Final Report: Music Genre Identification using Machine Learning

Noah Willis and James Wang

December 2023

Abstract

We explored three different machine learning algorithms with three different levels of feature selection to determine which would be most effective at identifying the genre of a short music sample. We also experimented with an ensemble method and with varying numbers of features. After running all of our experiments, we determined that an MLP algorithm with a high level of feature selection is most accurate, although a random forest algorithm with the same feature selection is able to compete with the results from the MLP algorithm.

1 Introduction

The main goal for out project is to investigate using machine learning to identify the genre of music. We thought this would be an interesting project because it can be hard to pinpoint exactly what genre any given song can be attributed to many different genres, oftentimes with different people disagreeing on what those genres are. As such, we wanted to evaluate if a computer would be able to perform this task effectively, and were curious what the best algorithm would be. Overall, we found that machine learning algorithms paired with good feature selection on a dataset guesses the genre of a given audio sample with an 85% accuracy, with a multilayer perceptron algorithm being the most accurate overall.

2 Problem Description

The specific problem that we tackled in this final project was to identify the genre of all of the songs in the GTZAN dataset. This set can be thought of as the MNIST set of music-related machine learning problems. It contains 30 second long samples of songs from ten music genres, with over 1000 samples from each genre. We pre-processed the data and then used it to train three popular machine learning models. After training and running the algorithms we calculated the accuracy and confusion matrices for all three algorithms and compared the results. We also experimented with combining multiple algorithms to improve accuracy and tested out the effect of using a different number of features.

3 Approaches

The first step in our project was to perform feature selection on our dataset in order to improve our experimental results. The dataset came with an initial feature set, and we performed our own feature selection on the dataset to get three processed datasets for use in testing. After performing feature selection, the next step in our project was to select which algorithms we would run on the dataset. The paper we based our project on used a support vector machine algorithm and a random forest algorithm, so we decided to use these algorithms as well. We wanted to see if we would be able to reproduce or improve the results from the paper through the use of better feature selection, as we mentioned earlier. Additionally, we wanted to run a new algorithm on our dataset, so we decided to use a multilayer perceptron algorithm. We then ran all three algorithms on all three levels of feature selection, comparing the results to see which combination of feature selection and algorithm proves to be most accurate at guessing the sample genre. Additionally, we experimented with using different numbers of features, from ten to six thousand. Lastly, we also blended together the MLP and RF methods to test if this would have any improvement on accuracy given that both of these methods achieved different accuracies on some genres.

4 Experimental Setup

The most important aspect of our project was our feature selection. As mentioned earlier, we had three datasets that we initially used, and after running tests we created an additional algorithm, which is essential an ensemble method using our support vector machine and random forest models. The first step was to actually get our dataset, which is provided by Kaggle. The dataset came with the data preprocessed, giving up a set of sixty features to train our algorithms on. These features are all used often in music recognition, such as the Chroma Short-Time Fourier Transform (STFT) mean and the zero-crossing rate mean. The next step we took was using the Python library openSMILE to extract more features from our dataset. openSMILE is a library specifically designed for audio analysis and processing, and we were able to extract 6000 features from our data samples. Lastly, we wanted to sort these features by how meaningful they are, so we used a random forest analysis technique to do so. This was a fairly simple process, and involved training and running a random forest algorithm on the dataset, and using the `RandomForest.feature_importances_` property to sort the features in order of importance and then taking the top 500 of the features from each audio track to represent our dataset. We also experimented with using a different number of important features from 10 to the 6000 total

After processing our datasets, we initialized, train, and ran all three of our algorithms. We used a 80/20 train/test split for all three algorithms, and used similar hyperparameters for all three. Table 1 below shows the hyperparameters used for each algorithm. We decided to hold the hyperparameters constant across all datasets and algorithms because the goal of our project was to compare the algorithms to each other and to compare the effectiveness of different types and numbers of features, not to investigate the effect hyperparameters have on each algorithm or dataset.

Table 1: Hyperparamter values used.

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

5 Experimental Results

5.1 Overall Results

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

Table 2: Overall accuracies on three main datasets.

Dataset	**SVM**	**MLP**	**RF**	**Ensemble**
Original dataset	70%	75.5%	75.5%	76.5%
openSMILE features	81%	83.5%	80%	81.5%
Important features	81.5%	85%	84.5%	83.5%

For reference, Table 3 below shows the accuracies achieved by the paper[1] we based our project on. We have only included their accuracy for easier comparison to our results.

Table 3: Accuracies from paper.

