaussian function to define the relationships between points in the high-dimensional space, allowing for an effective low-dimensional mapping. The t-SNE plot of audio features shows a clearer separation of data points compared

to the PCA plot, suggesting that t-SNE was effective in capturing non-linear relationships between the audio features. There appear to be clusters forming,with some overlap, which implies that while t-SNE captures some underlying structure in the data. The visualization highlights potential groupings that may correspond to different classes or genres, indicating that the audio features contain useful information.

Figure 3 shows the plot of the t-SNE.

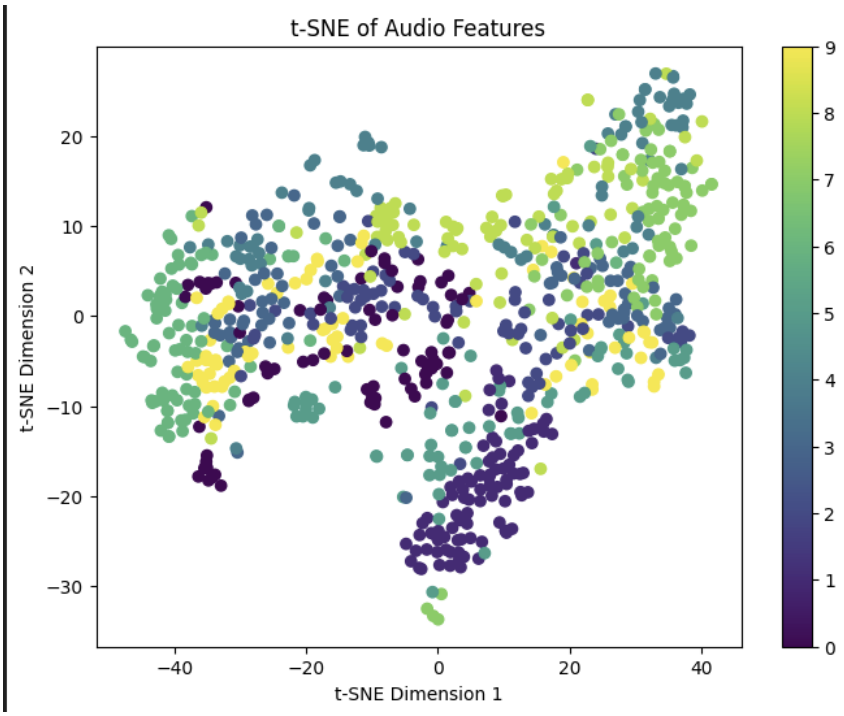


Figure 3: t-SNE plot

3.1.4 Data Standardization

Standardization is a scaling technique that centers values around the mean and scales them to have a unit standard deviation. This ensures that the mean of the attribute is zero, and the distribution has a standard deviation of one. The formula for standardization is as follows:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{1}$$

where $\bar{x}$ is the mean of the feature values, s is the standard deviation of the feature values.

4 Methodologies

Our objective was to classify audio samples into one of the ten predefined classes. For this classification task, we employed a multi-class prediction approach to distinguish between the ten classes. Various machine learning models were evaluated to determine which performed best in accurately classifying the audio samples.

4.1 Classification

We conducted a multi-class classification on the dataset, using an 80:20 train-test split. The training set consisted of 80 songs per class, while the test set included 20 songs per class.

We employed several classification models, including Decision Trees, ADA Boost with Decision Trees, LightGBM (LGBM) Classifier, XGBoost, RBF SVC, MLP Classifier, Random Forest and K-Means clustering to observe the clusters. To optimize the model parameters, we utilized GridSearchCV along with cross-validation techniques.

4.1.1 Clustering

We did K-Means clustering to determine the optimal number of clusters using both the Elbow Method and Silhouette Scores. The "elbow" point, indicating the optimal k, was identified using the KneeLocator. We then applied K-Means with the chosen k, calculated its silhouette score, and used PCA to reduce the dataset to two dimensions.

