Music Genre Classification using Machine Learning Models

Raiyan Bhuiya
Department of Computer Science and Engineering
Texas A&M University
College Station, United States
rynbhuiya@tamu.edu

***Abstract*—With music production becoming widely available to the public, the average person can now create and publish songs onto various platforms. The widespread availability has resulted in millions of songs being created that can cross over multiple genres. It is necessary to organize these songs appropriately and make them easier to be found for users. Machine learning models will be used to study and predict the genres of songs to find the most accurate model.**

I. Introduction

With music production becoming more widespread and available for the average person, more songs can be published on various music platforms on the internet. This has resulted in millions of songs being released on platforms such as Spotify, Apple music, and Soundcloud. The large number of songs makes it difficult for users to navigate through the services and find the songs they like. As music progresses and becomes more available, it starts to cross multiple genres or even start to create new ones themselves. This further makes navigation difficult for the users.

This is where machine learning comes in. With the use of machine learning models, we can train them to accurately classify songs and place them into a genre. Within machine learning, there are different types of learning: unsupervised, supervised, and reinforcement. In unsupervised learning, the training data doesn't contain the labels or 'correct answer', so the model needs to figure out the relationships between the data. Supervised is the opposite, where we have the answer and need to focus on minimizing the model's error. In reinforcement learning, the model is placed into an environment with a possible set of actions. Each action may reward, punish, or do nothing. The model needs to figure out the optimal method for maximizing the rewards. For this project, I will be using supervised learning based on the dataset retrieved from GITZAN that contain the genre labels and audio files. The goal is to create a model that can accurately be provided with an audio file's information and return the correct genre label.

When it comes to identifying songs and predicting lyrics, humans are very good at it. Neural network is a machine learning model that mimics the way we learn about something. Much of it is a Blackbox but they are very good at learning precise relationships between the datapoints to make differentiations.

II. Literature Review

Raval, Meet, et al [1] approached the problem of music classification in 2020 with a similar solution as before. Convolutional neural networks have become the norm for music genre classification and that is the model they have chosen for their research. Before training their model, they needed to pre-process the data. First, they started by normalizing the input data which would the Mel spectrograms. These spectrograms contain the loudness of the audio files at different frequencies. This can be used to find the energy levels of the song which helps the model find patterns for the different music genres. Depending on whether the spectrograms contain color, they need to be adjusted so that they can be inputted into the CNN. Lastly, the music genres needed to be one hot encoded to be used as the target values. One difference in their methodology was their method that they split the data in. Instead of randomly splitting the audio files, they did a stratified split. So, for each genre they took an 80:20 split for the training and testing sets respectively. The training set was then used from one to thirty epochs each train the model. The resulting model had a training error ranging from 0.04-0.14 which performed well against older models but not with the current standard.

Castillo and Flores [3] approached this same problem as well but using a different dataset. Instead of the GITZAN dataset, they opted to sample their data from YouTube and then splitting it into 10-second audio files. Before using the entire dataset to train their models, they decided to have a short experimental phase trying out different models. They trained a Decision Tree, a Naïve Bayes classifier, a linear SVM, and neural networks. From the experimental phase the decision tree was taken out due to poor performance. From their research project, they have discovered that the Naïve Bayes and LSTM models were the best performing. With these models, they were able to achieve a precision score in the range of 0.30 to 0.40. In their conclusion, they had an interesting suggestion and that was having YouTube users answer the correct genre for the music videos to help collect labeled data.

Pelchat and Gelowitz [2] build upon the research of another project done by Despois. They take the neural network that was used and add some changes to it. For their research project, they have increased the number of music genres used, increased the number of spectrogram slices per genre used, and have used multiple unique music datasets for their neural network. This is similar to the method that I will be using for my project. The thirty second audio files data is split into three second intervals which allows for more data to be fed into the neural network model. The neural network setup that I have is similar to what I will be using, where there will be multiple convolutional layers with a softmax for classification and a dropout to prevent overfitting. Originally, their model had overfitted resulting in higher training scores but lower test scores. They made the following changes to improve the accuracy: limiting number of genres used, changing the activation function to ReLU, and changing the spectrogram dimensions. Changing the activation function to ReLU resulting in a 5% accuracy in their predictions, which is what I will be using for my model as well. Their final model was able to achieve a test accuracy of 75% which is lower than the other models that I have studied.

