**Music Genre Classification using Neural Networks**

**Team 7 Members:**

**i) Anastasios Kleniatis  
ii) Gloria Muthumbi  
iii) Lee Sangyoon  
iv) Park Chanho**

# **Problem Description**

Music Classification and Selection has to be one of the most cumbersome problems to solve. To manually categorize music, one must listen to all the songs, and only then can they choose the genre it belongs to. The objective of the project is to make selection of songs easier and quicker. This is done by identifying trends in the audio signals and other significant features and classifying them in the different classes.

# **Dataset**

For our study, we used the GTZAN Dataset from Kaggle. The dataset comprises of 1000 audio recordings that are each 30 seconds long. They represent 10 genres of music each containing 100 audios. The genres include **Classical, Country, Blues, Hip-hop, Jazz, Pop, Reggae, Metal, Blues and Rock**. It contains:

* A “Genre Original” folder that holds a collection of the audios in their respective genres.
* 2 CSV files for features of the audio files. 1 file is for 30 second audios and the other one is for 3 second audio files.
* A “Images Original” folder that contains spectrograms for each of the audio files.

**Machine Learning Models**

* **Simply Using Basic ML Classifier & Neural Network**

XGBoost is gradient boosting algorithm with regularization and making faster result. XGBoost makes good result and also it is easy to find importance of features.

Neural Network is machine learning algorithm with fully connected(dense) layers.

* **Train a CNN using the spectrograms**

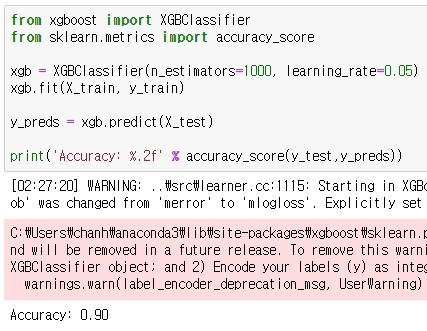
Convolutional neural networks are a specialized type of artificial neural networks that use a mathematical operation called convolution in place of general matrix multiplication in at least one of their layers. They can successfully capture Spatial dependencies using relevant filters and they can perform better fitting to the image dataset due to the reduced number of parameters. In this way they can better understand the sophistication of an image than a traditional NN. A very important feature of a signal is its spectrogram. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. From the spectrogram we can see how energy levels vary over time for songs of different genres.

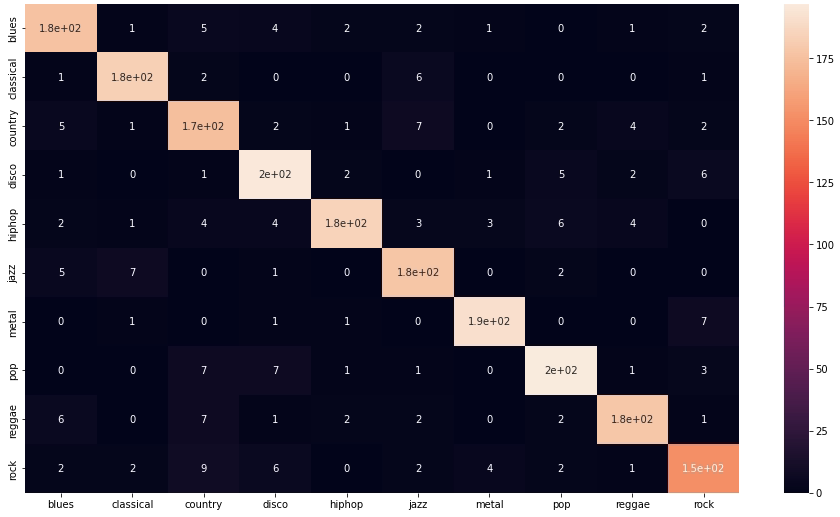
* **LSTM with Audio File as Sequential Data**LSTM, Long Short Term Memory Network is an enhanced RecurrentNeural Network Model that allows endurance of information. It resolves the vanishing gradient issue found in RNNs. RNNs are able to retain the prior knowledge and apply it to the incoming data for processing. However, they are unable to recall long-term dependencies hence are used for the sole purpose of preventing long-term dependence.