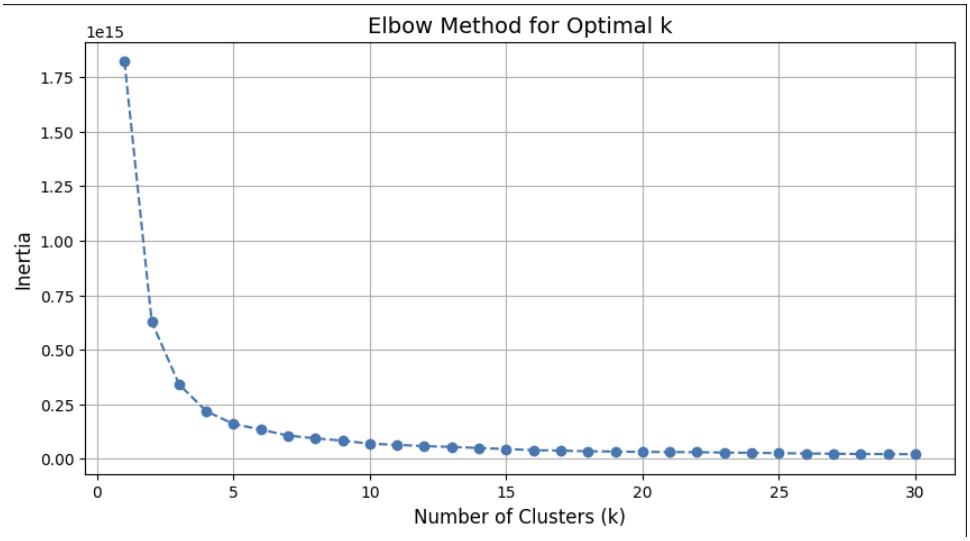


Figure 4: Elbow method for optimal k

The graph above depicts the Elbow Method for determining the optimal number of clusters (k) in a dataset. It plots the inertia (within-cluster sum of squares) against the number of clusters. The "elbow point," where the curve starts to flatten, indicates the ideal number of clusters.

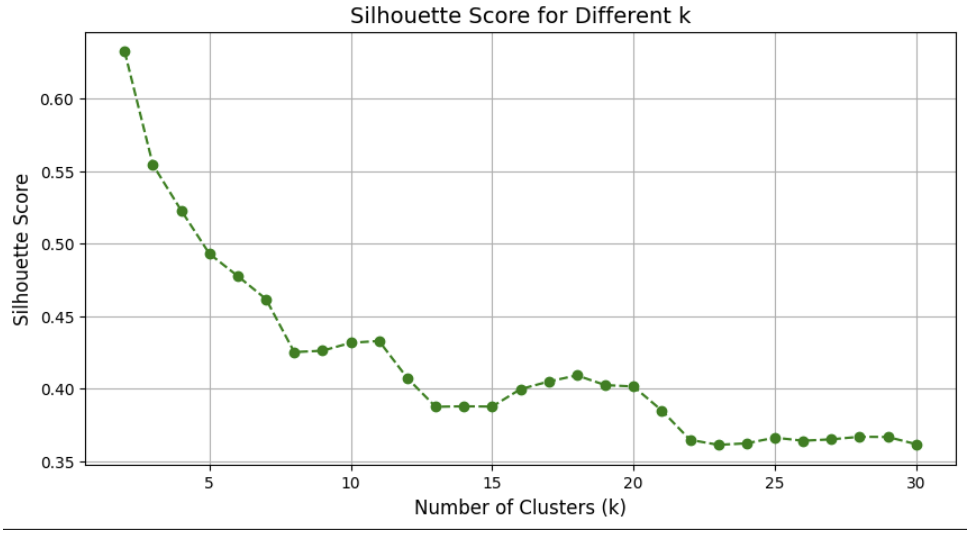


Figure 5: Silhouette score vs number of clusters

The graph illustrates the Silhouette Score for different numbers of clusters (k). The Silhouette Score measures the quality of clustering, with higher values indicating better-defined clusters.

