A combined Approach to Music Genre Classification

Arav Patel

Abstract

This project explores the problem of music genre classification using machine learning techniques using the GTZAN dataset. The process involves preprocessing the dataset, extracting meaningful features through advanced audio signal processing methods, and evaluating multiple classification models. Out of the chosen models, we find that Support Vector Machines report the best classification. A more careful feature selection contributed significantly to the success of many algorithms. Our models outperformed similar implementations of similar algorithms. The SVM algorithm outperformed several Deep learning models with a fraction of the computational resources and features.

Introduction

Music is one of the most widely consumed forms of Media. The rise of streaming platforms like Spotify and Apple Music has dramatically improved access and affordability of music, providing unprecedented access to over 100 million songs. These platforms rely heavily on genre classification to organize music libraries, generate playlists, and recommend similar songs to users. Song Recommendations have now become one of the most popular ways for people to discover new music.However, music genres are inherently subjective, and often labeled manually by humans, which can be inefficient and inconsistent. Automating this process using machine learning techniques can significantly enhance the efficiency and accuracy of genre classification.

Genre Classification is a crucial sub-problem in the overall goal of Music recommendation. By categorizing tracks based on shared characteristics, it can create a personalized recommendation for each user. Despite its potential, genre classification presents unique challenges due to the overlap definition of a genre and the complexity of audio signals. Automatic classification of audio has a long history, originating from speech recognition systems. One of the most widely used features is Mel-Frequency Cepstral Coefficients (MFCCs) because they are meant to mimic human audio perception. Other features include Spectral features such as roll off, Bandwidth and the Chroma Features.

Objective:

This primary objective of the project is to classify music into different genres based on its unique sound characteristics. It aims to find an effective approach to music classification by comparing different feature extraction methods and algorithms.

The dataset:

The GTZAN dataset, developed by Tzanetakis and Cook is one of the most widely used and historically significant datasets for music genre classification, making it an ideal choice for this study. It contains 10 distinct genres: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock. Each genre includes 100 audio tracks included as .wav files, with each track of 30 seconds each, sampled at 22,050 Hz, and recorded as 16-bit mono audio.The Gtzan dataset was intentionally used due to its popularity as well as historical relevance. This would allow us to compare the performance of our model with certain benchlines.

Methodology:

Feature Extraction: The first step of the project was to extract features from the audio dataset. The audio was loaded using the Librosa Library, and the tracks were truncated to match the same length to ensure consistency.

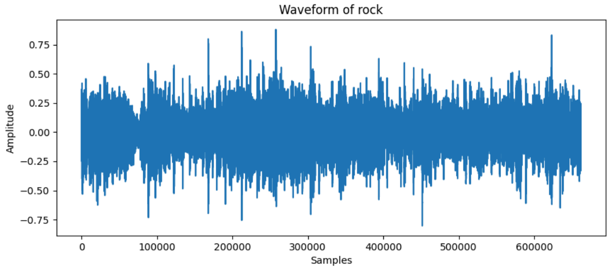

Fig 1. A wave form of a Rock song

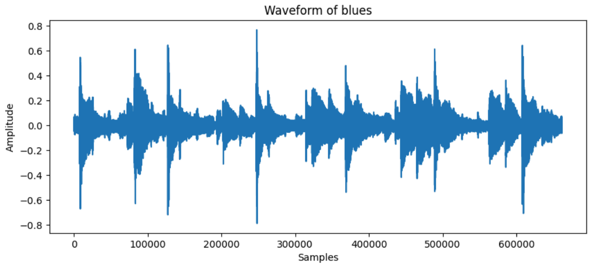

Fig 2. A wave form of a blues song.

The following features were initially extracted:

1. **Zero-Crossing Rate (ZCR)**
2. **RMS Energy**:

3. **MFCCs**.
4. **Spectral Features**: Including centroid, bandwidth, contrast, and rolloff.
5. **Chroma Features**

Zero-Crossing Rate: This represents the rate at which a signal changes its sign from negative to positive, and vice versa. This is crucial to distinguish between low voiced and high voiced sound. It is commonly used to classify percussive sounds.