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

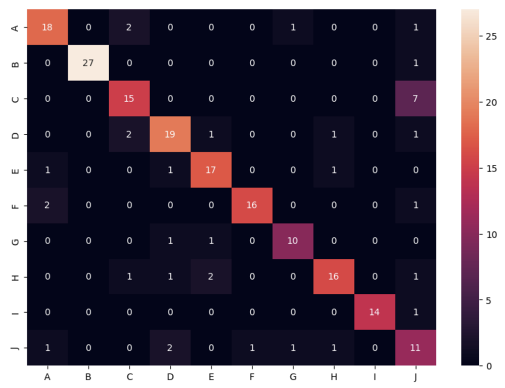
As can be seen from the tables above, we were able to achieve far better accuracies than the authors in both the algorithms we both implemented along with MLP and our ensemble method.

5.2 Individual Results

In the next few sections, we will exploring the calculated confusion matrices from our support vector machine, random forest, multilayer perceptron, and ensemble method with the feature selected dataset containing the 500 most important features.

5.2.1 Support Vector Machine

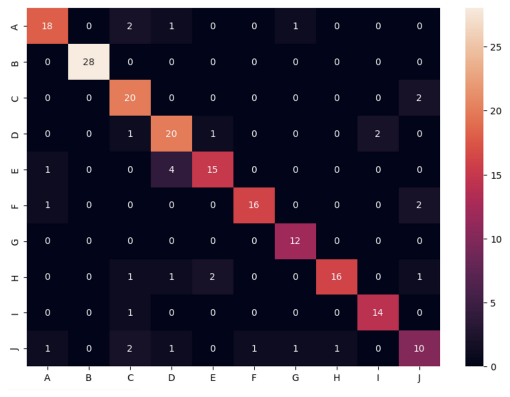
The support vector machine algorithm gave an overall accuracy of 81.5%, correctly assigning the genre of 163 songs. The most accurately guessed genres with this algorithm where metal, represented by B, and country, represented by I. Both of these genres had a 100% success rate. On the other hand, jazz music, represented by J, had the worst accuracy, with only 45% success, followed by disco, represented by C, with a 75% accuracy.



	A	B	C	D	E	F	G	H	I	J
A	18	0	2	0	0	0	1	0	0	1
B	0	27	0	0	0	0	0	0	0	1
C	0	0	15	0	0	0	0	0	0	7
D	0	0	2	19	1	0	0	1	0	1
E	1	0	0	1	17	0	0	1	0	0
F	2	0	0	0	0	16	0	0	0	1
G	0	0	0	1	1	0	10	0	0	0
H	0	0	1	1	2	0	0	16	0	1
I	0	0	0	0	0	0	0	0	14	1
J	1	0	0	2	0	1	1	1	0	11

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

5.2.2 Random Forest



	A	B	C	D	E	F	G	H	I	J
A	18	0	2	1	0	0	1	0	0	0
B	0	28	0	0	0	0	0	0	0	0
C	0	0	20	0	0	0	0	0	0	2
D	0	0	1	20	1	0	0	0	2	0
E	1	0	0	4	15	0	0	0	0	0
F	1	0	0	0	0	16	0	0	0	2
G	0	0	0	0	0	0	12	0	0	0
H	0	0	1	1	2	0	0	16	0	1
I	0	0	1	0	0	0	0	0	14	0
J	1	0	2	1	0	1	1	1	0	10

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

The random forest algorithm improved on the results of the SVM method by 4%, giving an overall accuracy of 84.5% and correctly recognizing the genre of

169 songs. Once again, metal (B) achieved an accuracy of 100%, with jazz (J) remaining the least accurately guessed genre, this time coming in at 66.66%. Despite jazz still being the least accurate, the random forest algorithm still performed over 20% better on the jazz genre than the support vector machine algorithm.

5.2.3 Multilayer Perceptron

The multilayer perceptron algorithm proved to be the most effective overall, giving an accuracy of 85% across all genres, however it was only marginally better than the random forest method, only guessing one additional song than random forest. For the multilayer perceptron method, classical music, represented by F, was the most accurately guessed genre at 100%. The least accurate method was once again jazz, coming in at 64.3%, 2% lower than the random forest algorithm achieved.

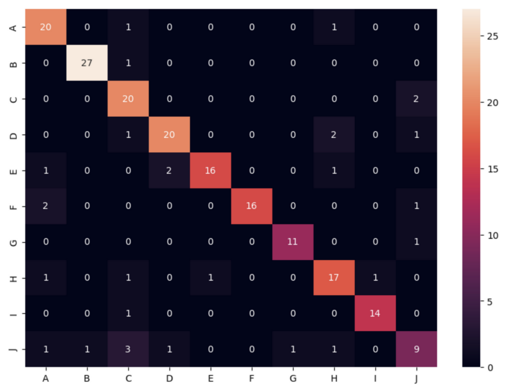


Figure 3: Confusion Matrix for Multilayer Perceptron with Feature Selected Dataset

5.2.4 Ensemble Method

The last method we will be discussing in depth is the ensemble method. This algorithm combines the support vector machine and random forest algorithms using the `StackingClassifier` function. The ensemble method was able to achieve an overall accuracy of 83.5%, strangely lower than both the random forest and multilayer perceptron methods individually. It correctly identified 167 songs, with jazz still being the least accurate genre at 68.75% accuracy and disco (F) achieving 100% accuracy.