When approaching the problem of music genre classification, researchers unanimously agree that neural networks are the optimal choice. For this project, a convolution neural network will be used to tackle this problem.

A. *Maintaining the Integrity of the Specifications*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Problem Formulation

Since anyone with a laptop and the correct software can make songs, the internet is now overflowing with all sorts of unique music. This is a benefit for music because more people can contribute and expand upon the art. The availability of music productions has increased music databases significantly with millions of songs that crossover multiple genres. While this is a benefit for music, it becomes difficult for listeners to find music that they enjoy. There could be thousands of songs out there, but the user won't know how to find it. By implement machine learning models to analyze and predict the genres of these songs, it can be used to organize the databases. Users will be able to easily navigate through the database of services like Spotify, Apple Music, and SoundCloud. More artists will be discovered and listened to since they can be easily found. By introducing machine learning into music databases, both listeners and artists will benefit.

IV. Proposed Solution

From the previous research projects mentioned, the researchers only used spectrogram images generated from the audio files for their model. The main difference is that my project will be using multiple different features of the audio file for the models. The dataset that I have contains a csv file with feature information for each of the audio files. This dataset is for each of the thirty second audio files which is then split into three second audio files giving us ten times the data to be used for training. The training and testing datasets are split randomly without considering the genres for each. Given the large amount of dataset that we have, stratifying using genre isn't necessary.

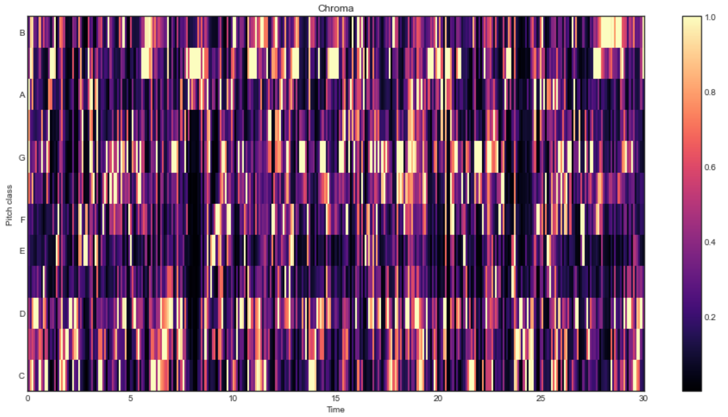
Two models will be used on the dataset, a convolutional neural network, and a support vector machine. Based on previous research history, there is no point in studying the usage of SVMs since neural networks always outperformed or was equal to other models. Still, it is necessary to test other models for comparison.

The SVM model will be ran with the two different classifiers, the one-vs-rest, and one-vs-one classifiers. The one-vs-rest classifier independently models each class against each other. It then creates a binary label whether the data point is within that class. The one-vs-one classifier is meant for multi-class classification, so it is expected to perform better than the other. The linear, poly, and rbf kernels will be used on the dataset to see which performs the best. For the poly model, the degree will be increased to see its effect on the test accuracy. PCA was performed on the dataset which will be used for the rbf and poly kernels SVMs to see if it has any beneficial effects.

