Music Genre Classification: Accuracy Scores and Feature Selection

James Wang

March 2024

Abstract

In this study, I conducted an extensive examination of three distinct machine learning algorithms, each subjected to three levels of feature selection, to ascertain their efficacy in classifying short musical samples by genre. Additionally, the investigation included an exploration of an ensemble method alongside a variable number of features. Upon the completion of my empirical analyses, it was determined that the Multilayer Perceptron (MLP) algorithm, when applied in conjunction with a substantial degree of feature selection, exhibited the highest accuracy. It is noteworthy, however, that the Random Forest algorithm, under equivalent conditions of feature selection, demonstrated a competitive performance relative to the MLP algorithm.

1 Introduction

The principal objective of this project was to scrutinize the application of machine learning techniques for music genre classification. Given the subjective nature of music genre categorization, wherein a single composition may be ascribed to multiple genres, often with divergent opinions on such classifications, this investigation sought to determine the computational capability to perform this task with efficacy. My inquiry revealed that, through the application of judicious feature selection, machine learning algorithms can predict the genre of an audio sample with approximately 85% accuracy, with the Multilayer Perceptron algorithm emerging as the most precise.

The focal challenge of this scholarly endeavor was the classification of musical genres within the GTZAN dataset, often likened to the MNIST dataset within the realm of music-related machine learning inquiries. This dataset comprises 30-second excerpts from tracks spanning ten distinct music genres, with each genre represented by over a thousand samples. The preliminary phase involved meticulous data preprocessing, subsequent to which three eminent machine learning models were employed for training. Post-training, the models' performance was rigorously evaluated through the computation of accuracy metrics and confusion matrices. Moreover, the research ventured into algorithmic fusion to enhance predictive accuracy and probed the impact of varying the feature set dimensionality on the classification outcomes.

2 Methodology

The inaugural phase of the inquiry entailed a rigorous feature selection process applied to the dataset, aimed at refining the experimental outcomes. Initially equipped with a predefined set of features, my endeavor extended to the derivation of three distinct datasets, each a product of the bespoke feature selection methodology. Subsequent to this preparatory stage, my attention pivoted to the selection of computational algorithms for analysis. Drawing inspiration from a foundational study, my selection included support vector machine and random forest algorithms, with aspirations to either replicate or surpass the benchmark results through enhanced feature selection. Furthermore, my exploratory ambit expanded to incorporate the multilayer perceptron algorithm, a novel addition to the algorithmic repertoire. This comprehensive analysis encompassed the application of each algorithm

across varying feature selection tiers, meticulously evaluating their efficacy in genre classification. The methodology also encompassed an examination of the impact of feature quantity, spanning a broad spectrum from ten to six thousand features. In a concluding venture, I amalgamated the multilayer perceptron and random forest approaches to assess potential synergistic effects on accuracy, particularly in genres where each method demonstrated disparate levels of precision.

The paramount facet of the investigation centered on the meticulous selection of features. Initially engaging with three datasets, the original dataset of 30 seconds provided by GTZAN, all features, and selected features, my analytical endeavors led to the genesis of an additional algorithm, fundamentally an ensemble method amalgamating support vector machine and random forest models. Procurement of the dataset from Kaggle marked the inaugural step, availing us of a preprocessed compendium equipped with sixty features integral to music recognition endeavors, including but not limited to Chroma Short-Time Fourier Transform (STFT) mean and zero-crossing rate mean. Subsequently, the employment of the openSMILE Python library facilitated the extraction of an expansive array of 6000 features from the audio samples, underscoring the library's utility in audio analysis. The ensuing phase involved the stratification of these features by their significance, employing a random forest analytical framework. This entailed the deployment of a random forest model, leveraging the RandomForest.feature_importances_ attribute to curate a hierarchy of feature importance, culminating in the selection of the 2000 most salient features for each audio track. Furthermore, my exploration extended to varying the quantity of pertinent features, ranging from a minimal ten to the entirety of the 6000 extracted. Post-dataset refinement, the initialization, training, and execution of the algorithms were conducted, adhering to an 80/20 training/testing partition and uniform hyperparameters across all models, as delineated in Table 1. This methodological constancy was pivotal, ensuring the comparability of algorithms and the assessment of diverse feature types and quantities, distinctly separate from the influence of hyperparameters on algorithmic performance or dataset integrity.

