AML Project Deliverable #3 - Final Report
Music Genre Classifier – Group 22

Overview:
With the ever-increasing digitalizing of music from creation to distribution and streaming platforms, the amount of music available has exploded in recent years. This has made it more challenging to manage and navigate through the vast collection of music available. To create a better and easier user experience, companies like Spotify, Apple Music, Soundcloud, and others have utilized machine learning to generate recommendation and search systems to provide users with a more personalized and relevant experience based on their preferences. These companies use music genre classifiers as part of their recommendation system.

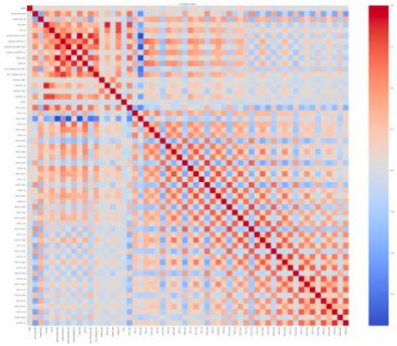
Datasets Used:
We have used two datasets in our project:

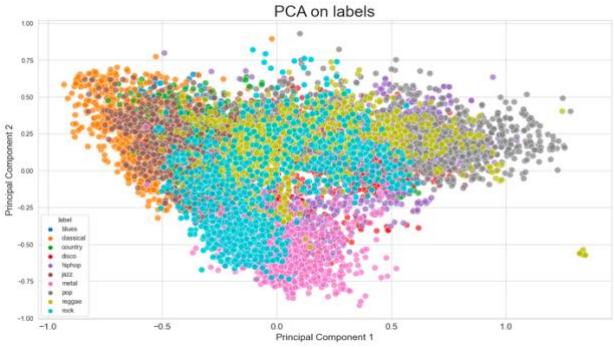
1. **GTZAN Music Classification Dataset -** A collection of around 10000 30-second-long audio files split into ten genres. The genres are - blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock.

   We have certain numerical features for 3-second segments of the audio files. In addition to this, we converted the audio files into frequency-domain representation to generate three kinds of image data that can be used to train CNNs:
   - Wave show plots
   - MFCC plots
   - Mel Spectrogram

2. **Spotify Data -** Using the Spotify API, we retrieved features for a set of 500 songs belonging to the ten genres in the GTZAN dataset. We extracted features such as danceability, energy, and tempo, commonly used in music classification. The features in this dataset are different from those in the GTZAN dataset.

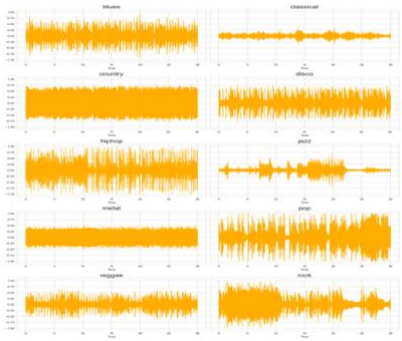
Exploratory Data Analysis:
1. **3-Second features:** This dataset contains the mean and variance of multiple features of an audio file, like spectral centroid, spectral bandwidth, roll-off, harmony, etc. These features describe different properties of a sound wave. There is a correlation between Spectral centroid and spectral bandwidth with roll-off, so we drop the roll-off column.

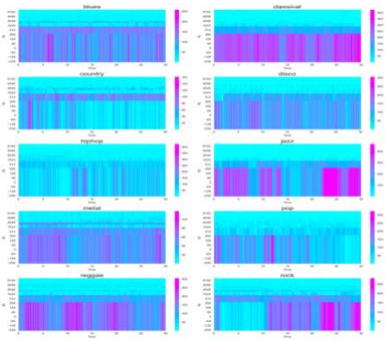


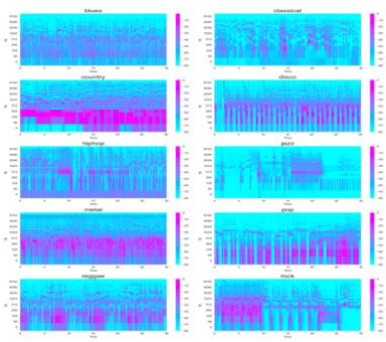


The PCA plot in 2 dimensions on the features dataset gives us some separation between the different classes. This indicates that modeling the feature data could provide us with good results. We'll explore this further in the modeling section.