The peak Silhouette Score occurs around $k = 2$, suggesting that two clusters provide the best separation and cohesion for this dataset. Beyond this point, the score decreases, indicating diminishing cluster quality as k increases.

Optimal number of clusters (k) determined by Elbow Method: 4
Silhouette Score for k=4: 0.52

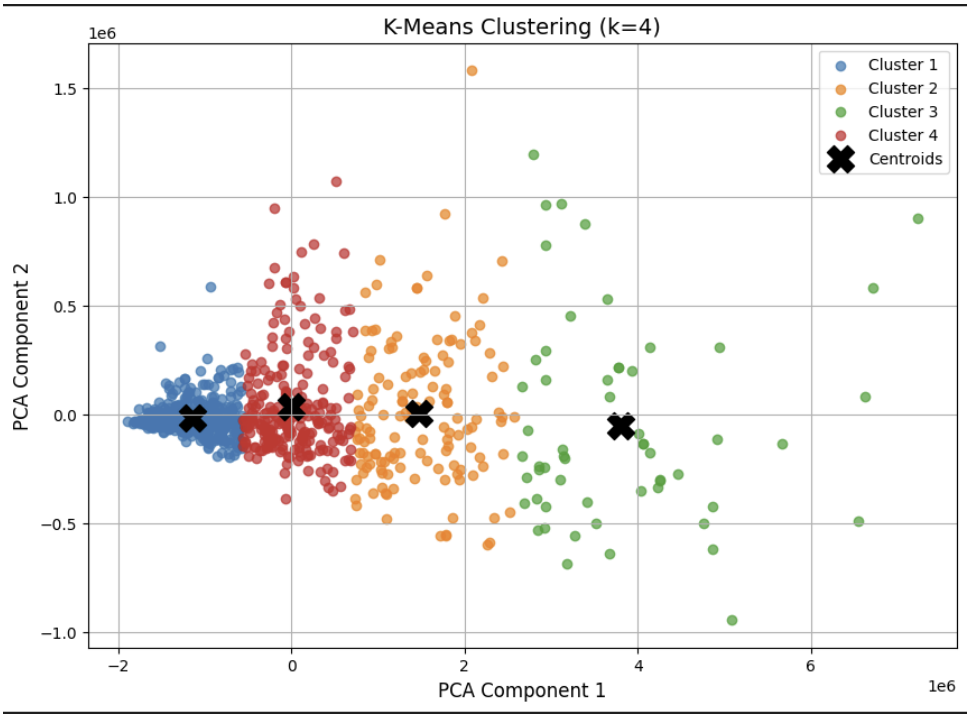
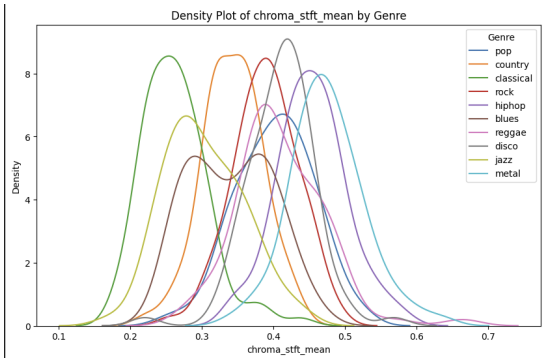


Figure 6: K-means clustering graph for optimal k

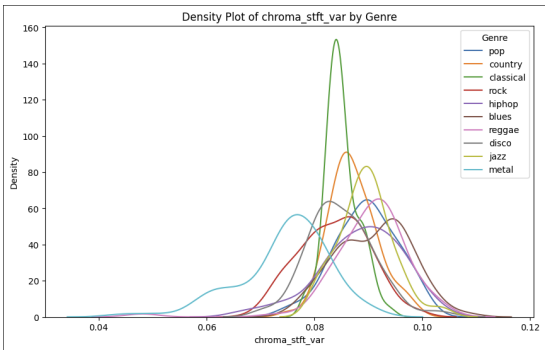
5 Results and Analyses

5.1 Features analysis

The distribution of various classes across different features is shown below, highlighting the characteristics of each class in relation to different audio features.



