Music Genre Classification

by

Abdullah SAHIN

INTRODUCTION

Many facts make intelligent systems of automatic music genre classification (AMGC) vital these days. The ease of listening to music on devices, the high availability of albums on the Internet, peer-to-peer servers and the fact that artists now actively distribute their songs on their websites make music database management a must. In addition, searching genres and generation of smart playlists to select specific tunes between gigabytes of songs on personal portable audio players are essential tasks that facilitate music mining.

On the other hand, the classification of music genres is as mentioned above, an ambiguous and subjective activity. It is also a field of research that is being challenged, either because of low classification accuracy or because some say that one is not capable of classifying genres that do not even have clear definitions.

However, end users are already accustomed to browsing both physical and online music collections by genre, and this strategy is proven to be at least relatively successful. In particular, a recent survey [1] found, for example, that end-users are more likely to browse and search by genre than by recommendation, artistic similarity or music similarity, although these alternatives were also common.

Innovative companies such as Spotify and Shazam have been able to leverage music data in a clever way to provide amazing services to users. An automatic genre classification algorithm could play important role on efficiency for music database and help music recommender systems and playlist generators that companies like Spotify and Pandora.

In this work, we implemented a variety of classification algorithms admitting two different types of feature: time variant and time invariant. We deliberately separate extraction process of features in two ways because we want to compare traditional algorithms (Random Forest, Support Vector Machines, Logistic Regression etc.) with recurrent neural network.

DATASET

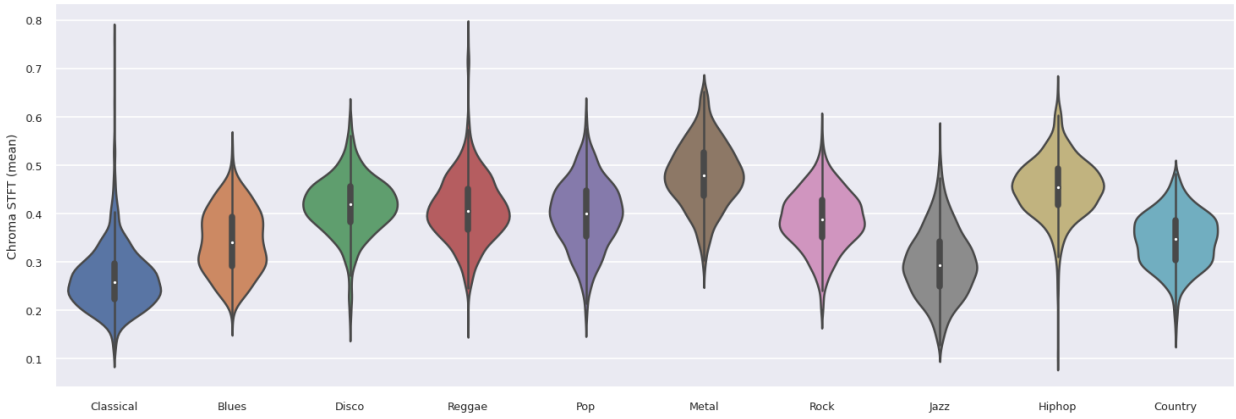
The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. The files were collected in 2000-2001 from a variety of sources including personal CDs, radio, microphone recordings, in order to represent a variety of recording conditions. From each clip, we sampled a contiguous 3-second window at 10 equally distanced locations, thus augmenting our data to 10000 clips of three seconds each. Since this data was sampled at 22050HZ, this leaves us with 66150 features for the raw audio input. Thus, after pre-processing our input is of shape (10000, 66150), where each feature denotes the amplitude at a certain timestep out of the 66150. The dataset is available on http://marsyas.info/downloads/datasets.html

EXPLORATORY DATA ANALYSIS

There are numerous characteristics of a sound wave that might be helpful to distinguish genre of music from one another. For the sake of simplicity, we will not go through all the characteristics here.