Table 1: Hyperparameter values used.

Algorithm	**random_state**	**max_iter**
Multilayer Perceptron	42	1000
Support Vector Machine	42	n/a
Random Forest	42	n/a

2.1 About openSMILE

OpenSMILE is a widely recognized and versatile framework for audio processing and feature extraction, designed to facilitate the comprehensive analysis of speech and music signals. It is renowned for its ability to extract a vast array of features from audio inputs, spanning from basic signal properties to complex spectral, voice, and emotion-related characteristics. The framework is engineered to support researchers and developers in tasks related to audio analysis, including but not limited to music information retrieval, speech recognition, and effective computing, by providing an extensive, configurable toolkit for feature extraction from audio data.

3 Empirical Results and Discussions

3.1 Overall Results

After training and testing the main three data sets on the three algorithms and the ensemble algorithm, I achieved the accuracies contained in Table 2 below.

Table 2: Overall accuracies on three main datasets

Dataset	**SVM**	**MLP**	**RF**	**Ensemble**
Original dataset	70%	75%	75.5%	77%
openSMILE features	81.5%	82.5%	86.5%	80.5%
Selected 2000 features	83%	84%	86%	83%

3.2 Compared to Similar Research

In the paper authored by Shah et al. 2022[1], the authors using the same dataset delineate the accuracy levels as enumerated in Table 3. Contrary

to employing openSMILE, the tool of choice for Shah et al. is Librosa. Notably, openSMILE is a tool that is specifically tailored for the extraction of features from audio signals, with an emphasis on applications pertaining to speech and emotional recognition. It enjoys considerable esteem within the scholarly community, particularly for its applications in emotional analysis and speaker identification, among others. This tool is capable of extracting an extensive array of audio features, which includes but is not limited to energy, spectral, voice quality, and prosodic characteristics. It is crafted to facilitate real-time processing as well as analysis of audio on a large scale. Librosa, in contrast, is a Python library that is devoted to the analysis of audio and music. It is designed with an emphasis on user-friendliness and simplicity for tasks related to music and audio analysis. The functionalities it encompasses comprise audio and waveform analysis, extraction of spectral features, beat tracking, and additional tasks. It is particularly adept for use in music information retrieval.

Table 3: Accuracies from the reference

MLP	**RF**	**XGB**	**CNN**
69.1%	66.2%	71.2%	74.1%

Upon review of the data presented in the tables, it is evident that the accuracies achieved by my methodologies surpass those reported by Shah et al. This holds true for the algorithms that were mutually implemented, as well as for the Multilayer Perceptron (MLP) and the proprietary ensemble method.

3.3 Individual Results

In the next few sections, I will explore the calculated confusion matrices from the support vector machine, random forest, multilayer perceptron, and ensemble method with the feature selected dataset containing the 2000 most important features.

3.3.1 SVM - Support Vector Machine

The Support Vector Machine (SVM) model exhibits a strong ability to classify 'metal' (B) with perfect accuracy, indicating that the features separating

'metal' from other genres are well-captured by the SVM. The model also performs well on 'pop' (A), 'disco' (C), and 'hiphop' (H), although there is some confusion between these genres and 'jazz' (J) and 'country' (I). The most significant challenge for the SVM model is the classification of 'jazz' (J), which is often misclassified as other genres, suggesting that the boundaries between 'jazz' and other genres are not as clear-cut in the feature space used by the SVM.