5.3 Exploring Feature Selection

The last result we would like to touch on is our exploration of differing numbers of features. As explained in section 4, we used a random forest analysis method

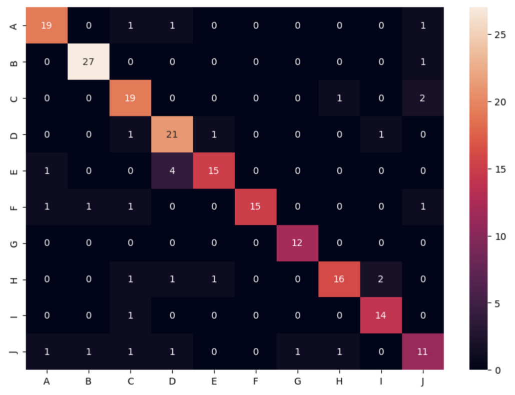


Figure 4: Confusion Matrix for Ensemble Method with Feature Selected Dataset

to sort the 6000 features generated by openSMILE in order of importance. Our final dataset from the above testing used 500 features, but after completing this round of testing we decided to run all of the algorithms again using a varying number of features. To do so, we simply created more datasets based on the original ranking of features, selecting the top n features for any given dataset. The number of features we selected ranged from 10 to 6000. Figure 5 below shows the accuracies of all four of our algorithms (the ensemble method is labelled as "stacking") plotted against the number of features in the set. We used the same test split and hyperparameters as before. As you can see, the

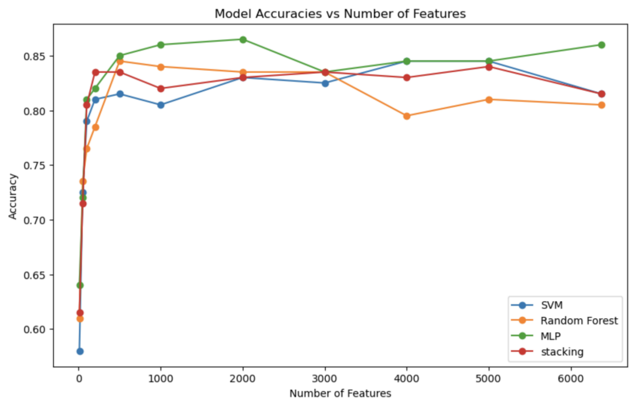


Figure 5: Model Accuracy Plotted Against Number of Features

accuracy sharply increases at 10 features, but then largely flattens out. It seems that the most accurate results came from the MLP method with around 2000 features, which is not surprising based on the previous results. In fact, MLP outperforms the other three algorithms at nearly every number of features.

6 Discussion

Based on all of the evidence collected, it would seem that a multilayer perceptron algorithm is by far the most effective algorithm to use for genre recognition. It outperforms all other algorithms at every level, and while it may not be the best at recognizing every individual genre, its overall accuracy shows that it is the best choice. It is interesting to note that using an ensemble method and combining random forest, the second best algorithm, with our support vector machine algorithm performs worse than RF would have done on its own. This suggests that an ensemble method is not effective for this purpose.

7 Conclusions

Overall, we achieved solid results with all three algorithms. For the first run using the processed dataset that was given to us, we were able to achieve a maximum accuracy of 75.5% using both a multilayer perceptron algorithm and a random forest algorithm. For the second round after we began to process the data, we were able to achieve a maximum accuracy of 83.5% with a multilayer perceptron algorithm, and for the final round we also achieved a maximum accuracy using a multilayer perceptron algorithm, hitting 85% accuracy. Overall, multilayer perceptron proved to be the most effective, however a random forest algorithm slightly outperformed the MLP algorithm on some genres. In all cases, a support vector machine algorithm performed the worst. The conclusion we drew from this project is that overall, machine learning algorithms can do a fairly good job at recognizing music, however it's accuracy is highly dependent on which genre the music actually is. An interesting avenue for future research would be further investigation into which algorithms work best for specific genres and using them concurrently to achieve better overall accuracy.

References

[1] Mitt Shah et al. *Music Genre Classification using Deep Learning.* 2022. DOI: 10.1109/ICCMC53470.2022.9753953.