(a) Mean of chroma features

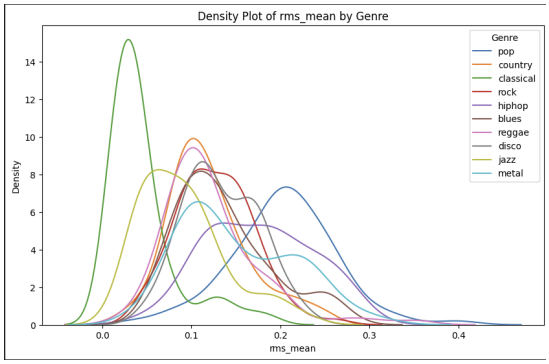


(b) Varience of chroma features

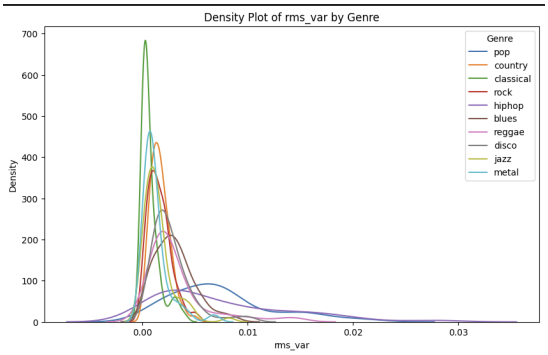
Figure 7: Chroma stft features

The two density plots illustrate the distributions of the mean and variance of chroma features across different music genres. The plots indicate that genres exhibit distinct tonal characteris-

tics, as reflected in both the average chroma values and their variability. Classical music shows a high consistency in both mean and variance, suggesting a defined harmonic structure, while genres like jazz and hip-hop display broader distributions, indicating greater diversity in tonal content. These differences highlight how chroma features can effectively capture and distinguish the tonal profiles of various music genres.



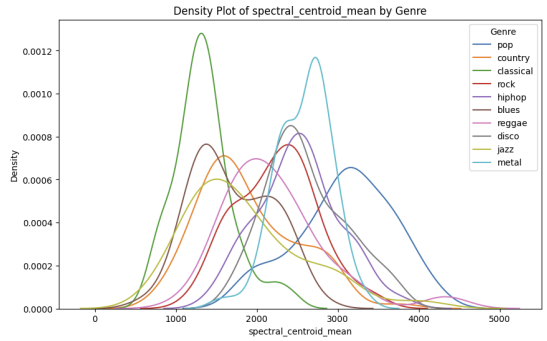
(a) Mean of root mean square features



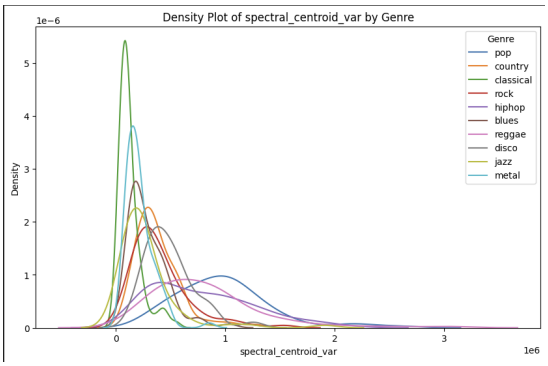
(b) Varience of root mean square features

Figure 8: Root mean square features

The density plots for rms mean and rms varience highlight the differences in loudness and dynamic range across music genres. Classical music shows lower RMS mean and variance, indicating it is generally quieter and more consistent in loudness. In contrast, genres like rock, pop, and hip-hop have higher RMS means and greater variance, suggesting these genres are typically louder and have more dynamic variation. These distinctions illustrate how RMS features can effectively differentiate genres based on their energy and intensity profiles.