Chroma Features

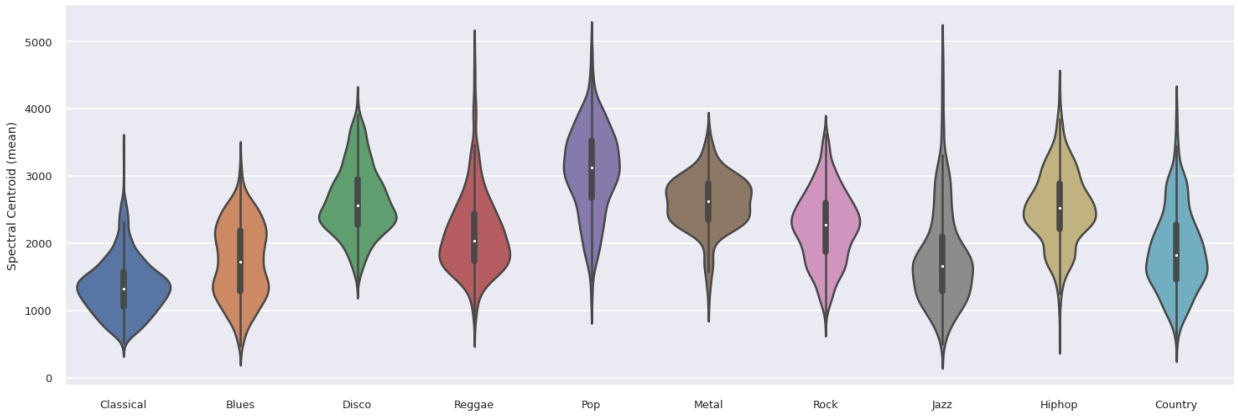
Chroma frequency vector discretizes the spectrum into chromatic keys and represents the presence of each key. We take the histogram of present notes on a 12-note scale as a 12-length feature vector. The chroma frequency have a music theory interpretation. The histogram over the 12-note scale actually is sufficient to describe the chord played in that window. It provides a robust way to describe a similarity measure between music pieces



The figure above shows the mean of chroma feature of short time Fourier transformation which is the intensity of musical notes over time. So, we can see there is definitely a difference across our Genres when comparing the Chroma mean. Metal has the highest while classical has the lowest. Reggae and classical music have a few interesting outliers that go far up as well. We evaluated outliers in the jupyter notebook.

Spectral Centroid

The spectral centroid is commonly associated with the measure of the brightness of a sound. This measure is obtained by evaluating the “center of gravity” using the Fourier transform’s frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes. Consider two songs, one from blues and one from metal. A blues song is generally consistent throughout its length while a metal song usually has more frequencies accumulated towards the end part. So spectral centroid for blues song will lie somewhere near the middle of its spectrum while that for a metal song would usually be towards its end.



Per definition of spectral centroid above, it is expected that classical music genre has the minimum spectral centroid mean value while pop genre has the highest. This time, we would like to evaluate outlier of hip-hop and jazz music.

FEATURE ENGINEERING

All the feature we used in modeling can be extracted librosa Python library. We used two different sets of features: time invariant and time variant. The former feature set is computed on windows of 2048 samples spaced by hops of 512 samples (except zero-crossing rate). Seven statistics were then computed over all windows: the mean, standard deviation, skew, kurtosis, median, minimum and maximum. For the time variant feature, we keep time varying component instead of aggregating per summary statistics mentioned above. The table below describes dimensionality of feature sets for time variant and invariant cases.