Root Mean Square (RMS) Energy: This computes the square root of the mean of the squared amplitudes of the signal in a frame. It captures the power of a signal over time. It therefore reflects the loudness of the song.

MFFCs: The Mel-frequency cepstral coefficients are a set of features that represent the short term power spectrum of the sound. They are calculated by first taking the fourier transformation of a signal and mapping it onto a mel scale. This is then logged, which is called a Mel Spectrogram. A discrete cosine transformation of this is the MFCC. It is thought that this most closely mimics the sound range of humans. There are up to 20 MFCCs, we have taken the first 13, which are the ones which are the standard for capturing the core spectral features of sound.

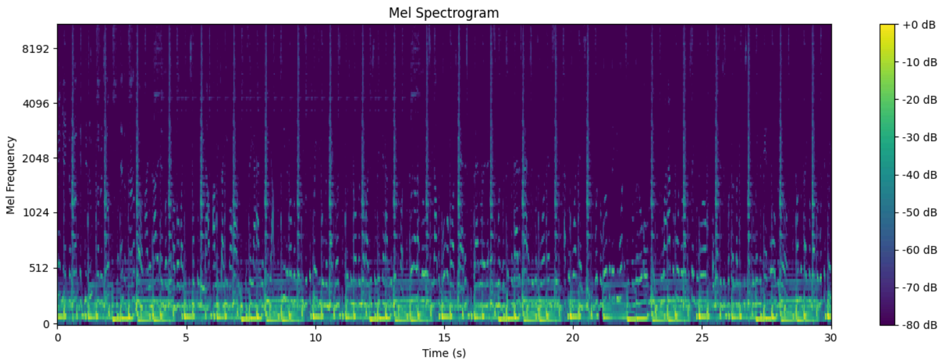

Figure 3. Example Mel Spectrogram of a hiphop song

Spectral Centroid: The Spectral Centroid is a frequency-domain feature that represents the "center of mass" of a spectrum. It is calculated by using weighted mean of the frequencies where the weights are their respective magnitudes of the frequencies. The Spectral Centroid can be taken as the "brightness" of a sound, with higher values indicating brighter or sharper tones and lower values indicating a dull or bass-heavy sound.

Spectral Bandwidth: Spectral Bandwidth measures the spread of frequencies around the Spectral Centroid, indicating the songs harmony and percussiveness . A wider bandwidth indicates more diverse frequencies and is characteristic of energetic genres like Metal or Disco, while a narrower bandwidth corresponds to simpler, harmonic-rich genres like Classical or Jazz.

Spectral Rolloff: The Spectral Rolloff is the frequency below which 85% of the total spectral energy is concentrated. This feature indicates whether the energy is concentrated in lower frequencies, such as bass heavy genres like Reggae or higher frequencies like in pop.

Chroma Features: Chroma Features represent the intensity of the 12 pitch classes in an octave, mapping the audio's harmonic content.