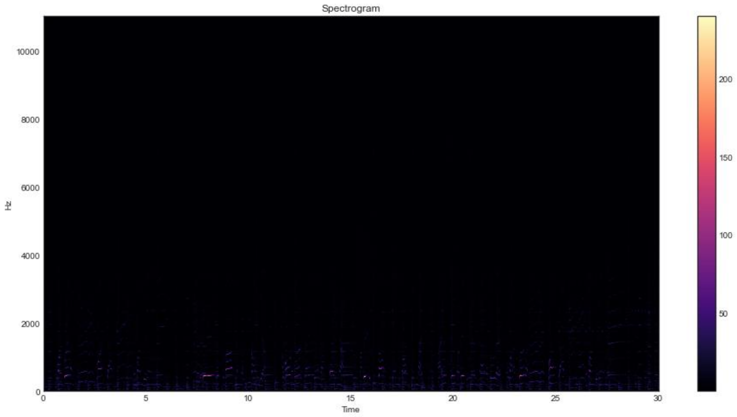
The CNN model contains six convolutional layers with a ReLU activation functions. After each layer, there is dropout layer with a probability of 0.2 was used to prevent overfitting. Finally, there is a softmax layer with an output of 10 for each of the music genres. The model is running with cross entropy for its loss function and adam for its optimizer. Various epochs will be tested to find the best testing accuracy.

V. Data Description

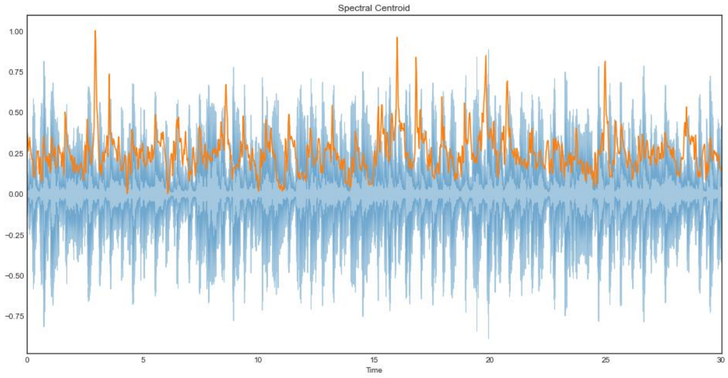
The dataset contains information regarding each of the features for the audio files. The csv file is then split even further to give us more data to be used for training and testing. The first feature is the chroma which represents the musical harmony of the song. It is categorized into the twelve semitones of the octave thus making them ignore the changes from using different instruments. Utilizing the chroma will allow for the model to capture the patterns in the frequency of the notes that are being played.



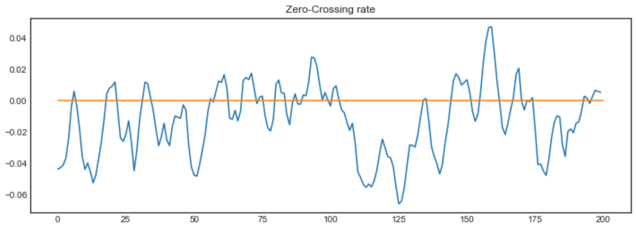
Next, the spectrogram will be used which contains the loudness of the audio files at various levels of frequencies. The spectrogram will allow the models to find patterns in the energy levels for each of the genres.



The spectral centroid contains the location of the sound within the song. It finds the location by calculating the average of the frequencies that are available.



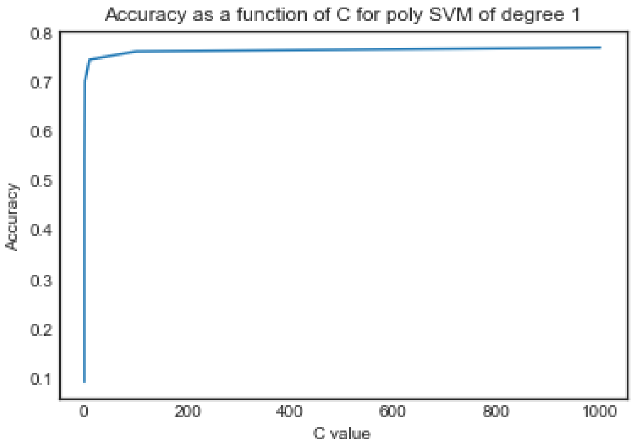
Lastly, we have the zero-crossing rate which is just the number of times that the frequency crosses zero in the audio file.