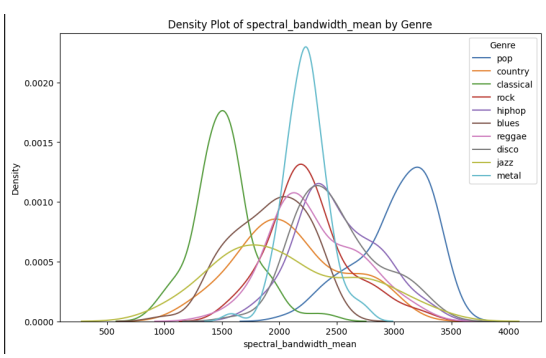
(a) Mean of spectral centroid features



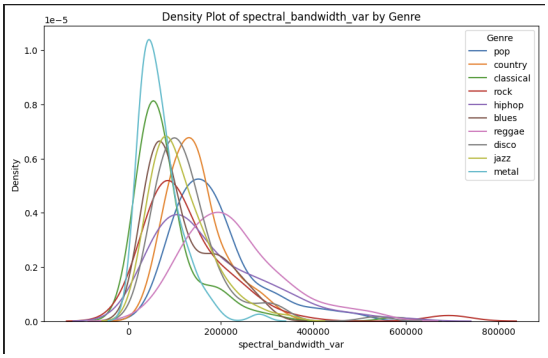
(b) Varience of spectral centroid features

Figure 9: Spectral centroid features

The density plots for spectral centroid mean and spectral centroid var highlight the differences in brightness and tonal dynamics across music genres. Genres like metal and pop have higher spectral centroid means, indicating a brighter, sharper sound, while classical music has a lower spectral centroid mean, reflecting a warmer tone. The variance plot shows that classical music maintains consistent brightness, whereas genres like rock and jazz exhibit greater variability, indicating dynamic tonal shifts. These features are effective for distinguishing between genres based on their brightness and dynamic tonal properties.



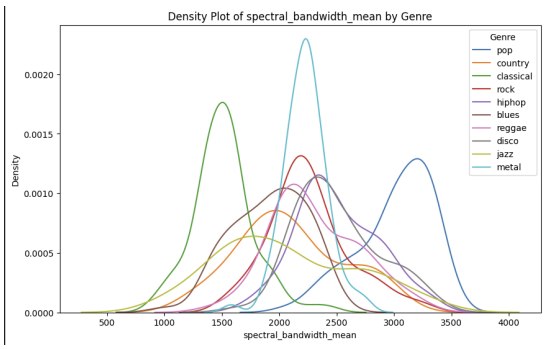
(a) Mean of spectral bandwidth features



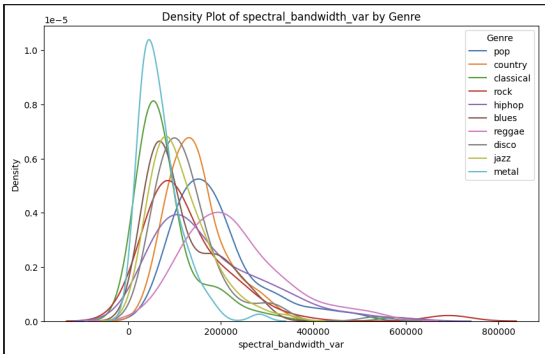
(b) Varience of spectral bandwidth features

Figure 10: Spectral bandwidth features

The density plots for spectral bandwidth mean and spectral bandwidth var reveal differences in the frequency range and its variability across music genres. Genres like metal and pop show higher spectral bandwidth means, indicating a broader frequency range, whereas classical music has a lower mean, reflecting a narrower, focused sound. The variance plot shows that pop and metal have more variability in their frequency range, suggesting dynamic shifts, while classical and country maintain a consistent frequency profile. These features help in distinguishing genres based on the complexity and consistency of their sound spectrum.



