MUSIC GENRE CLASSIFICATION OF AUDIO SIGNALS

NAMAN GUPTA

B tech 3rd year , Jaypee Institute Of Information Technology, Noida

VAIBHAV JAIN

B tech 3rd year , Jaypee Institute Of Information Technology, Noida

ADITYA AGGARWAL

B tech 3rd year , Jaypee Institute Of Information Technology, Noida

SIDDHANT PATHAK

B tech 3rd year , Jaypee Institute Of Information Technology, Noida

SUPERVISED BY:

DR. CHETNA DABAS

ABSTRACT:

The topic of music genre classification has been studied previously using various classification algorithms. It is a popular problem in the domain of Music Information Retrieval (MIR), which has been used in many music streaming platforms, such as Pandora. In this paper, we have done a comparative study using various machine learning algorithms to classify the music into its various genres namely, blues, classical, country, disco, hip-hop, jazz, metal and pop respectively. We have used various audio features, such as Mel Frequency Cepstral Coefficients (MFCC), Delta, Delta- Delta and temporal features, including beats and tempo to featurize our data. Various classification algorithms, such as Support Vector Machine(SVM), Decision Tree, k-Nearest Neighbors(KNN), Random Forest and Gradient Boosting Classifier are used in the classification of the data.

INTRODUCTION:

Musical genres are created and used by humans to describe and categorize various categories of music. They do not possess any strict definition and boundaries as they are created through cultural and marketing factors by the common public. This has led to the curiosity of some researchers to systematical study of various genres of music. However, even with current musical genres, it is evident that the members of a particular genre may share certain characteristics typically related to instrumentation, rhythmic structure and pitch content of the music. Automatic music analysis is one of the service that the music content providers uses to attract people to buy their product. Examples include Saregama Carvaan. We have provided a framework for the development and evaluation of feature which describe musical content. We have proposed various audio analysis techniques for the music. The paper compares the accuracies of various classifiers which we have used for the classification and suggests the best method for classifying the dataset. We have used the technique such as Fast Fourier Transform(FFT) and Mel Frequency Cepstral Coefficients (MFCC). The techniques we have suggested can be used by any application to give utility to a user to listen to a song based of his choice. We believe that there is no prior work done to analyse the algorithms using features such as Delta, Delta-Delta and tempo.

MUSIC DATASET:

The dataset we have used for our music genre classification is GTZAN[]. This dataset contains 1000 song files, each of which is 30 seconds long. These songs are classified into 8 genres, namely, blues, classical, country, disco, hip-hop, jazz, metal and pop respectively. The sampling rate we have used for our data files is 22050 Hz. All these files were in .au format

which were converted to .wav using online converter. We divided our dataset into training and testing data in the ratio 7.5 : 2.5. Spectogram of songs of different genres are depicted.

DATA PROCESSING AND NORMALISATION:

To convert our audio data, for processing we have used 5 different algorithms, namely, Support Vector Machine(SVM), Decision Tree, k-Nearest Neighbors(KNN), Random Forest and Gradient Boosting Classifier respectively. Librosa toolkit has been used to implement these algorithms on our dataset.

i. FAST FOURIER TRANSFORM:

A FFT is a mathematical method to obtain DFT(Discrete Fourier Transform) for a sequence or the inverse of a sequence. A Fourier analysis is performed to obtain a frequency domain representation of the original domain. Rapid computation of this transform by the factorization of Discrete Fourier Transform Matrix into a sparse factors' product is job done by an FFT. Because of which, the complexity of obtaining a DFT is reduced to O(n log n) from O(n^2) to , where n represents the data size. Let x0, ..., xN-1 denote complex numbers. The DFT is obtained by the formula X_k = ∑n to n=1 xe-i2πkn/N, where k = 0,1,….., N-1.

ii. Mel Frequency Cepstral Coefficient

Mel Frequency Cepstral coefficients emphasize on obtaining the exact structure of the audio signal to extract linguistic features and discard the background noise. The linear cosine transform of a logarithmic power spectrum on a Mel scale which is non-linear is the basis of its calculation. A collection of Mel Frequency Cepstral Coefficients form a Mel frequency cepstrum. The Mel scale is calculated as M(f) = 1125 ln(1+ f/700).

