Slotify: Ensemble Music Genre Classification

Data Science Project | November 13, 2020

Lucas Parzianello
University of Notre Dame
Notre Dame, Indiana
lbarbosa@nd.edu

Eric Tsai
University of Notre Dame
Notre Dame, Indiana
ctsai@nd.edu

ABSTRACT

Music genre classification is one of the best way of organizing music archives in a way that is interpretable and searchable by humans. However, teaching a machine to perform it similarly to a human remains a challenge. Our proposed method trains and evaluates two sets of traditional classifiers for music genre classification named Ensemble One and Ensemble Two. Ensemble One has the best 5 models ranked by their average F-1 score on 5 folds using cross validation. Ensemble Two has only three members that are selected to increase the ensemble diversity. With both ensemble models, we evaluate them against each other and against their components' individual performances gaining insights of what can be done to improve genre classifiers.

1 INTRODUCTION

In recent years, the increasing availability of music in several streaming services and personal libraries makes automation a requirement for properly organizing potentially millions of audio tracks. Genre classification is one of the ways we can use to organize such data. In order to automatically categorize the tracks into such categories, we need to first extract audio features from them. This extraction can be hand-crafted (i.e. designed by a specialist) or automated – in the case of the increasingly popular deep neural networks.

This work explores both methods of feature extraction. Our main contributions are:

- Answer which features are more determinant when solving the problem of music genre classification.
- Compare the confusion matrices of handcrafted and automated approaches in order to create an ensemble classifier more accurate than its components in isolation.

2 RELATED WORK

2.1 Common classification methods

In Bahuleyan [1], they explore the application of machine learning (ML) algorithms to identify and classify the genre of a given audio file. The conventional ML models that are often seen are Gradient Boosting, Random Forests (RF), Logistic Regression (LR), and Support Vector Machines (SVM). This paper mainly compares the performance of two different classes of methods:

- The first is to make prediction of the genre solely based on its spectrogram as input.
- The second approach is to make prediction of the genre based on features from frequency and time domain.

They train the four conventional ML classifiers mentioned above with these different features and compare their performances. The experiments are conducted on the Audio Set dataset [4] and have an AUC value of 0.894 for an ensemble classifier which combines the two proposed approaches mentioned above.

2.1.1 Hierarchical Taxonomy. In Li and Ogihara [8], they mainly focus on automatic music genre classification based on hierarchical classification with taxonomies. This paper introduce the concept of taxonomy. The hierarchical taxonomy identifies the connection between different genres and provides valuable sources of information for genre classification. This experiment displays different accuracy based on Flat- and Hierarchical-classification, and the Hierarchical-classification has a slightly higher performance in both of their testing datasets A and B.

With this technique, classifiers are able to take care of an easier separable problem and utilize an independently optimized feature set; this leads to improvements in accuracy apart from the gain in training and testing speed. The benefit of applying taxonomy makes the classification errors become more acceptable than in the case of flat classification, which is a type of Divide-and-Conquer approach that makes those errors fall within their level of the hierarchy.

2.1.2 CNN. In Zhang et al. [15], they proposed two ways to improve music genre classification with CNN:

- Method 1: Integrating max- and average-pooling to yield more statistical information to upper level neural networks;
- Method 2: Utilizing "shortcut connections" to bypass one or more hidden layers, a method inspired by residual learning method.

The methodology of their improved CNN is to implement a pile of CNN module, which is used as the feature extractor, for learning mid- and high-level features from the Spectrogram, and followed by a fully connected module, which is utilized as the classifier. CNN's are also used for music genre prediction by Bahuleyan [1].

3 DATA SOURCES

From our research, we have found a quite few options of datasets containing music tracks (excerpts or integral) with music genre

Data Science, 2020, University of Notre Dame, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

labels, enabling us to perform feature extraction for genre classification. Due to the public availability, ease of access, and good documentation, we have selected the FMA dataset. We also describe some of the alternatives we have considered below.

3.1 FMA

The Free Music Archive (FMA) dataset [3] is a publicly available alternative containing over 100 thousand audio tracks with four dataset versions of varying track number, lengths, and genres, ranging from 8 thousand tracks of 30 seconds of 7.2GB in total size; to over 106 thousand untrimmed tracks across 161 genres summing 879GB. The audio tracks are under a Creative Commons license and it appears to be the best documented alternative.