## **Results of the Model**

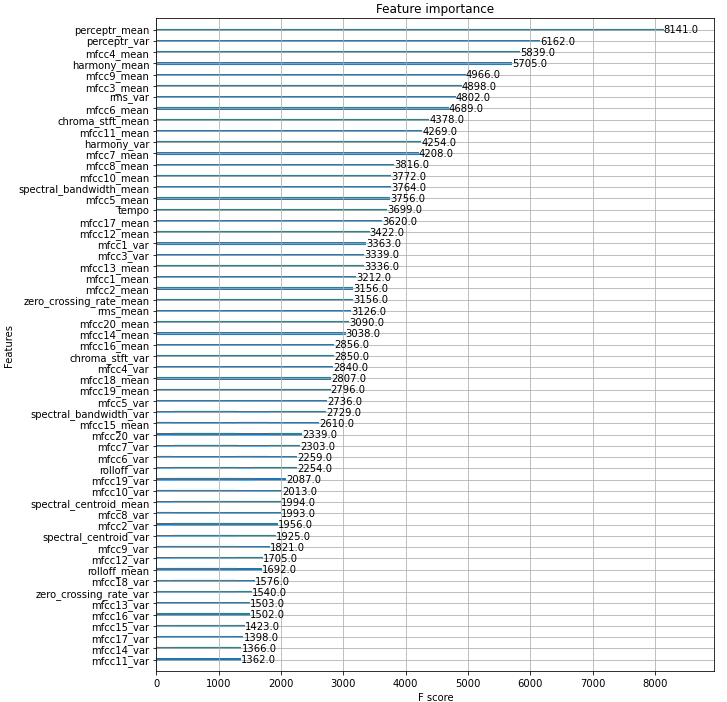
Simple ML Algorithm

Used XGBoost for ML Classification (<https://drive.google.com/file/d/1_nGX9MX14OF28pJqpUL9Qq6zU75YZJPc/view?usp=sharing> )

Code & Accuracy

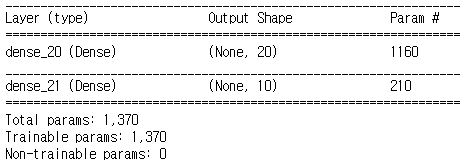
Confusion Matrix

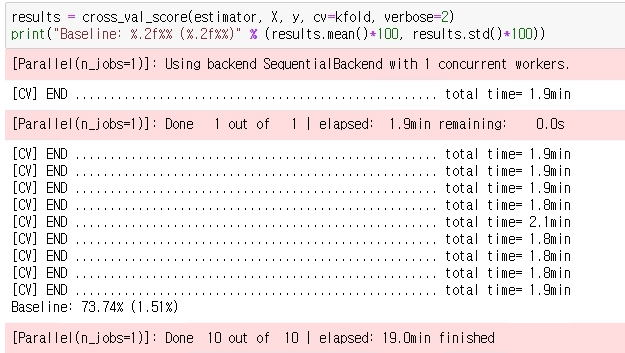
Feature Importance



Simple Neural Network

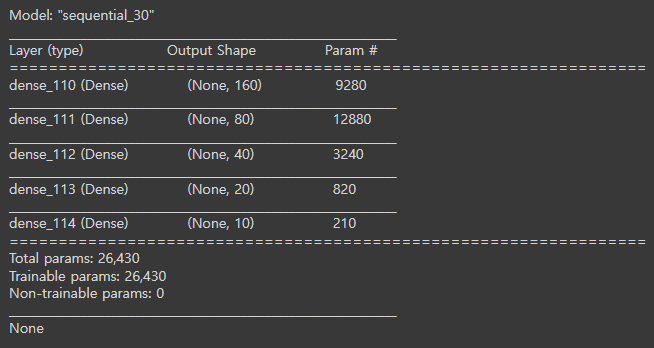
The model consists of 1 hidden layer, and relu for activation. For classification, softmax is used as activation function.

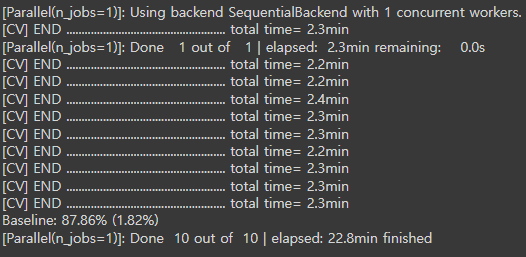
Model 

Accuracy : 73.74%

Deeper Neural Network

The model consists of 4 hidden layers, and relu for activation. For classification, softmax is used as activation function.