The following computations are made to obtain the Discrete Fourier transform of the frame - Si(k) = ∑n to n=1 s(n)h(n)e-j2πkn/N where 1≤ k ≤ K h(n) – is the Analysis window for N samples K- Length of the Discrete Fourier Transform The power spectral estimate based on the Periodogram for si(n) , which is the speech frame is specified by Pi(k) = |Si(k)|2 This periodogram is further processed to obtain 26 cepstral co-efficients, DCT is applied on 26 log filter bank energies which is called the MFCC.

iii. Delta

The extraction of the cepstrum via the Inverse DFT from the previous section results in 12 cepstral coefficients for each frame. We next add a fourteenth feature: the energy from the frame. Energy correlates with phone identity and so is a useful cue for phone energy detection (vowels and sibilants have more energy than stops, etc). The energy in a frame is

the sum over time of the power of the samples in the frame; thus for a signal x in a window from time sample t1 to time sample t2, the energy is:

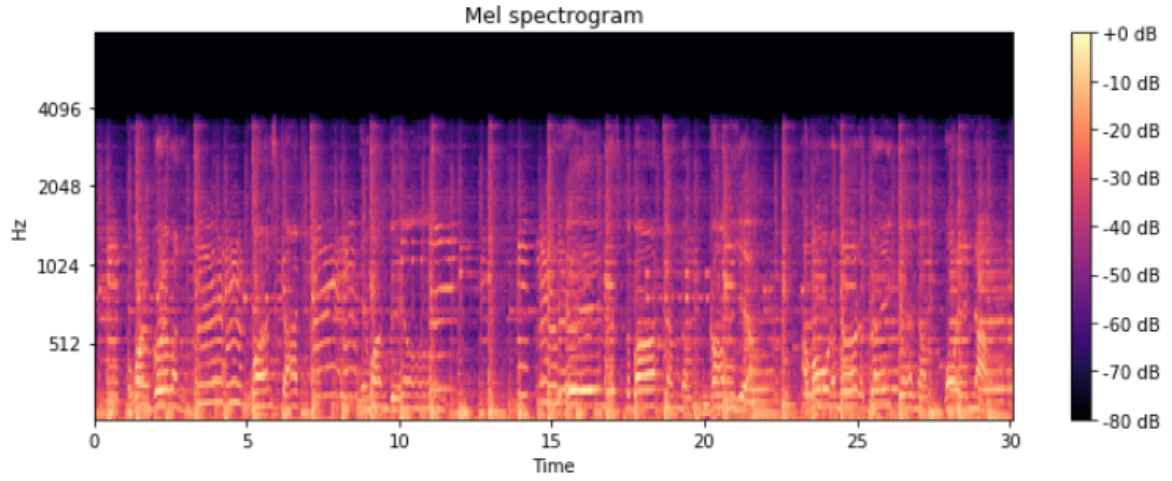
Energy = $\sum_{t=t1 \text{ to } t2} x^2[t]$