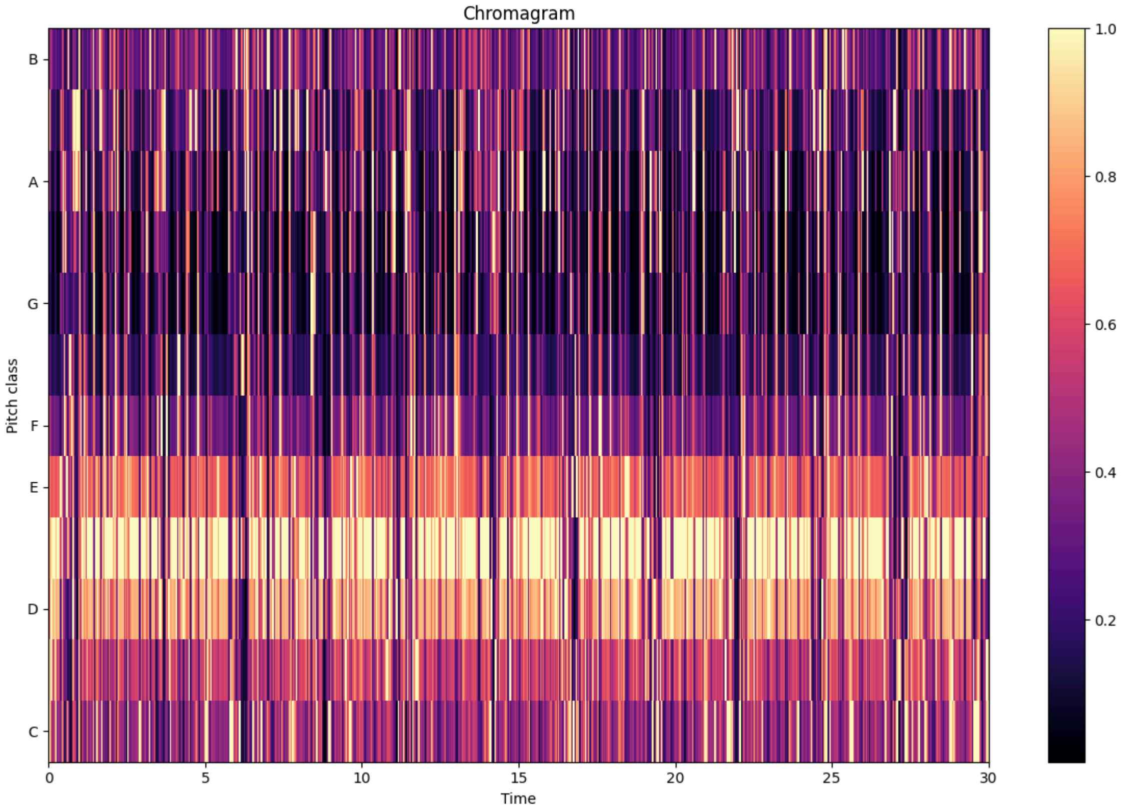

Fig 4. Chroma Features of a hiphop song

All of these features are derived from a time-series representation of the audio signal. Therefore each feature would need to be represented as its own vector. This would be computationally intense and incompatible with many of the models that we use. We therefore aggregate these features into a fixed-length representation. This not only reduces the dimensionality but ensures consistency of the audio. They are also robust against sudden changes in tempo. We aggregate all the features by calculating the mean in order to capture the central tendencies of the data. For MFCCs we also calculate the standard deviation, in order to measure the spread of the data.

Feature	**Aggregation Method**	**No of Extracted Features**
Zero-Crossing Rate	Mean	1
RMS Energy	Mean	1
MFCCs	Mean and std	26(13 mean + 13 std)
Spectral Centroid	Mean	1
Spectral Bandwidth	Mean	1
Spectral Contrast	Mean	7
Spectral Rolloff	Mean	1
Chroma Features	Mean	12

This leads to a total of 50 features for training.