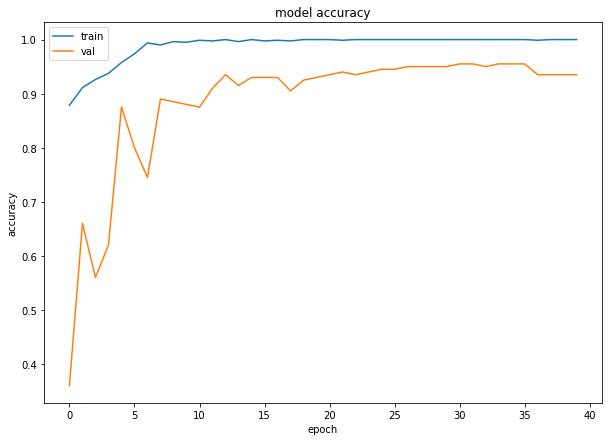
Model

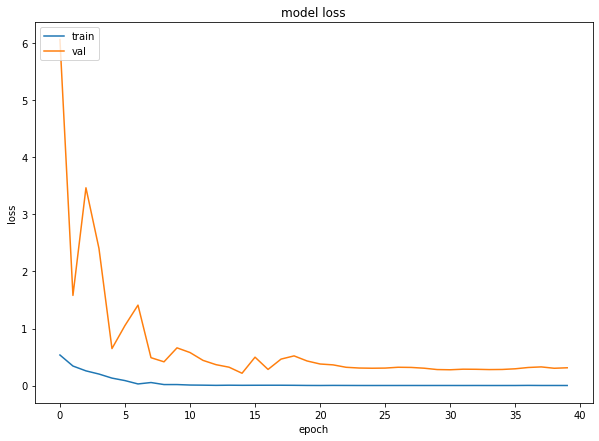
Accuracy

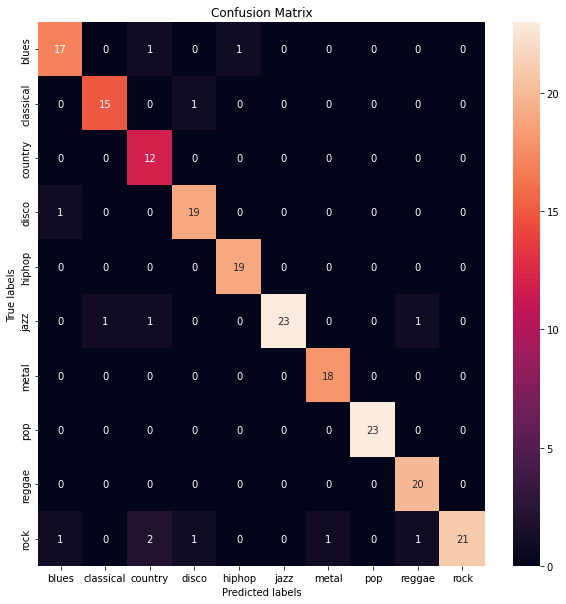
CNN Model

The model consists of convolutional, batch normalization, pooling, dropout and dense layers. The batch normalization layers are used to accelerate the training process by normalizing the data. The pooling layers decrease the spatial resolution which lowers the complexity and can help in processing larger parts of the image. The dropout is used to avoid overfitting by randomly disabling some neurons. Then the extracted feature matrix passes through the flatten layer and becomes a one dimensional vector which in turn passes through a dense layer with a softmax activation function. The output is a vector of the probabilities of each class. The loss function that is used is Sparse Cross Entropy and the metric is the evaluation accuracy.

In order to gain a better picture of the effectiveness of the CNN in this classification, K-Fold Cross Validation was used. The dataset was split in 5 folds, and the CNN used a different fold as the validation data in each iteration. But due to late implementation only 3 of the 5 iterations were executed. The average accuracy of these 3 iterations is 97.5% which is surprisingly good (but raises concern about implementation errors).







Code Link:<https://colab.research.google.com/drive/1czYoDIcaP4DwIep4W_6Nwgkxmd63snF5?usp=sharing>