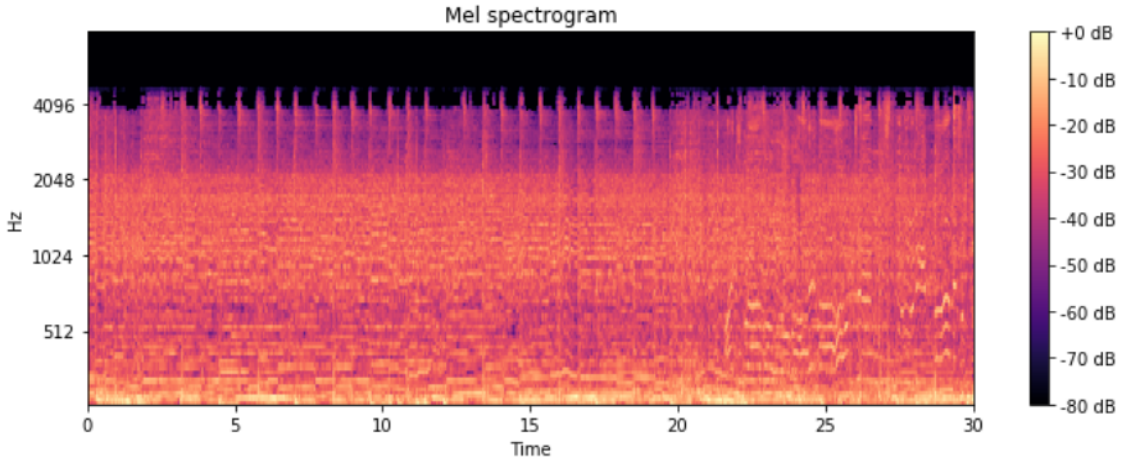
iv. Delta-Delta

Another important fact about the speech signal is that it is not constant from frame to frame. This change, such as the slope of a formant at its transitions, or the nature of the change from a stop closure to stop burst, can provide a useful cue for phone identity. For this reason we also add features related to the change in cepstral features over time. We do this by adding for each of the 13 features a delta or velocity feature, and a double delta or acceleration feature. Each of the 13 delta features represents the change between frames in the corresponding cepstral/energy feature, while each of the 13 double delta features represents the change between frames in the corresponding delta features. A simple way to compute deltas would be just to compute the difference between frames; thus the delta value d(t) for a particular cepstral value c(t) at time t can be estimated as:

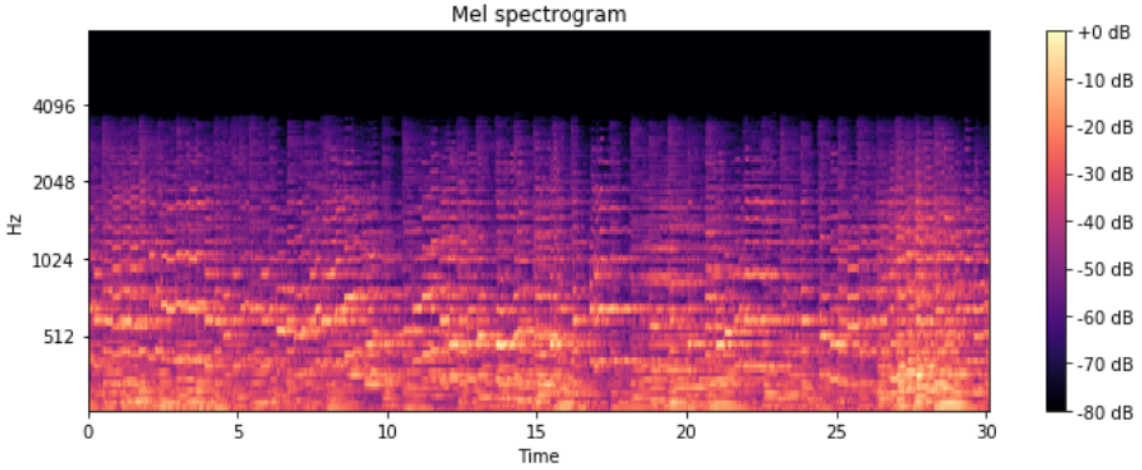
$$d(t) = [c(t+1) - c(t-1)] / 2$$

Instead of this simple estimate, however, it is more common to make more sophisticated estimates of the slope, using a wider context of frames.