Features	**Time Variant**	**Time Invariant**
Chroma (STFT, Cens, CQT)	(130, 36)	252
Tonnetz	(130, 6)	42
MFCC	(130, 13)	91
Spec. Centroid	(130, 1)	7
Spec. Bandwidth	(130, 1)	7
Spec. Contrast	(130, 7)	49
Spec. Rolloff	(130, 1)	7
RMSE	(130, 1)	7
Zero-crossing Rate	(130, 1)	7

The first component in time variant column, 130, represents the temporal part of the feature which is calculated by:

$$\frac{sampling\ rate}{hop\ length} * time\ window = \frac{22050}{512} * 3 \approx 130$$

MODELING

Once the feature vectors are obtained, we train different classifiers on the training set of feature vectors. Following are the different classifiers that were used:

- Logistic Regression (Baseline Model)
- Random Forest Classifier
- RBF Kernel Support Vector Machine
- Recurrent Neural Network with LSTM Layer

Logistic Regression

Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership. For this multi-class classification task, the LR is implemented as a one-vs-rest method. That is, 10 separate binary classifiers are trained. During test time, the class with the highest probability from among the 10 classifiers is chosen as the predicted class.

Random Forest Classifier

Random forests are an ensemble learning method for classification. They are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree. They operate by constructing a lot of decision trees during training time and outputting the class that is the mode of classes output by individual trees. The accuracy of a random forest depends on the strength of the individual tree classifiers and a measure of dependence between them. Given an ensemble of classifiers and with the training set drawn at random from the distribution of the random vector we define a margin function. The margin measures the extent to which the average number of votes at the random vector for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. The result that random forests do not over-fit as more trees are added but rather produces a limiting value of the generalization error follows from the strong law of natural numbers (probability theory). Therefore, an upper bound on the generalization error can thus be determined.

Support Vector Machine

Support vector machines are a popular tool in supervised learning if no prior knowledge about the domain is available. As mentioned in [2], three properties make SVMs attractive:

Maximum Margin Separation: SVMs compute the decision boundary in such a way that the distance to the closest datapoint on either side is maximized, which helps SVMs to generalize well.

Kernel Trick: Generally, SVMs create linear hyperplanes to separate data, but not always can such a linear separator be found in the original input space. The original data can be mapped to a higher-dimensional space using kernel functions. In this space, data is more likely to be linearly separable. Since the linear separator in the high-dimensional space is nonlinear in the original space, the hypothesis space is greatly expanded.

Nonparametric Method: SVMs are nonparametric, which gives them the flexibility to represent complex functions. On the other hand, because most of the time only a small fraction of the training data is retained, they are also resistant to overfitting.