Exploratory Data Analysis

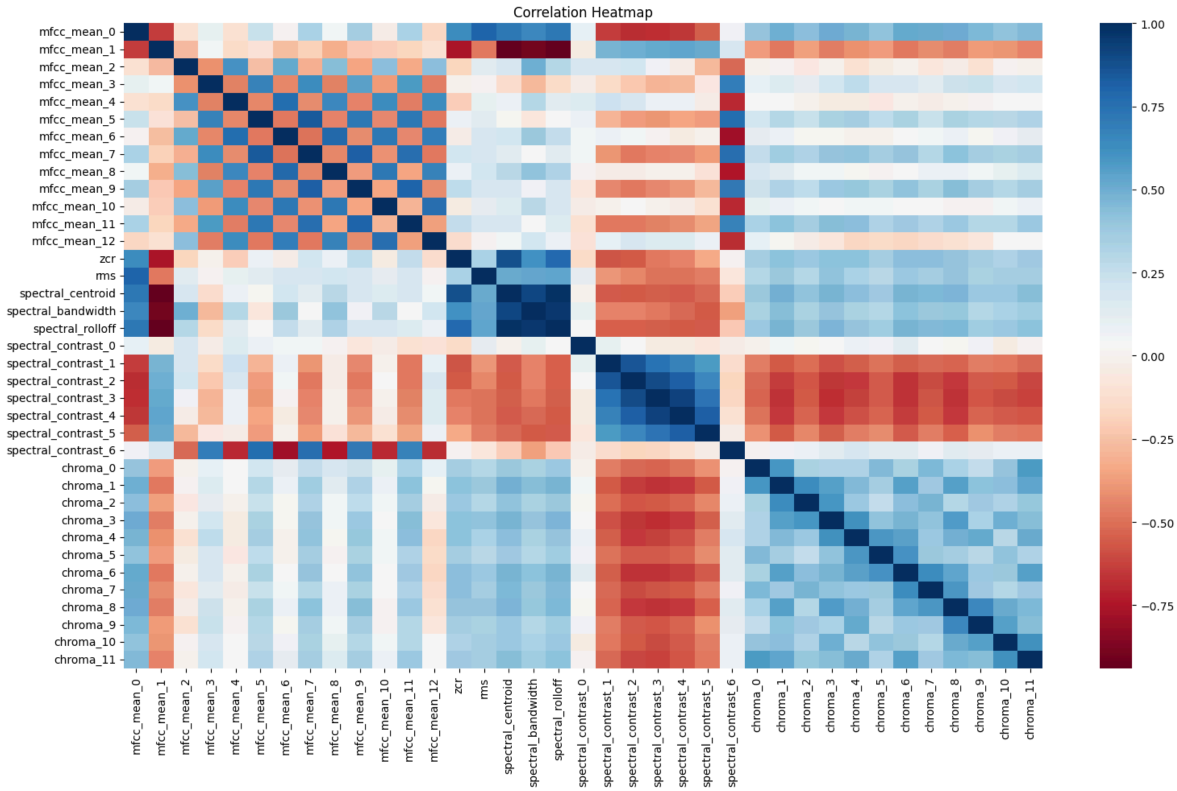

Fig 5. Correlation heatmap of the Data

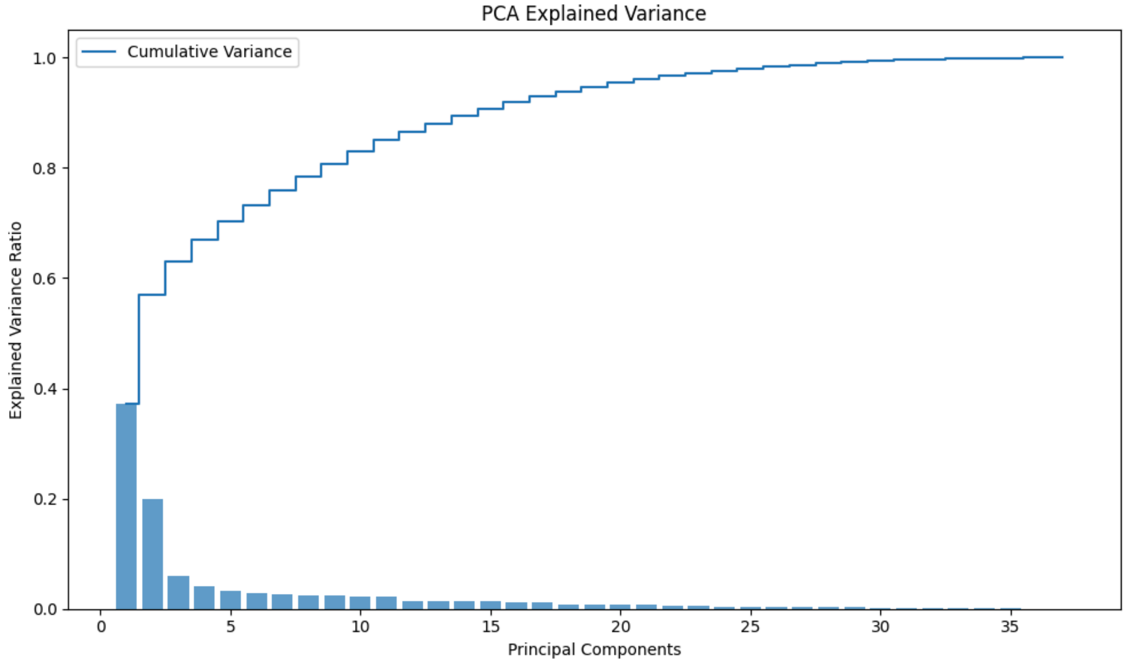

Fig 8. PCA

Results

Logistic Regression

A basic soft max regression was used as a baseline for comparison. After some tweaking, it managed to achieve a test accuracy of 73.67% and a cross validation accuracy of 66.71%. This was surprisingly high given the simplicity of the algorithm. This was already higher than the original 61% achieved in the GTZAN study. This showcases the importance of feature selection, which can elevate even basic models. To explore the role of dimensionality, we repeat the exercise using features from the Principal Component Analysis achieving a test accuracy of 67.33% and a cross-validation accuracy of 67.43%. This highlights the trade-offs between dimensionality reduction and model performance. However, this means that there isn't a significant overfitting in the original model.

K-Nearest Neighbors (KNN)

KNNs were chosen for its simplicity and intuitive modelling. It works by identifying the **k** nearest neighbors of a data point in the feature space and assigning the majority class among the neighbors to the point. Since we are working within a supervised learning context, there is no training as such, making it a good test of the feature selection. Using the default Euclidean Distance, it achieved a Test Accuracy of 65.00% and a Cross-Validation Accuracy: 64.14% (5-fold CV). It performed decently, well given that KNN is sensitive to high dimensional data and does not weigh certain features over others.

Support Vector Machines (SVM)

Support Vector Machines (SVMs) are among the most effective and widely used algorithms for classification tasks, including music genre classification. SVMs work by finding a hyperplane that best separates the data into different classes while maximizing the margin between the classes. We initially trained using an RBF kernel to handle non-linearity, finding an accuracy of 74%. We then trained multiple models using Grid Search to find the best hyperparameters for the model. This further improved the accuracy to 76.67% and a Cross-Validation Accuracy of 73.71%.

Multilayer Perceptron (MLP)

A Multilayer Perceptron is one of the simplest neural networks, and was employed to explore the potential of deep learning for music genre classification. MLPs consist of multiple layers of interconnected neurons, where each layer applies an activation function to the weighted sum of its inputs, allowing the model to capture non-linear patterns in the data. We trained several different combinations of neurons and activation functions, with the best one achieving an Accuracy of 72.67%.