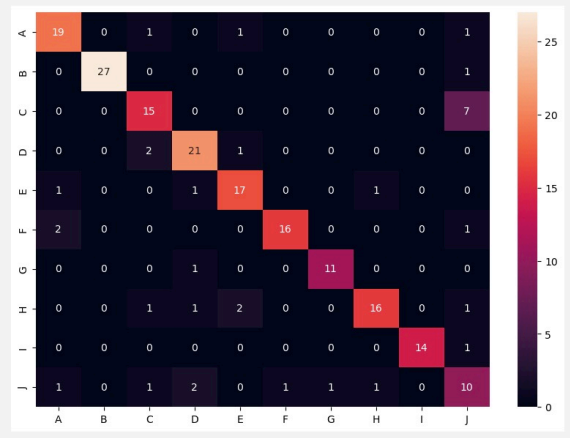

Figure 1: Confusion Matrix for Support Vector Machine with Feature Selected Dataset

3.3.2 RF - Random Forest

The Random Forest (RF) classifier maintains the trend of accurately classifying 'metal' (B), with no misclassifications observed. However, the model shows a slightly different pattern of confusion compared to the SVM, particularly between 'rock' (G) and 'hiphop' (H), indicating that these genres share similar features that the RF model finds difficult to distinguish. 'Disco' (C) and 'blues' (D) are classified with high accuracy, showing that the RF model captures their unique characteristics well. Yet, 'jazz' (J) continues to be problematic, with misclassifications distributed across multiple genres.

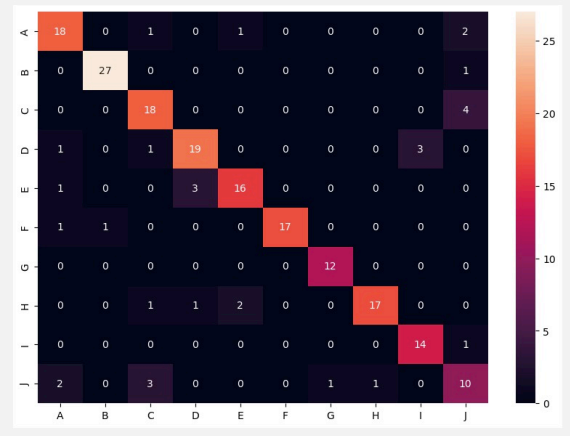

Figure 2: Confusion Matrix for Random Forest with Feature Selected Dataset

3.3.3 MLP - Multilayer Perceptron

The Multi-Layer Perceptron (MLP) model, a type of neural network, generally improves on the classification of 'jazz' (J) compared to the previous models, though it is still the most misclassified genre. The model excels in classifying 'disco' (C) and 'blues' (D), with the highest correct classifications among the models for these genres. An interesting note is the misclassification of 'classical' (F) with 'pop' (A), which the other models did not exhibit, pointing to a unique aspect of how the MLP model interprets the features of these genres.

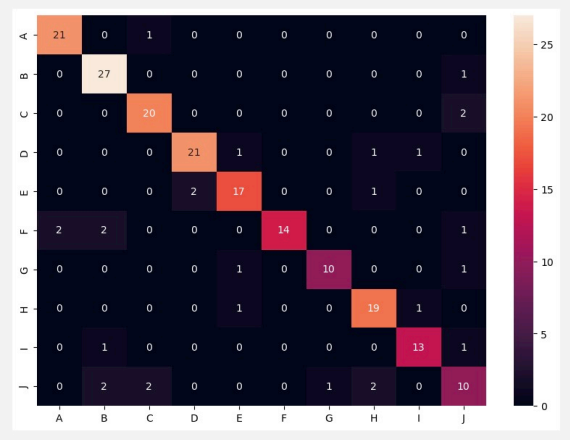

Figure 3: Confusion Matrix for Support Vector Machine with Feature Selected Dataset

3.3.4 Stacking - Ensemble Method

The stacking classifier, which combines the predictions of the SVM, RF, and MLP models, shows a robust performance across most genres, especially 'metal' (B), which it classifies with perfect accuracy. 'Pop' (A) and 'disco' (C) also see strong performance, reinforcing the stacking model's ability to capture the defining features of these genres. However, 'jazz' (J) remains a challenge, underscoring its complex nature which may require a more nuanced approach or additional features to accurately classify.