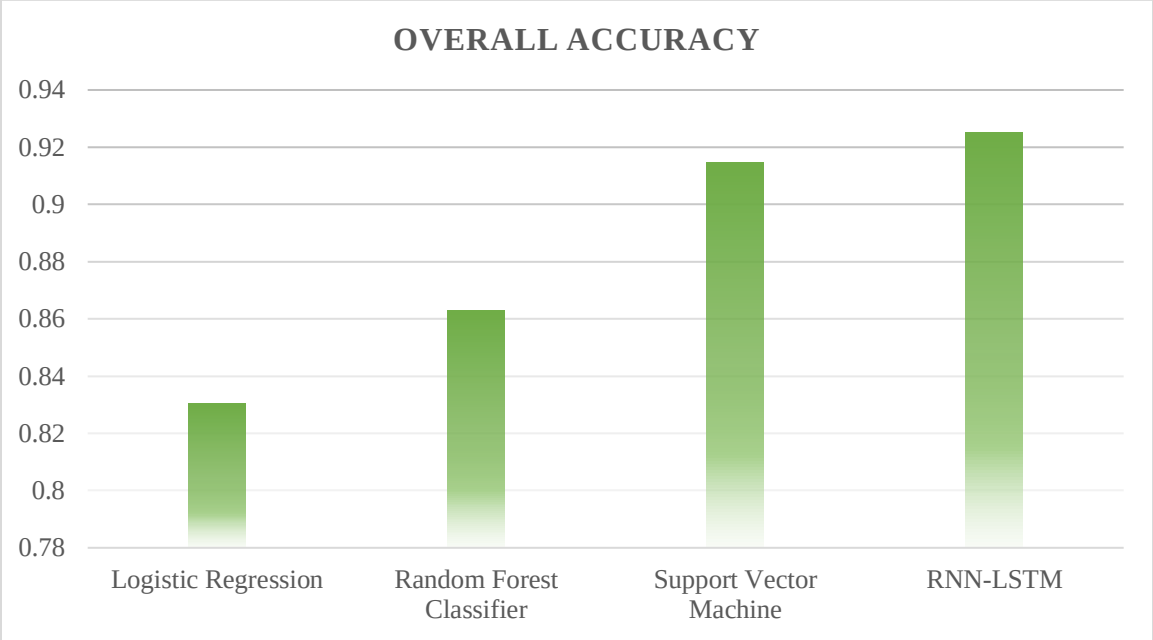
Recurrent Neural Network

Recurrent neural networks are mostly used on tasks with sequential data, such as speech recognition, grammar learning, or text prediction. Music shares a sequential nature with speech and text, as the flow from one note to the next determines the mood of melody and hence the genre. Given this knowledge, recurrent neural networks seemed like a logical next step. We settled on Tensorflow as our choice for a machine learning library because of its balance between ease-of-use and full control. The optimum network topology was decided by trial and error and It is observed that the following network is sufficient to differentiate genres from one another.

Network Topology		
Layer (Type)	Output Shape	Param #
LSTM (LSTM)	(None, 130, 796)	2747792
LSTM_1 (LSTM)	(None, 130, 384)	1814016
LSTM_2 (LSTM)	(None, 130, 128)	262656
LSTM_3 (LSTM)	(None, 64)	49408
Dense (Dense)	(None, 10)	650

RESULTS

The best performance in terms of accuracy is observed for the RNN-LSTM model that uses time varying features as an input to predict the music genre with a test accuracy of 92.5%. All other models used time invariant features which is seven summary statistics applied over the time components. Among those, support vector machine has the highest overall accuracy score.



The clearest trend identifiable here is the dramatic jump in performance as we increase the model complexity. We see a substantial improvement in a support vector machine model, which suggests that it is relatively the best estimator in terms of computational resources/time even with aggregating time component of the features.

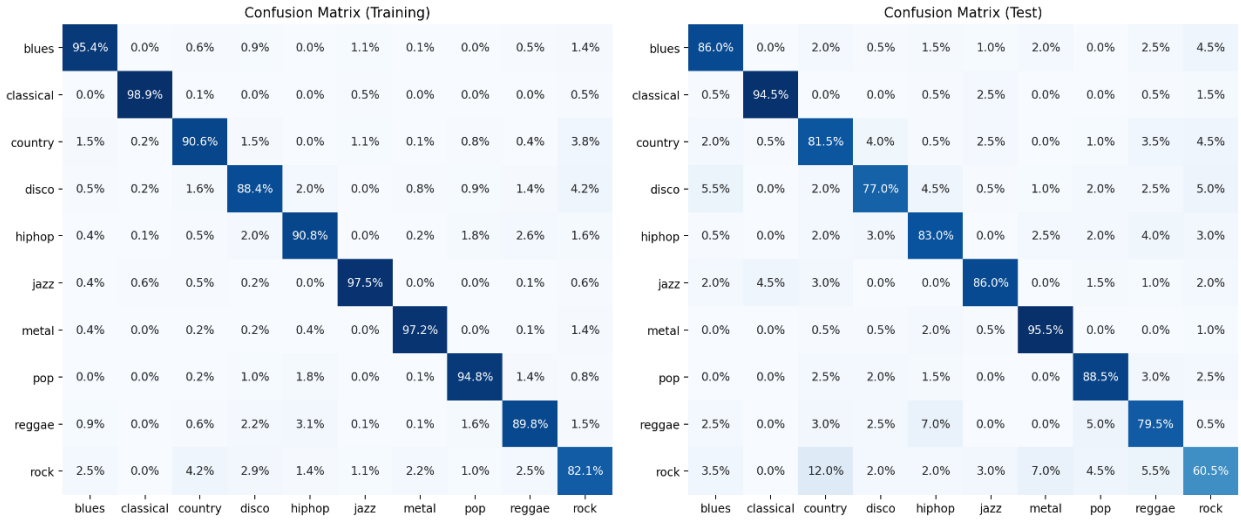


Figure: Confusion Matrix (Logistic Regression)