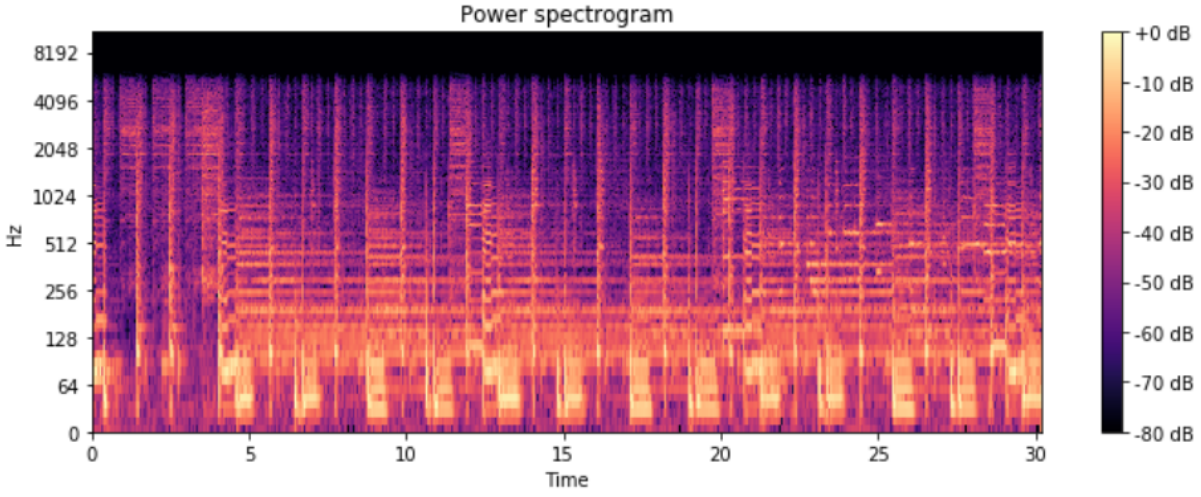




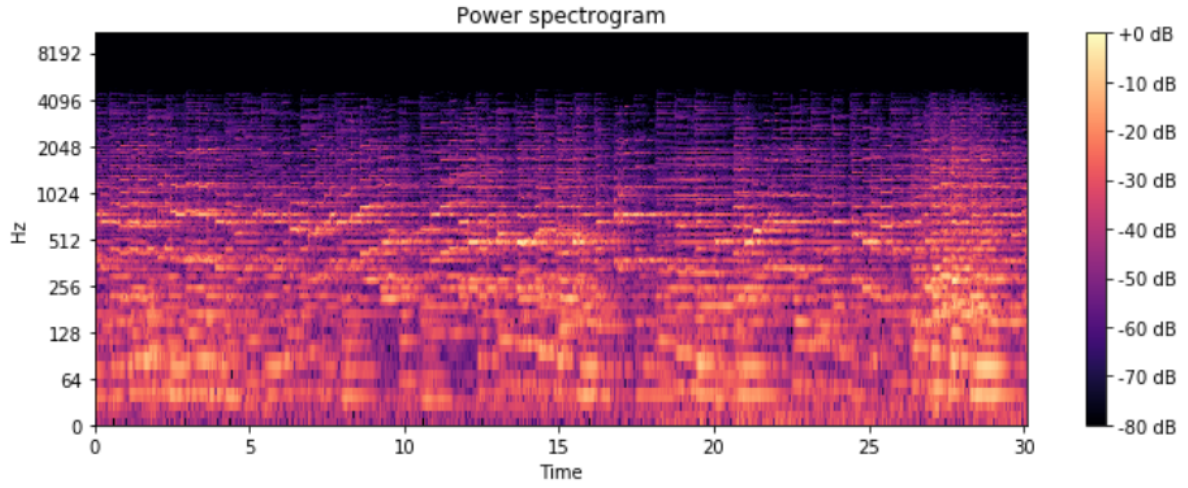
Power Spectrogram of Disco Genre files



Mel Spectrogram of Classical Genre files



Power Spectrogram Of Disco Genre files



Power Spectrogram Of Classical Genre files

METHODOLOGY:

The architectural structure of our dataset is pre-processed. The pre-processed data is used for the training of each of our classifiers. The classified data is then tested using the test data.