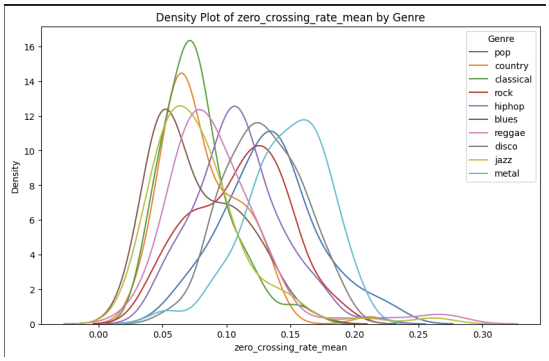
(a) Mean of rollof features



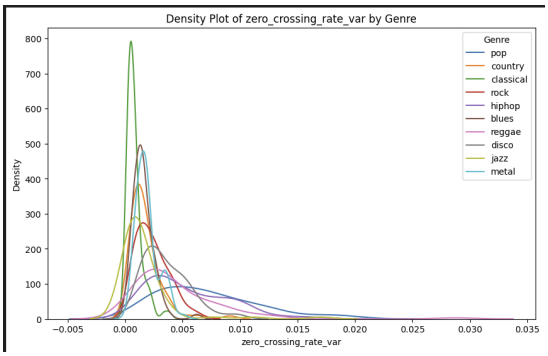
(b) Varience of rollof features

Figure 11: Rollof features

The density plots for rolloff mean and rolloff var illustrate the differences in high-frequency content across music genres. Genres like metal and pop have higher spectral rolloff means, indicating more high-frequency energy and a brighter sound, while classical music has lower rolloff values, reflecting a smoother, mellower tone. The variance in spectral rolloff shows that classical and country music maintain consistent high-frequency content, while pop and rock exhibit greater variability, suggesting more dynamic shifts. These features help differentiate genres based on their brightness and dynamic changes in high-frequency content.



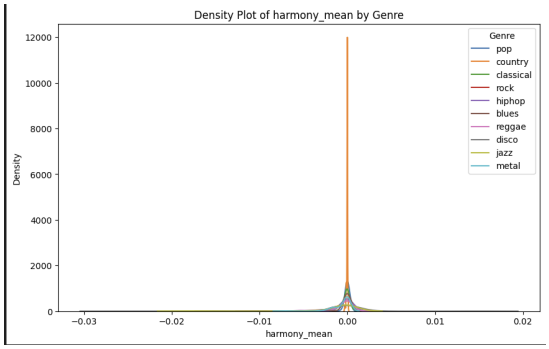
(a) Mean of zero crossing rate features



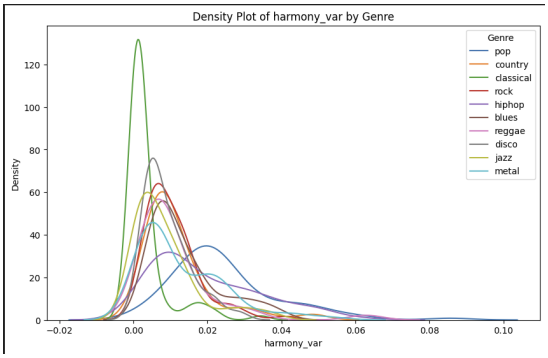
(b) Varience of zero crossing rate features

Figure 12: Zero crossing rate features

The density plots for zero crossing rate mean and zero crossing rate var illustrate differences in the percussive nature and variability across music genres. Genres like metal, pop, and disco have higher ZCR means, indicating more frequent transitions in their waveforms, which results in a noisier and more percussive sound, while classical music has a lower ZCR, reflecting its smooth and melodic nature. The variance plot shows that classical and country music maintain consistent waveform characteristics, whereas rock and metal exhibit more dynamic changes. These features help in differentiating genres based on their percussiveness and the variability of rhythmic content.