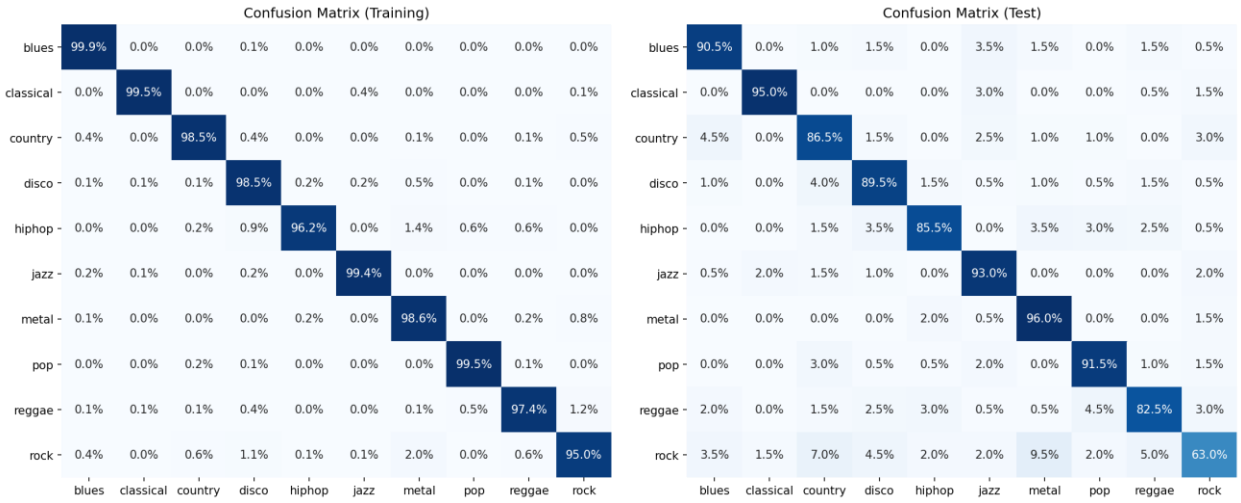


Figure: Confusion Matrix (Random Forest Classifier)