Algorithm

1. Convert data from .au format to .wav format.
2. Preprocess the data using FFT.
3. Feature Extraction using MFCC, Delta, Delta- Delta and rhythmical features.
4. Feature Reduction using Principal Component Analysis.
5. Optimisation of hyper-parameters using Grid Search with Cross-Validation.
6. Training the classifier using various classification algorithms.
7. Testing the data and predicting the genre of our data files.
8. Compare performances of different classifiers using different benchmarks.

FEATURE REDUCTION (PCA)

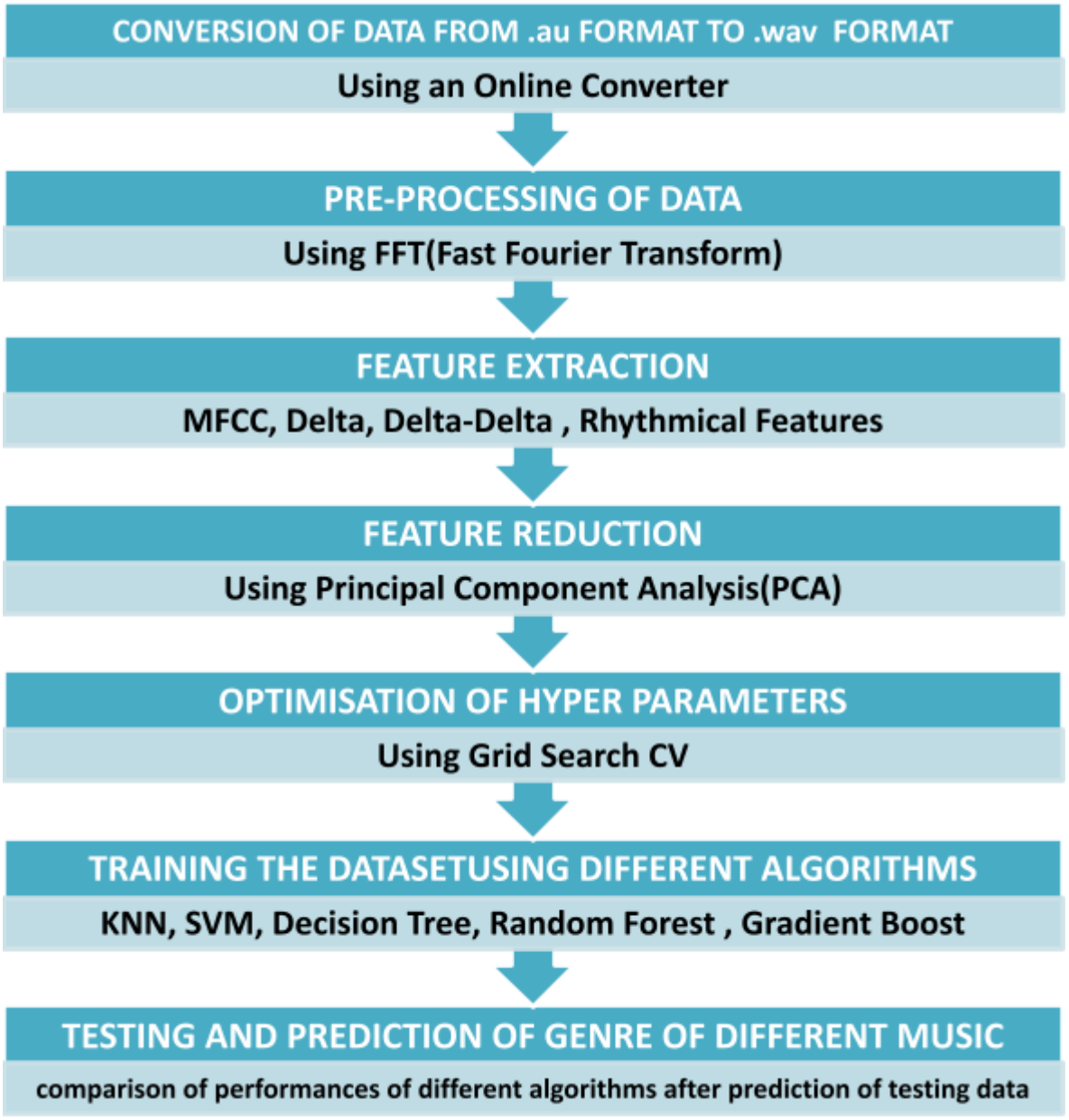
Principal Component Analysis (PCA) is the most important technique to visualize the data of dataset. It is a statistical procedure that is entirely based on orthogonal transformation ,