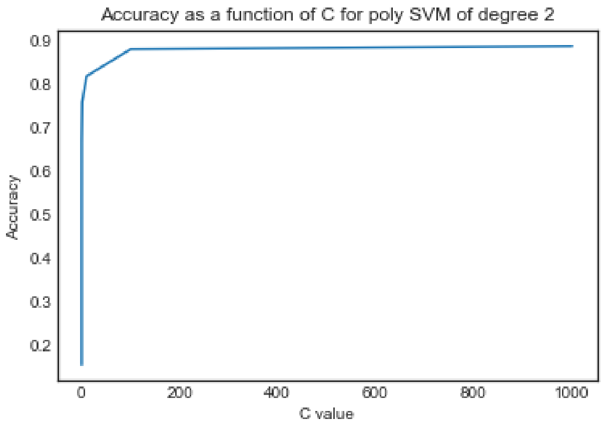


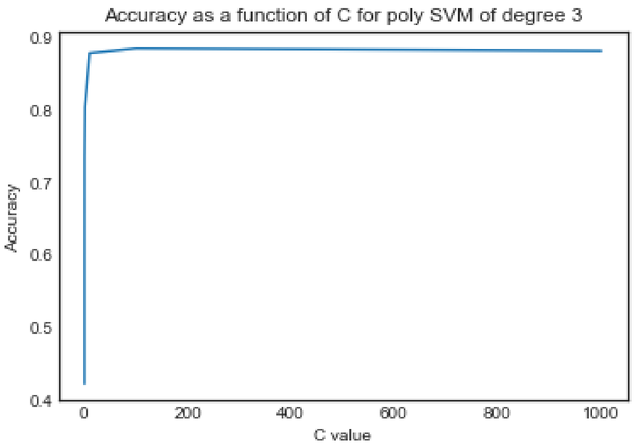
Compared to the other projects, this project will be using much more features in order to train the model.

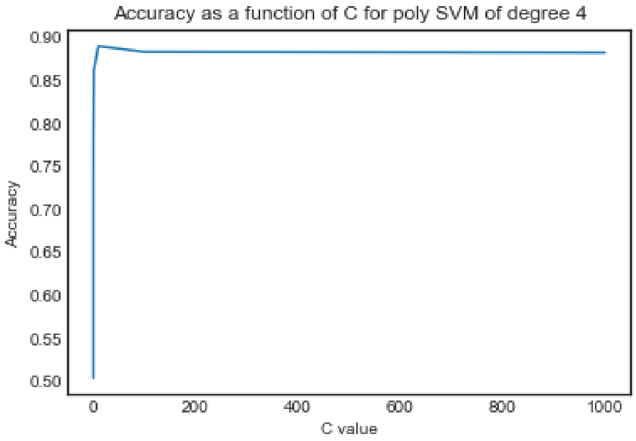
VI. Results

First, the linear kernel svm was ran on the dataset which resulted in a test accuracy of 0.73. The test accuracy is not good when compared to previous research. Next, a poly kernel svm was used with degrees ranging from one to four at different levels of C and the test accuracies were recorded.