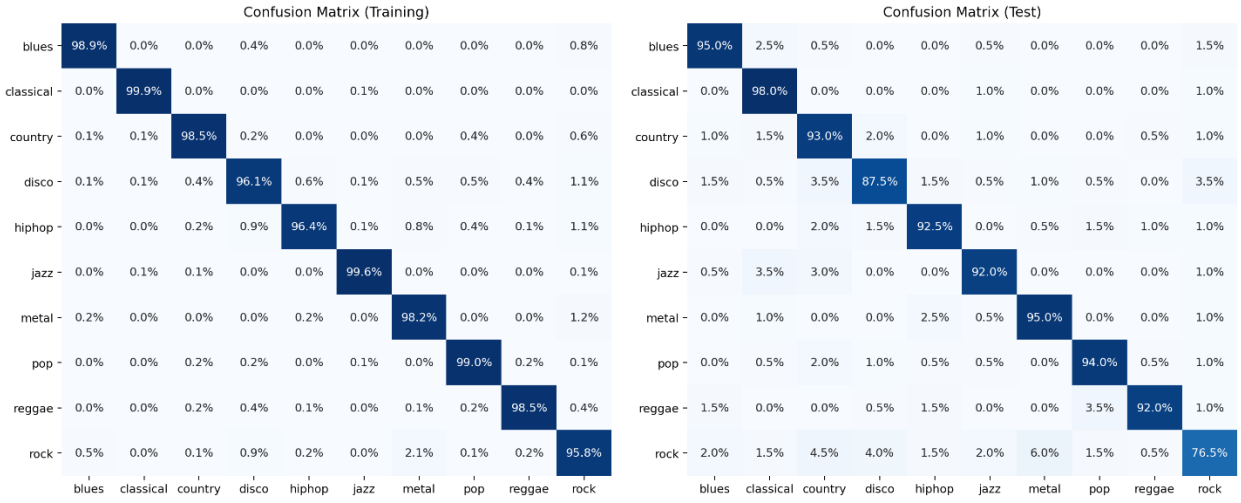


Figure: Confusion Matrix (Support Vector Machine)

Looking more closely at our confusion matrix, we see that all models struggled most with the rock genre. It only managed to correctly classify 76.5% at its best (SVM), labeling the others as mainly country or metal. Additionally, it incorrectly classified some disco, as well as a small fraction of reggae and jazz, as rock music. While it's not all that surprising that rock was a challenging genre – a qualitative inspection of rock mel-spectrograms implies that many rock music excerpts lack the easily visible beats that other genres such as hip-hop and disco possess. Additionally, rock is a genre that both encapsulates many different styles (light rock, hard rock, progressive rock, indie rock, new wave, etc.) and heavily influences many other derivative genres.

RNN

The LSTM network used in this project is a subclass of RNN. RNN is different from the traditional neural networks. It can memorize the past data and is able to predict with the help of the information stored in the memory. Moreover, LSTM solves the RNN long term dependencies problem. Although RNN model can make use of the past information to predict the current state,

the RNN model may fail to link up the information when the gap between the past information and the current state is too large. We believe that recurrent neural network with LSTM layers can fit multiclass classification problem that has time component. After long trials to tune hyperparameters of the model such as batch size, number of layers, and units, we believe that the optimum number of epoch and batch size is to be 150 for both. From the figure below, we can say that the model is overfitting after 110 epochs. We used a checkpoint callback function so that we load it later.

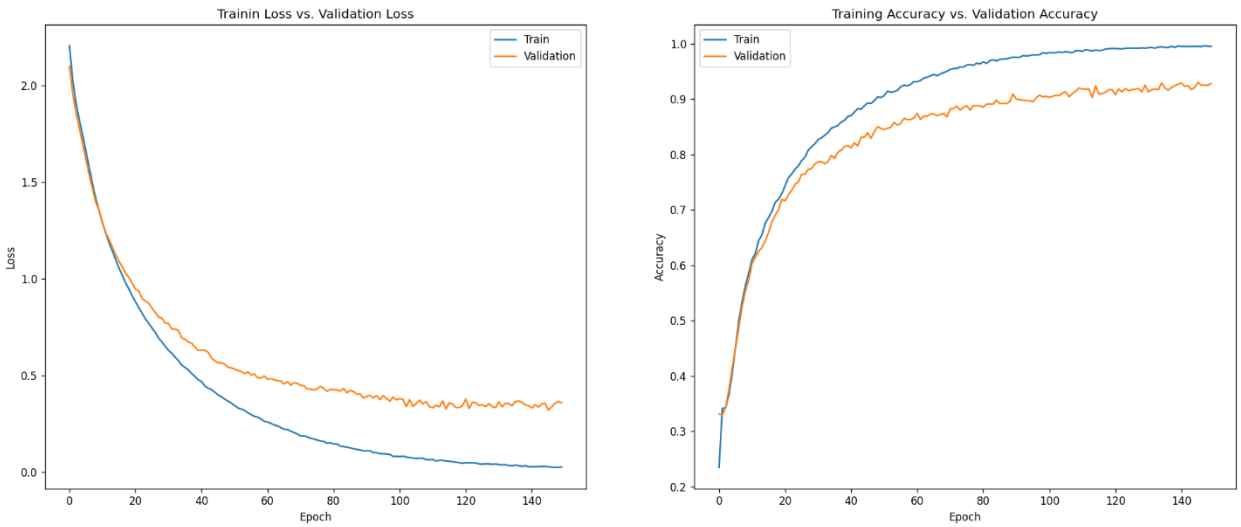


Figure: Training vs Validation Loss (RNN-LSTM)