which converts some corelated values in the data set into set of linealry uncorelated data set.

We have used PCA in our work so as to carry out the dimension reductibility. We have reduced from 40 features (13 mfcc , 13 delta ,13 delta-delta , 1 tempo) to 7 important features by applying explained variance to our data set.

OPTIMISATION (GRID SEARCH)

The most effective way of hyperparameter optimisation (c, gamma, kernel in case of SVM) is by the use of grid search. It basically does an exhaustive searching through a specified subset of hyperparameter. For our dataset we got the best results for '**C**': 10000, '**Gamma**': 0.0005, '**Kernel**': 'rbf'. There are many metrics for its evaluation but the most effective is cross validation , so we used it.



CLASSIFICATION ALGORITHMS

i. k-Nearest Neighbors

K nearest neighbors is an algorithm that stores all the currently classified new cases based on pervious available cases. It belongs to supervised learning domain and is non-parametric. It does not make any underlying assumptions about the distribution of data. A majority vote of the case's neighbors classify it and each of the case being assigned to the class most common among its k-nearest neighbors is computed by a distance function. We have used Euclidean distance for computing our distance function. Generally, a higher k value reduces the overall noise. In our classification, k = 17 turned out to be the best k value.

ii. Support Vector Machine(SVM)

SVM training algorithm builds a model which is a non-probablistic binary linear classifier used for classification. SVM finds a hyperplane which maximises the decision boundary between 2 classes. The vectors which define a hyperplane are known as support vectors. The data is mapped to a higher dimensional space after defining hyperplane. We have used the Radial Basis Function(RBF) kernel, which gives best results in case of non-linearly separable data points. This kernel in our case gives the best results.

iii. Decision Tree

A decision tree constructs a tree structure based on regression or classification models. Dataset is divided into smaller subsets and simultaneously a tree is built. A node of the tree has 2 branches. A leaf node represents a decision or a classification. The root node corresponds to a predictor. The algorithm used for building decision trees is CART(Classification and Regression Tree). CART adopts a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. CART uses gini index as the splitting attribute for the tree. The gini index measures the impurity in a data partition. It considers the best binary split for each attribute. The point giving the minimum gini index for an attribute is taken as the split point of that attribute.

$$Gini(D) = 1 - \sum_{i=1 \text{ to } m}(pi^2)$$

The reduction in impurity that would be incurred by a binary split on an attribute A is –

$$\Delta Gini(A) = Gini(D) – Gini_A(D)$$

iv. Random Forest

Random forest algorithm is a supervised classification algorithm which is an ensemble of decision trees. There are more trees (dataset is divided into smaller partitions) in a forest which makes this algorithm more robust. It uses divide-and-conquer approach to improve the performance. The decision trees are weak learners whereas random forest is a strong learner. When a new input is entered, it is run down in all of the trees. The result may be average, or weighted average of the terminal nodes which are reached. Random forest is able to deal with unbalanced and missing data as well.

v. Gradient Boosting Classifier

A boosting algorithm which is used for reducing bias and variance in supervised learning. It is a weighted algorithm where each input class is given a weight and then is evaluated based on predicted output. The important data is given more weightage then low predicting data. This algorithm boosts the basic bagging algorithm by taking strong predictor model. The learning rate must be kept in range [0.00001 , 0.1] so that it takes precised inputs.