2. **Image Data:** We plotted the Waveshow plots, MFCC plots, and Mel Spectrograms for the ten genres to see if there was a difference in the plots for the different classes. Comparing the first few audio files tells us that some of these plots are distinct for the class they belong to.







AML Project Deliverable #3 - Final Report
Music Genre Classifier – Group 22

3. **Spotify Data:** Some of the features and their meaning in this dataset are:
   - **Danceability**: a measure of how suitable a song is for dancing based on a combination of musical elements, including tempo, rhythm stability, beat strength, and overall regularity.
   - **Energy**: a measure of the intensity and activity of the music
   - **Key**: the key the song is in, numbered 0-11
   - **Loudness**: the overall loudness of a track in decibels
   - **Mode**: the modality (major or minor) of a track [0 represents Minor, 1 represents Major]
   - **Speechiness**: the presence of spoken words in a track, etc.

Observations & Results:

Given the different kinds of data available, we have used various models and techniques. Since this is a multiclass classification problem, we have used AUC and Accuracy as our metrics for measuring model performance. After EDA, all the datasets went through a similar preprocessing step of scaling, dropping correlated features, encoding the target variable, etc. The datasets were then split into train-val-test sets in the ratio 60-20-20. The training and validation datasets were used for model training and hyperparameter tuning. The test set was used to get the model performance.

1. **3-Second features:**

We trained four models using the original feature data. The Hist Gradient Boosting Classifier performed exceedingly well on the test dataset with a very high accuracy score. There was overfitting in the Decision Tree Classifier model, so we carried out a cost complexity pruning step which slightly increased the Accuracy.

Model	*Accuracy – Test Set*	*Optimal Hyperparameter*
Linear Discrimination Analysis (LDA)	67.91%	-
Decision Tree Classifier (Before Pruning)	67.41%	-
Decision Tree Classifier (After Pruning)	68.31%	Alpha: 0.0001181
Hist Gradient Boosting Classifier	92.64%	LR: 0.2, Max Depth: 7, Max leaf nodes: 100
XGBoost Classifier	90.44%	Eta: 0.2, Max Depth: 7, Gamma: 0.1

We calculated feature importance using the optimal Decision Tree and XGBoost models. The three most important features from this dataset based on:

- Decision Tree Classifier - perceptr_var, mfcc4_mean, and chroma_stft_mean
- XGBoost Classifier - perceptr_var, spectral_bandwidth_mean, and mfcc4_mean