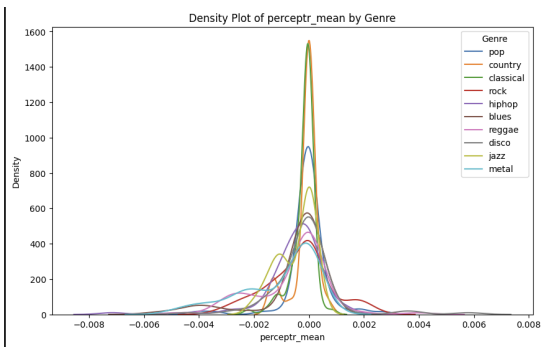
(a) Mean of harmony features



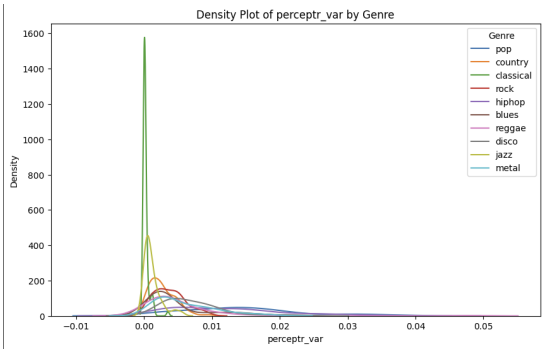
(b) Varience of harmony features

Figure 13: Harmony features

The density plots for harmony mean and harmony var illustrate the differences in harmonic content and its variability across music genres. The harmony mean plot shows that most genres cluster closely around zero, with country music showing a particularly sharp peak, indicating high harmonic stability. The harmony var plot reveals that classical and country music maintain consistent harmonic structures with low variance, while genres like pop, rock, and metal exhibit greater variability, suggesting more dynamic harmonic changes. These features highlight the differences in harmonic stability and fluctuation among various genres.



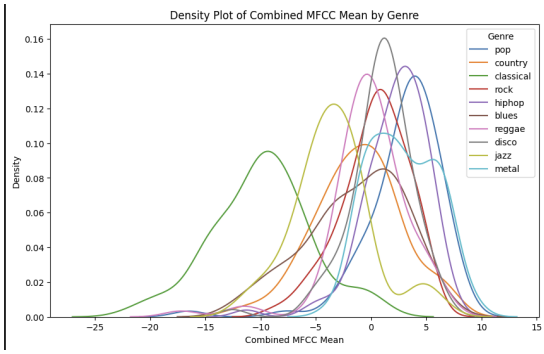
(a) Mean of Percussive features



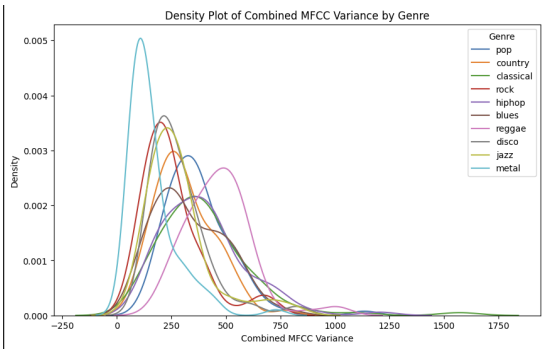
(b) Varience of Percussive features

Figure 14: Percussive features

The density plots for perceptr mean and perceptr var reveal differences in percussive content and its variability across music genres. Both plots show that country and classical music tend to have minimal and highly consistent percussive elements, reflected by sharp peaks near zero. In contrast, genres like pop, rock, and metal show broader distributions in both mean and variance, indicating greater percussive content and dynamic changes. These features help differentiate genres based on the level and variability of percussive elements.



(a) Mean of mfcc features



(b) Varience of mfcc features

Figure 15: Mfcc features

The density plots for Combined MFCC Mean and Combined MFCC Variance highlight differences in the timbral characteristics across music genres. Classical music has a more negative MFCC mean, indicating a smoother and lower-energy timbre, while pop and metal are characterized by a brighter, energetic sound. The variance plot shows that genres like pop and metal have greater timbral variability, reflecting more dynamic changes, whereas classical and country music maintain a consistent timbral texture. These features help differentiate genres based on their overall tone and dynamic timbral shifts.