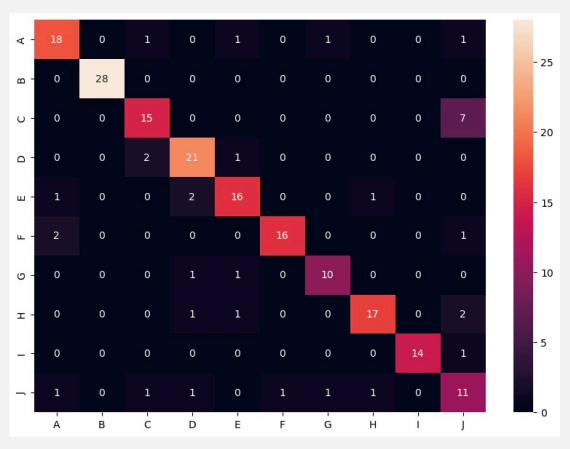

Figure 4: Confusion Matrix for Ensemble Method with Feature Selected Dataset

3.4 Discussion

Overall, 'metal' (B) is the genre best recognized by all models, suggesting that its musical characteristics are distinct and well-represented by the features used in the classification task. The consistent difficulty in classifying 'jazz' (J) across all models indicates that 'jazz' may share overlapping features with other genres or that the genre is more varied within itself, making it harder for models to find a clear boundary. The variability in the performance of 'classical' (F) and the confusion between 'rock' (G) and 'hiphop' (H) in some models could be due to the models' different handling of features or suggest that a more diverse set of features is needed to better differentiate these genres. This analysis highlights the importance of selecting or engineering features that capture the essence of musical genres and the potential benefits of ensemble methods like stacking to improve classification performance.

The MLP's layered structure allows it to learn hierarchical representations of data. In the context of music genre classification, this means that an MLP can learn to recognize complex patterns in the spectral and temporal features of audio that simpler models may miss. For instance, the MLP might be able to discern subtle differences in rhythm, melody, and harmony that define a genre, which could explain its improved accuracy in classifying 'disco' (C) and 'blues' (D). Additionally, MLPs are capable of feature learning, which means they can identify which aspects of the data are the most informative for

classification tasks, potentially giving them an edge in distinguishing between genres with overlapping characteristics like 'jazz' (J) and 'pop' (A).

MLP's deep learning framework is designed to handle large volumes of data and automatically extract features at multiple levels of abstraction. In music, this could involve learning low-level audio features in the early layers (such as pitch and timbre) and higher-level musical concepts (like tempo and genre-specific patterns) in the deeper layers. This hierarchical feature learning is particularly suited to music genre classification, where nuances can be subtle and deeply embedded in the structure of the music.

The mathematical models, such as SVM and RF, rely on predefined equations and decision rules to classify data. SVM finds the optimal separating hyperplane that maximizes the margin between different classes, and RF builds multiple decision trees to vote for the most popular class. While these models are powerful and have their own strengths, they may not capture the depth of complexity within musical data as well as an MLP can. Moreover, SVM and RF require careful feature engineering to perform well, whereas an MLP can automatically learn to represent the features through its multiple layers, which is particularly beneficial for high-dimensional data like audio signals.

The effectiveness of MLP in music genre recognition underscores the strength of deep learning models in capturing the complexities of audio data. Its ability to learn at multiple levels of abstraction and to handle a vast array of nuanced features makes it well-suited for tasks that involve rich, multi-dimensional datasets. While SVM and RF are grounded in solid mathematical principles and provide robust baseline models, the deep learning approach of MLP offers a pathway to unravelling the intricate patterns found in music, leading to potentially more accurate and insightful genre classifications.

4 Exploring Feature Selection

The final aspect warranting discussion pertains to the investigation of feature quantity variations within the dataset. Employing a random forest analytical framework, I meticulously ranked the 6000 features extracted via openSMILE by their relative significance. The initial testing phase utilized a dataset comprising 2000 salient features. Subsequent to this phase, a decision was reached to re-evaluate the performance of the algorithms, this time incorporating datasets with a varied feature count. This was accomplished

by generating additional datasets, each encompassing the uppermost n features as determined by the initial feature importance ranking. The spectrum of feature quantities examined extended from a minimal selection of 10 to the entire set of 6000. Depicted in Figure 5, the graph delineates the accuracies of the four algorithms (including the ensemble method, denoted as "stacking") as a function of feature set size. The same division for testing and consistent hyperparameters were applied. An initial surge in accuracy is observable at the inclusion of the first 10 features; thereafter, a plateau becomes evident. It is noteworthy that the Multilayer Perceptron (MLP) algorithm consistently delivered superior accuracy, particularly at the 2000 feature mark, corroborating its preceding performance. Indeed, the MLP algorithm demonstrated enhanced efficacy in comparison to its counterparts across nearly the entire range of feature quantities examined.