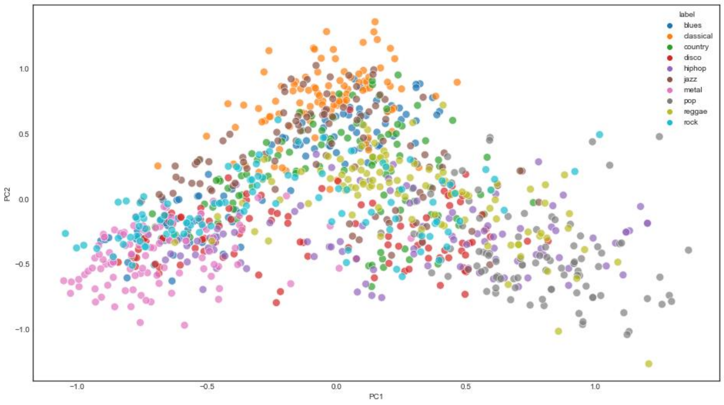






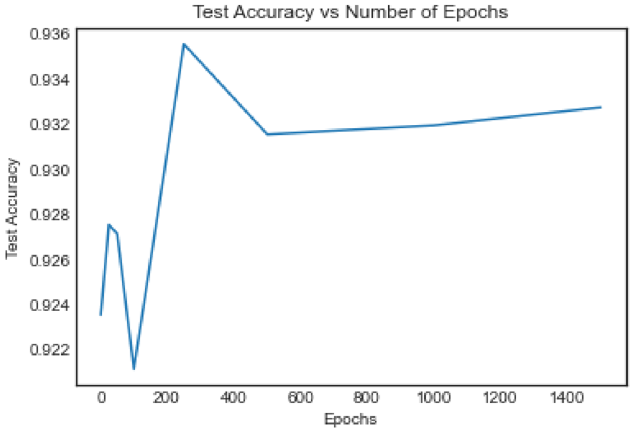


The poly kernel with a degree of 2 at C = 1000, performed the best with a test accuracy of 0.88. This same kernel was then used on the PCA dataset where it got a test accuracy 0.272. The following is the plot of data points after PCA is applied. The rbf kernel was also ran on this data and got a test accuracy of 0.344



Looking at the plot, there is a lot of overlap among the datapoints making it diffcult for support vector machines to make differentiations. This resulted in the models getting poor test scores.

Lastly, the CNN model was trained using the dataset at different epochs at 1, 25, 50, 100, 250, 500, 1000, 1500. The test accuracy was then recorded at then end of each training epoch. The CNN model ran at 250 epochs achieved the highest test accuracy at 0.93. The score is slightly better than the results from previous researchers but not enough to be considered state-of-art.



VII. Limitations and Future Directions

One significant issue is the assumption that there are only ten genres for music classification. In the introduction, it was stated that music production becoming widely available has resulted in genre crossing and even new genres being created. When the dataset was created 20 years ago, the internet was not developed as much as it is today. But now databases are overflowing with millions of songs. So, the next steps would be to have an updated dataset containing labeled data from the last five years or so. This will allow for a more accurate model and predictions. Another assumption is that there is only one genre available for a song. This is related to the previous statement, that with genre crossing there can be multiple correct labels for a song. This is an easy step in which the model can predict labels with varying confidence levels. There can be cutoff level, after that any predicted genres can be applied to the song.

Some limitations that were faced was the slow hardware that can't run the neural network in parallel. Having better hardware could have speed things up for training. Another issue is the small and outdated dataset that had to be used. For the model to be viable, a much larger and current dataset is required. Compared to the previous project, the model performed well but not against the state-of-the-art models. A large dataset such as Spotify's will allow for a well-trained neural network.

VIII. Conclusion

The purpose of my project was to discover alternatives to the current state-of-art models. The SVM machine performed decently on the dataset, but not enough to be considered for commercial use. PCA was an available option that resulted in poorer performance due to the complexity of the datapoints. Overall, SVMs aren't something that can be considered for future use in music genre classification. As expected, CNN performed the best with the highest test accuracy. This model can be further developed in the future using a more updated dataset. With a large and varying dataset, the CNN model can be further improved. Collecting and labeling the data while analyzing it becomes a huge cost issue. So, the help of music platforms is needed to create a more effective model.

References

[1] Raval, Meet, et al. "Music Genre Classification Using Neural Networks." *International Journal of Advanced Research in Computer Science*, vol. 12, no. 5, Sept. 2021, pp. 12–18. *EBSCOhost*, doi:10.26483/ijarcs.v12i5.6771.

[2] Pelchat N, Gelowitz CM. Neural Network Music Genre Classification. Canadian Journal of Electrical and Computer Engineering 2020;43(3):170-173.

[3] Castillo JR, Flores MJ. Web-Based Music Genre Classification for Timeline Song Visualization and Analysis. IEEE Access 2021;9:18801-18816.