RESULTS:

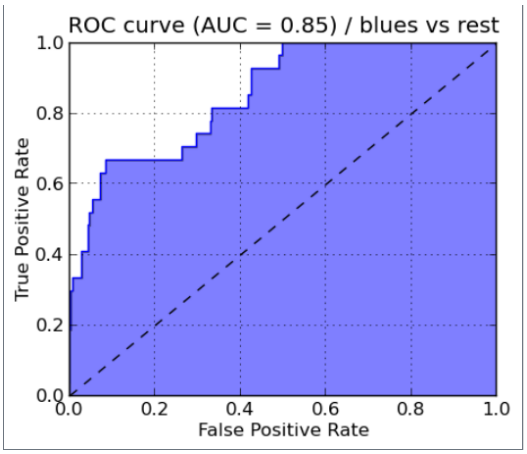
We tested the dataset on different classifiers. Different classifiers gave different accuracies. The various parameters we have used to evaluate the result of our testing data are accuracy, precision, F1 score and recall. We have tried to interpret the outcomes with the help of confusion matrices.

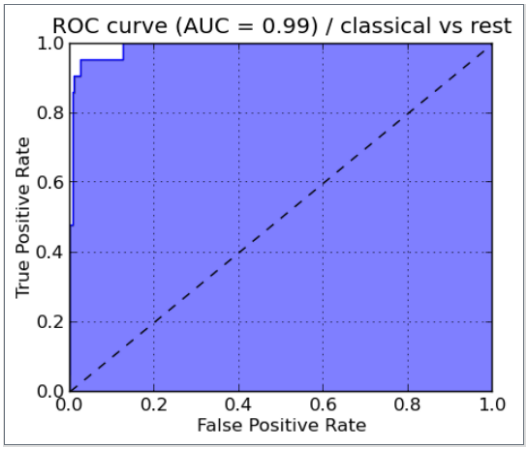
ier	on		›re	›rt
m Forest Classifier hbors lassifier ›n Tree ›st nt Boost /ith Grid Search CV				

The Gradient Boosting Classifier gave us the highest precision but after applying grid search optimisation SVM gave the highest precision value. The highest value of recall also was of gradient boost but after grid search SVM gave us the best recall value of 0.53.

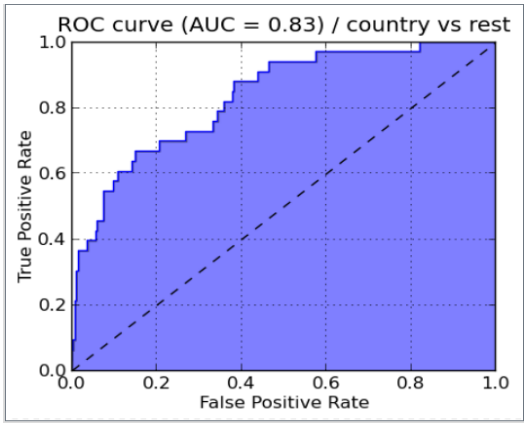
Furthermore, since F1 score is the harmonic mean of precision and recall so f1 score also followed the same pattern i.e before grid search gradient boost and after grid search SVM gave best results.

ONE VS ALL COMPARISION FOR ALL GENRES :