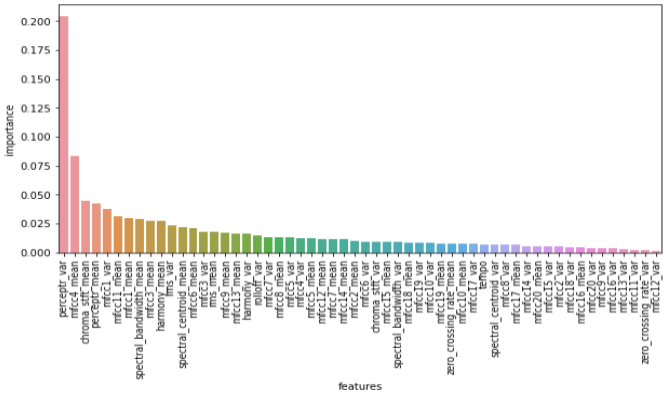


Figure 1 - Decision Tree Feature Importance

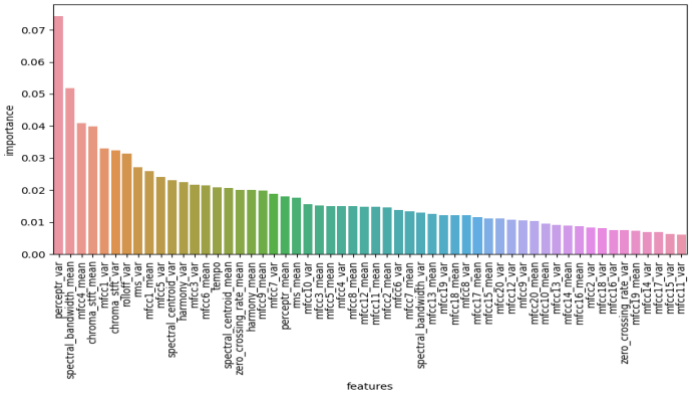


Figure 2 - XGBoost Feature Importance

2. **Image Data:**

For the image data, we scaled down the pixel values and resized the images based on the input shape of the models that we used (Ex: Inception V3 takes in a 256x256x3 image as input). Given that we had three kinds of Image data, we trained a CNN with InceptionV3 with ImageNet weights as the base model for each of the three image datasets to identify which representation of the audio file captures most of the information of the audio file and is suitable for classification tasks.

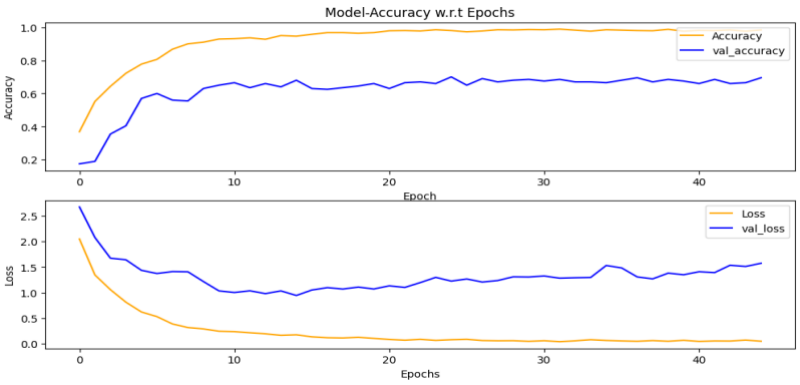
AML Project Deliverable #3 - Final Report
Music Genre Classifier – Group 22

```
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, 6, 6, 2048)        21802784

flatten (Flatten)            (None, 73728)             0

batch_normalization_94 (Bat  (None, 73728)             294912
chNormalization)

dense (Dense)                (None, 512)               37749248

dropout (Dropout)            (None, 512)               0

batch_normalization_95 (Bat  (None, 512)               2048
chNormalization)

dense_1 (Dense)              (None, 256)               131328

dropout_1 (Dropout)          (None, 256)               0

dense_2 (Dense)              (None, 10)                2570

=================================================================
...
Trainable params: 38,031,626
Non-trainable params: 21,951,264
_________________________________________________________________
```

Image Data	*Train Acc*	*Val Acc*	*Test Acc*
Wave show Plot	99.17%	48.50%	44.5%
MFCC Plot	96.66%	52.00%	41.5%
Mel Spectrogram	96.16%	60.00%	62.00%

Based on the model performance on the Test data, Mel Spectrogram is the representation of audio files that contains the most information and thus is a good dataset for genre classification if we are using image data.
After identifying Mel Spectrograms as the Image data to use, we replaced the InceptionV3 base model with a few different models, including creating a CNN from scratch to compare performance and identify the optimal model. VGG16 with ImageNet weights performed the best.