3.2 Other Datasets

3.2.1 GTZAN. GTZAN is one of the most popular public datasets for music genre recognition [13] and is composed of a thousand 30-second audio excerpts labeled across 10 music genres. Despite its popularity, the dataset was not created for music genre classification. Moreover, there are many critics about the dataset quality and whether its size is capable of allowing for accurate or significant results [13].

3.2.2 SYNAT. The SYNAT database [7] stores over 50 thousand 30-second music tracks in MP3 format, across 22 genres: Alternative Rock, Blues, Broadway & Vocalists, Children's Music, Christian and Gospel, Classic Rock, Classical, Country, Dance and DJ, Folk, Hard Rock and Metal, International, Jazz, Latin Music, Miscellaneous, New Age, Opera & Vocal, Pop, Rap and Hip-Hop, Rock, R&B, and Soundtracks. However, we were unable to find a working URL for download or a request form at the time, which forced us to choose an alternative.

3.2.3 MSD. The Million Song Dataset (MSD) [2] is a collection of one million songs for which over 190 thousand tracks have consistent genre annotations. Due to the large size of the dataset (around 300GB), MSD is publicly available for research purposes as an AWS EC2 snapshot, rather than a direct download. For our purposes a smaller collection would suffice, but MSD remains an adequate alternative.

4 FEATURE EXTRACTION

We are currently using the features listed below, extracted with the help of `librosa` [11] – a Python package for music and audio analysis.

- Zero Crossing Rate (ZCR) [9]
- Root Mean Square Energy (RSME) [14]
- Mel-Frequency Cepstrum Coefficients (MFCC) [6, 9, 10, 12]
- Spectral Centroid, Bandwidth, Contrast, Roll-Off [8, 9]
- Chroma Features
- Tonal Centroid Features (Tonnetz) [5]

In addition to those, the features that we have found in the literature but are currently not extracted:

- Daubechies Wavelet Coefficient Histogram [9]
- Central Moments (CM)
- Tempo

4.1 Feature Overview

4.1.1 Zero Crossing Rate. The zero-crossing rate is the rate at which a signal changes from positive to negative and from negative to positive. This is a key-feature in many audio processing tasks - from speech recognition to music information retrieval.

4.1.2 Root Mean Square Energy. The energy of a signal E_s relates to how loud the signal is and it is defined as the total magnitude of the signal. The RMSE is computed from this energy across a sliding window of N frames:

$$E_s = \sum_n |x(n)|^2$$

$$RMSE = \sqrt{\frac{1}{N} \sum_n |x(n)|^2}$$

4.1.3 Mel-Frequency Cepstrum Coefficients (MFCC). We have found MFCC to be one of the most relevant features in our classification. To understand the Mel-Frequency Cepstrum Coefficients, let's first define what Cepstrum means. Starting from a sound wave in the time domain, we can generate its spectrum of frequencies using the Discrete Fourier Transform. Because of the nature of sound waves, it makes more sense to visualize these frequencies in the log scale. Now with the log power spectrum of our original signal in the frequency domain.

By applying the Inverse Fourier Transform on this spectrum, we have the Cepstrum, which can be interpreted as the information about the rate of change in the different spectrum bands. Similar to a "spectrum of spectrum".

When the spectrum is first transformed using the Mel scale, the resulting information is called the Mel-Frequency Cepstrum or MFC, and its coefficients are called Mel-Frequency Cepstral Coefficients, or MFCC's. One key advantage of MFCC's is that they are able to describe the large structures of the spectrum, being able to capture the timbre of the tracks. This makes them a good option for genre classification.

4.1.4 Spectral Centroid, Bandwidth, Contrast, and Roll-Off. This set of spectral features help to make sense of different properties related to the track analyzed in the frequency domain. The spectral centroid characterises a spectrum by its center of mass, it is which frequency the spectrum is centered upon. It is defined as the weighted sum of frequency bins using their magnitude as weights. The spectral bandwidth is proportional to the difference of the frequencies at each bin and the spectral centroid. Depending on its hyperparameter p, it can be a weighted standard deviation of frequencies. The spectral contrast measures the difference of the spectral peak and valley for each frequency sub-band. Finally, the spectral roll-off is the frequency, below which a specified percentage of the total spectral energy lies. For example, 85% of the energy lies below 2500 Hertz.

4.1.5 Chroma Features. Chroma features are useful to analyze the pitch of an audio track along time. Chroma features capture harmonic and melodic characteristics of an audio track while being robust to changes in timbre and instrumentation. For example, the

same song played on a piano will have a similar chromagram than when performed on a guitar.