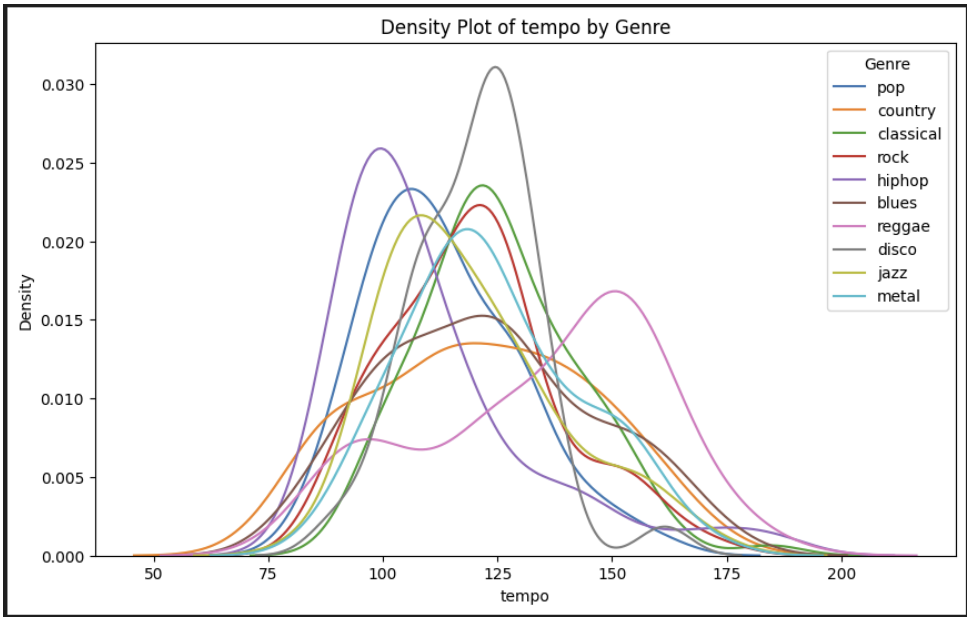


Figure 16: Tempo features

The density plot for the tempo feature across genres shows distinct tempo ranges that align with each genre's rhythmic characteristics. Blues tends to have a slower tempo, while disco and hip-hop are characterized by higher and more consistent tempos, making them suitable for dance. Genres like metal and jazz exhibit broad distributions, indicating a wide range of rhythmic possibilities. Overall, tempo serves as an important distinguishing feature, reflecting each genre's typical energy and rhythmic style.

5.2 Classification

For our multi-class classification task with 10 music genres, we utilized multiple models to predict song classifications. To optimize performance, we conducted a GridSearchCV to identify the best parameters for each model and applied cross-validation to ensure precise performance estimation.

5.2.1 Model training without fine tuning hyperparameters and cross validation

Model	Precision	Recall	F1	Acc
SVC	0.74	0.72	0.73	0.725
MLP	0.79	0.79	0.79	0.785
DT	0.49	0.49	0.49	0.495
RF	0.79	0.79	0.78	0.785
XGB	0.75	0.76	0.75	0.755
Lgbm	0.77	0.77	0.76	0.765
Adaboost on DT	0.54	0.47	0.48	0.475

Table 2: Table 3. Multi Classification Scores

From the above table, Random Forest and MLP achieved the highest performance, both with an accuracy of 78.5

5.2.2 Model training with refined parameters and using cross validation for the accuracy

Model	Precision	Recall	F1	Acc
SVC	0.77	0.77	0.76	0.765
MLP	0.76	0.75	0.75	0.75
DT	0.48	0.49	0.48	0.495
RF	0.77	0.77	0.77	0.77
XGB	0.76	0.76	0.75	0.755
Lgbm	0.77	0.77	0.77	0.77
Adaboost on DT	0.49	0.48	0.46	0.48

Table 3: Table 3. Multi Classification Scores

From the above table, after tuning hyperparameters and performing cross-validation for accuracy, the best-performing models are Random Forest and LGBM, both achieving an accuracy of 77