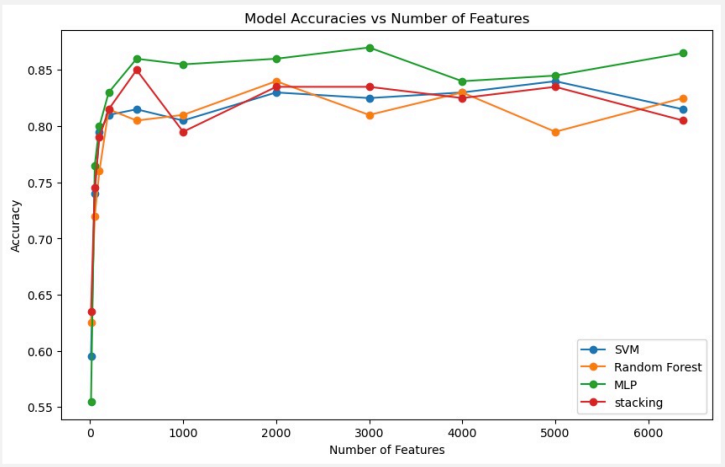

Figure 5: Model Accuracy Plotted Against Number of Features

5 Potential Bias and Future Work

In my music genre recognition machine learning project utilizing openSMILE for feature extraction and the GTZAN dataset from Kaggle, several potential biases need to be addressed to ensure the robustness and generalizability of my model. Firstly, the GTZAN dataset, while popular for music genre classification tasks, has been critiqued for its limitations, including a small

size, potential mislabelings, and a lack of diversity in terms of musical genres, cultural representation, and recording quality. These factors can introduce biases that might hinder my models' ability to generalize across different music types and recording conditions. Additionally, the inherent bias in the selection of tracks and genres can skew the model's performance, favoring more represented genres over others and potentially underperforming on musical styles that are not well-represented in the dataset.

Addressing these biases requires a multifaceted approach in future steps of my project. Expanding the dataset with additional genres, more diverse cultural music, and varying recording qualities can help mitigate some of the dataset-specific biases. It's also beneficial to employ data augmentation techniques to artificially increase the variety within the dataset, thereby enhancing the model's ability to generalize. Furthermore, critical evaluation of the model's performance across different subgroups within the data, followed by fine-tuning or employing more sophisticated models, can address potential biases introduced by the feature extraction process or the learning algorithm itself. Continuous validation of the model's performance on external datasets, preferably those with a broader and more diverse music collection, will be crucial for assessing the model's robustness and generalizability across various musical genres and recording conditions.

6 Conclusion

In the conducted study, satisfactory outcomes were achieved employing three distinct algorithms. Initially, utilizing a pre-processed dataset provided, an apex accuracy of 75.5% was attained employing both a Multilayer Perceptron (MLP) and a Random Forest algorithm. Subsequent to the initiation of data processing enhancements, the MLP algorithm demonstrated an improved maximum accuracy of 86.5%. In the concluding phase, the MLP algorithm consistently exhibited superior performance, achieving an accuracy of 86%. Comparative analysis revealed the MLP algorithm as the most efficacious overall, notwithstanding the Random Forest algorithm's marginally superior performance in classifying certain musical genres. Conversely, the Support Vector Machine (SVM) algorithm consistently underperformed relative to its counterparts. The findings from this investigation underscore the potential of machine learning algorithms in effectively recognizing musical genres, albeit with accuracy significantly influenced by the specific genre

and the selection of features. Future research could beneficially explore the optimization of algorithmic selection tailored to specific genres, potentially employing a combination of algorithms to enhance overall accuracy in music genre classification.

References

Shah, Mitt et al. (2022). "Music Genre Classification using Deep Learning". In: *IEEE*.