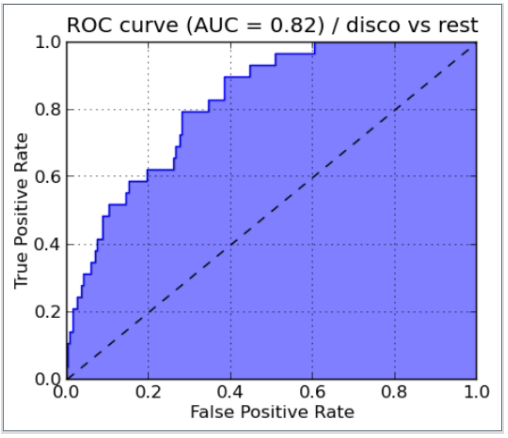




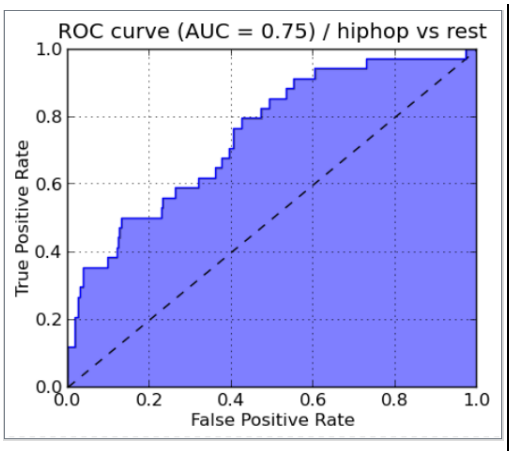
BLUES CLASSICAL



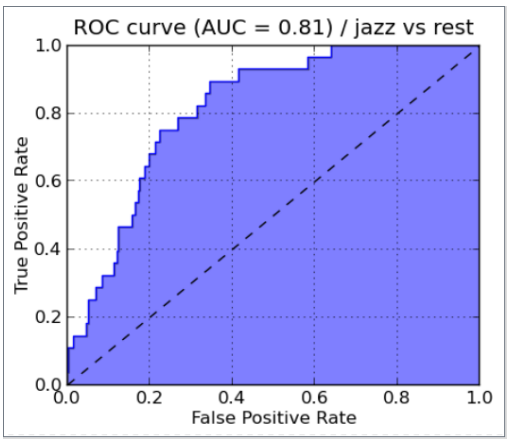
COUNTRY



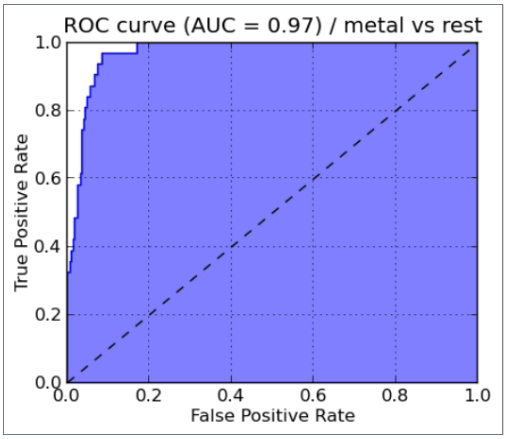
DISCO



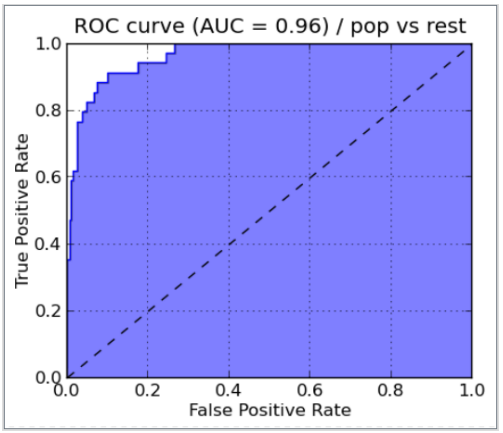
HIP HOP



JAZZ



METAL



POP

Our algorithm works best for classical genre class as compared to other classes.

REFERENCES:

1. G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302, 2002.

2. Khunarsal,P.;Lursinsap,C.;Raicharoen,T.:Veryshorttimeenviron- mental sound classi cation based on spectrogram pattern matching. 2013, (in press). [Online]. Available: http://www.sciencedirect.com/ science/article/pii/S0020025513003113

3. Long Short Term Memory RNN Wikipedia: https://en.wikipedia.org/wiki/Long_short-term_memory.

4. Spicy: Discrete Fourier Transform Pack: http://docs.spicy.org/doc/spicy/refernce/fftpack.html

5. Pradeep Kumar D, Sowmya B J, Chetan, K G Srinivasa , A Comparative Study of Classifiers for Music Genre Classification based on Feature Extractors.

6. http://marsyas.info/downloads/datasets.html