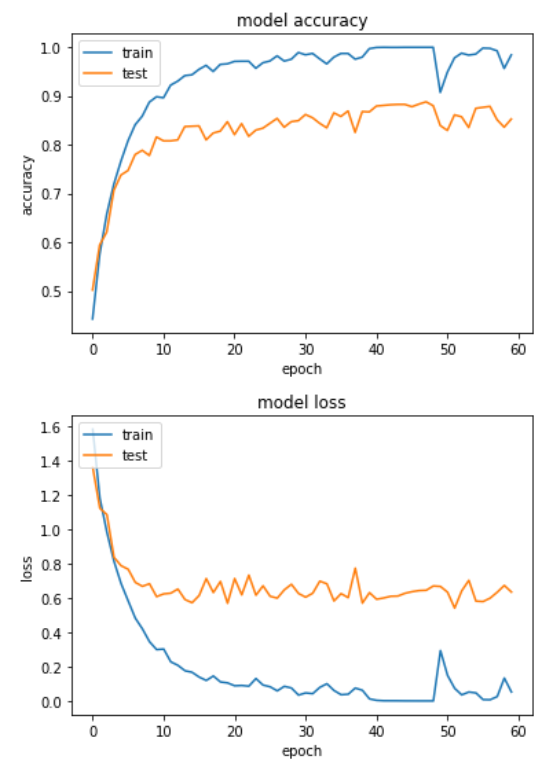
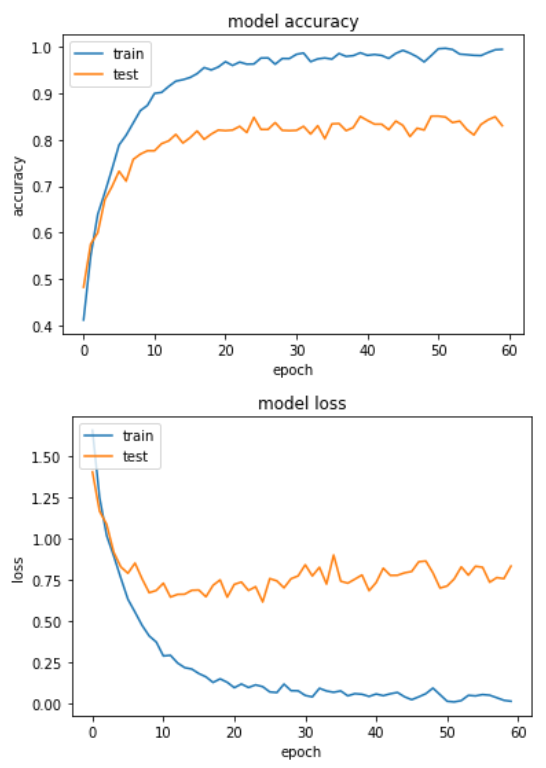
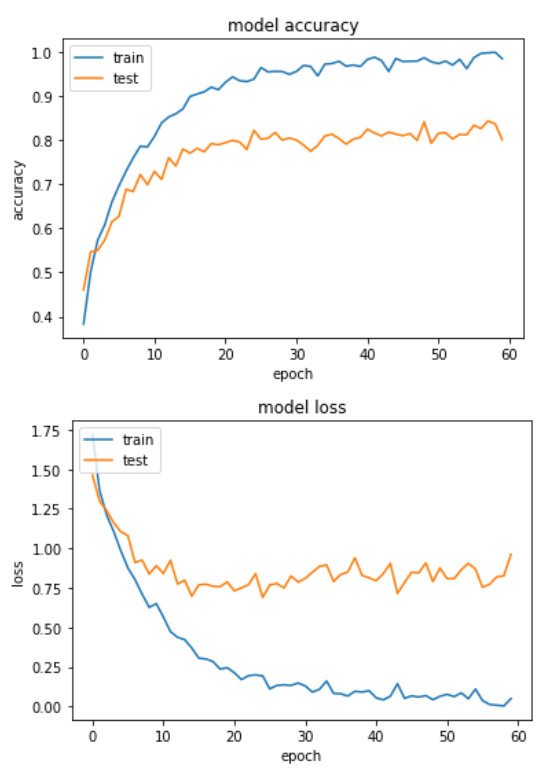
Sources:

<https://towardsdatascience.com/music-genre-recognition-using-convolutional-neural-networks-cnn-part-1-212c6b93da76>

<https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-k-fold-cross-validation-with-keras.md>

LSTM Model

The accuracy that was achieved was about 86%.



Result comparison

Given Results from <https://www.kaggle.com/code/andradaolteanu/work-w-audio-data-visualise-classify-recommend>

| Model | Accuracy |
| --- | --- |
| Naive Bayes (given) | 0.51952 |
| Stochastic Gradient Descent (given) | 0.65532 |
| KNN (given) | 0.80581 |
| Decision Trees (given) | 0.63997 |
| Random Forest (given) | 0.81415 |
| Support Vector Machine (given) | 0.75409 |
| Logistic Regression (given) | 0.6977 |
| Nerual Nets (given) | 0.67734 |
| Cross Gradient Booster (Random Forest) (given) | 0.74875 |
| XGBoost (given) | 0.90224 |
| XGBoost | 0.90 |
| Simple Neural Network | 0.7374 |
| Deeper Neural Network | 0.8786 |
| CNN | 0.975 |
| LSTM | 0.86 |