Random Forest

Random Forest is an ensemble method which combines the average prediction of several decision trees to come up with a classification. Each ensemble is trained on a randomly sampled subset of the data, and the final classification is determined by majority voting. This inherent randomness makes random forest robust to noise. We tried several depths and Tree combinations, eventually arriving at a accuracy of 71%.

Results Summary

Model	Test Accuracy	Cross-Validation Accuracy
Logistic Regression	73.67%	66.71%
KNN (k=10)	65.00%	64.14%
SVM	76.67%	73.71%
MLP	72.67%	72.57%
Random Forest	71.33%	69.14%

Conclusion

Support Vector Machines achieved a test accuracy of **76.67%** and a strong cross-validation accuracy of **73.71%**. It was able to leverage non-linear features in the data. This demonstrates the versatility of SVMs when it comes to music classification. The Logistic Regression performed surprisingly well, achieving a test accuracy of **73.67%**, outperforming models such as KNN and Random Forest. This was likely because the parameters were linearised to reduce dimensionality. It came close to Deep learning models on the same dataset which achieved about 75% accuracy. Even feature rich CNN models only got to about 81%. Achieving close to these numbers using simpler models highlights the effectiveness of feature selection. This was especially true because the dataset only had 100 examples for each genre, limiting what a complex algorithm can gather.

Limitations and Further Research

The GTZAN dataset contains only 1000 tracks across 10 genres. This limited size may not fully capture the diversity of music genres in real-world scenarios, restricting the generalizability of the models to other types of music. Music genres often share overlapping characteristics, making classification inherently challenging. This overlap limits the achievable accuracy, even with robust models and feature selection. The extracted features, such as MFCCs and spectral attributes, are aggregated over the entire track, losing temporal dynamics and transitions. This could hinder the model's ability to distinguish genres that rely on changes over time, such as Classical music. Attempts to capture multidimensional data were limited by computational power and thus this made it difficult to implement Deep Learning models, as originally planned.