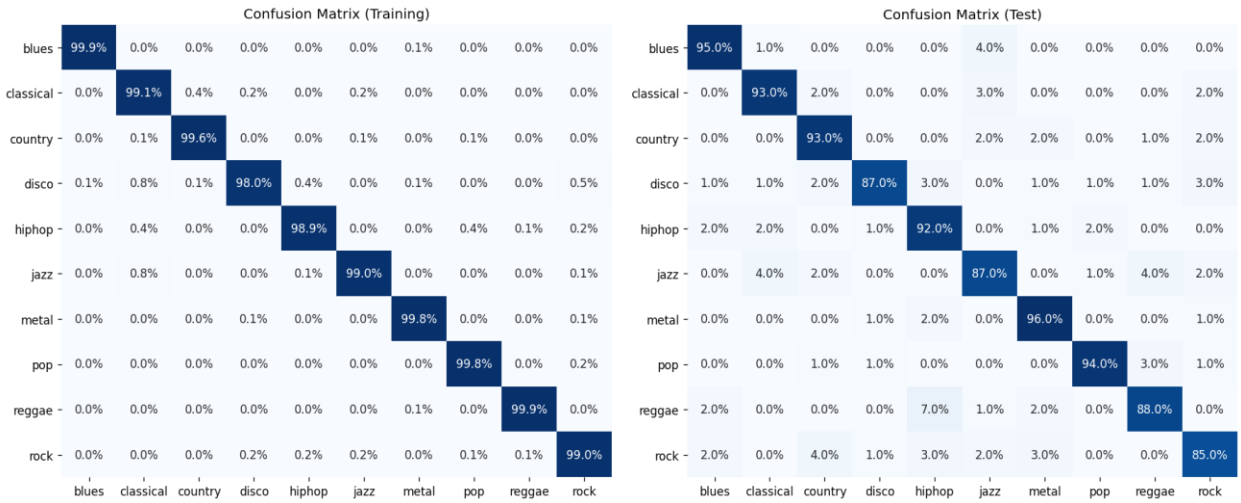


Figure: Confusion Matrix @110 epoch (RNN-LSTM)

RNN-LSTM model has the best overall accuracy, and it differentiates rock genre better than other models whereas it did not perform well for disco and jazz genre as much as support vector machine did.

FUTURE WORK

Since there are limited number of music samples, 1000, we augment our data to 10,000 by resampling. One drawback of this is that music mostly has chorus that repeat itself again and again. Therefore, our models might have seen some of the training data at testing stage. To overcome this issue, either number of resampling needs to be decreased or models need to be trained with a larger data.

We realized that each model performs better for certain genres. For example, RNN model has its best score for rock music, while reggae is classified best at support vector machine. This problem arises the need of ensemble models that averages out different models. We believe that this might make classifier more robust and accurate.

RECOMMENDATION TO CLIENTS

This model can classify 10 different genres by providing any length of music, the more is better. There are new genres introduced to music industry since this dataset was collected, 2001. Therefore, it cannot handle new genres or genres that are not in this dataset such as electronic, folk music etc.

REFERENCES

[1] Lee JH, Downie JS. Survey of music information needs, uses, and seeking behaviours: preliminary findings. In: Proceedings of the international conference on music, information retrieval; 2004.

[2] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach. Pearson Education, 2010.