Model	*Val Acc*	*Acc (Complete Dataset)*
InceptionV3	69.00%	93.69%
VGG16	**70.00%**	**93.89%**
ResNet50	65.50%	90.49%



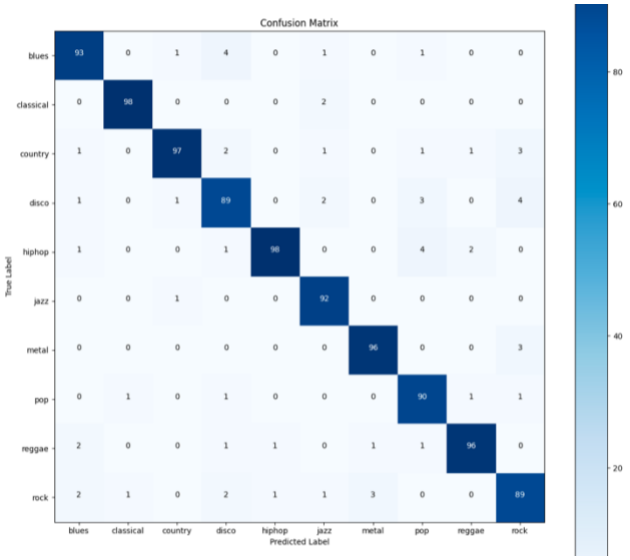
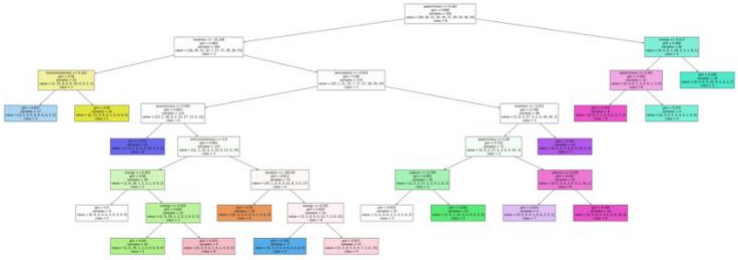


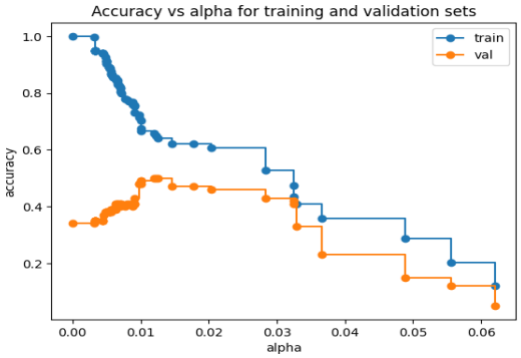
Figure 3 – Accuracy vs Epoch plot and Confusion Matrix for VGG16 model

3. **Spotify Data:**
Like the 3-second feature data, we trained multiple models after carrying out the standard preprocessing steps. The results from the model are as follows:

Model	*Val Acc*	*Test Acc*
Logistic Regression	56.00%	59.00%
Support Vector Machines	55.00%	50.00%
Decision Tree Classifier	50.00%	50.00%
Random Forest Classifier	**59.00%**	**61.00%**
XGBoost Classifier	54.00%	55.00%



We carried out cost complexity pruning for the Decision Tree Classifier to find the optimal alpha. The Accuracy vs. alpha plot gives us the optimal alpha for this model.



The three most important features from this dataset based on:

- Decision Tree Classifier – Speechiness, Loudness, Instrumentalness
- Random Forest Classifier – Speechiness, Danceability, Acousticness
- XGBoost Classifier – Loudness, Key_1, Speechiness

AML Project Deliverable #3 - Final Report
Music Genre Classifier – Group 22

Conclusion:

Based on our project, using a Hist Gradient Boosting model on the 3-second features of the audio files gives the best result with an accuracy of 92.64%. Training a model using the 3-second features related to the audio signals like MFCC and Spectral Bandwidth instead of the Spotify data, which is related to music features like Acousticness and Speechiness, gives us better results. However, the generalizability of this result might require further research since the audio files in the Spotify dataset are different from the ones in the GTZAN dataset.

Image representations of the audio files, specifically the Mel Spectrogram representation, also gives us a classifier with high Accuracy of 69% on the test data. This increases further when we replace the base pre-trained InceptionV3 model with a pre-trained VGG16 model.

Given the variation between songs, even in the same genre, and no clear definition of what a genre constitutes, classifying music into genres is a complex and dynamic problem. Our project partly explores and explains the complexity of categorizing music into genres.