4.1.6 Tonal Centroid Features (Tonnetz). The Harmonic Network, also known as Tonnetz, represents pitch relations and it's used in music theory for centuries. The Tonal Centroid Features used in our classifier is a projection of Chroma Features onto a 6-dimensional basis representing the close harmonic relations of fifths, minor and major thirds, each with a set of 2 coordinates. Chroma Features and Tonnetz help our classifier to see the music as a composition of notes and harmonics rather than a complex merger of frequencies.

5 EXPERIMENT SETTINGS

For training purposes, we split the FMA-Small dataset into 8:1:1, respectively training, validation, and testing. In these sets, as shown in Figure 1, the tracks are equally distributed in each genres, with 800 and 100 tracks per genre respectively on training and testing set. The exact numbers are 6400:800:800 respectively for training, validation, and testing.

5.1 Models

After the feature extraction, we make them as input and send them to 13 models for training process:

- Logistic Regression (LR)
- Random Forests (RF)
- K-Nearest Neighbors (KNN)
- 2x Multi-Layer Perceptron (MLP)
- Support Vector Machines (SVM)
  2x Linear, Polynomial, and RBF kernels
- Decision Tree Classifier (DT)
- AdaBoost
- Gaussian Naive Bayes (NB)
- Quadratic Discriminant Analysis (QDA)

6 FEATURE AND MODEL SELECTION

From the heatmap in Figure 2, we have found that the most decisive individual feature for genre classification across the classifiers tested is Mel-Frequency Cepstrum Coefficients (MFCC). Followed by the second most decisive feature is Spectral Contrast feature. The last rows of the heatmap combine these two features in sets with Spectral Contrast, Spectral Centroid, Chroma, Tonnetz and ZCR. When combined into feature sets, the experiments show a slight improvement on the accuracy and F1 measure on some classifiers, but a very similar performance to MFCC + Spectral Contrast overall. This performance is ratified by the F1 measure across different sets of features and classification models, shown in Figure 3. Finally, the combination of all features does not indicate a noteworthy improvement on the metrics.

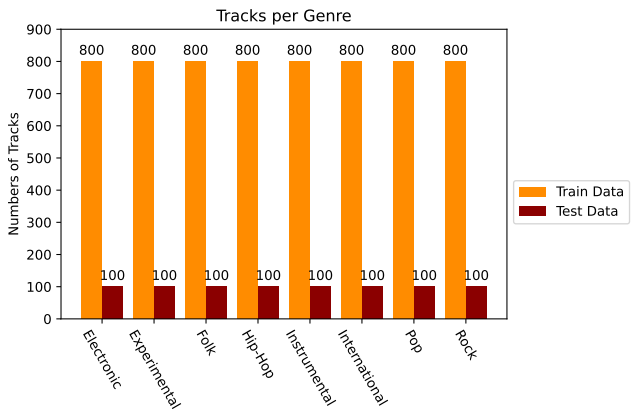

Figure 1: Tracks per genre.

Based on that, we select the **Mel-Frequency Cepstrum Coefficients** and **Spectral Contrast** as our fixed feature sets for our ensemble models. For model selection, we form two ensembles creatively named *Ensemble One* and *Ensemble Two*. Ensemble One has the best 5 models ranked by their average F-1 score on 5 folds using cross validation. These happened to be similar models: 4 out of 5 are based on SVM's. Ensemble Two has only three members that are selected to increase the ensemble diversity. We then evaluate these against each other and against their individual components.

Ensemble One:

- SVC RBF kernel
- Logistic Regression
- SVC Polynomial kernel
- Linear SVC 1
- Linear SVC 2

Ensemble Two:

- SVC RBF kernel
- Logistic Regression
- MLP 2

6.1 Voting Mechanism

For the ensemble decision, we use a majority voting mechanism. In case of a tie (e.g. 2-2-1 for an ensemble of 5 classifiers), we select the genre picked by the top F-1 classifier on the cross validation analysis.

7 EVALUATION

In order to compare the classification models and fulfill our contributions of (i) which features are most relevant in the handcrafted method, and (ii) build an ensemble model for music genre classification; we plan to extract a list of metrics from our classifiers. Firstly, our dataset will be split into training, validation, and testing sets with disjoint audio tracks and uniform representation across music genres, when possible. Then, once the classifiers models are built, we will extract the metrics below in isolation, and lastly, from our ensemble version.

7.1 Metrics

To evaluate our model we use the following metrics:

- Mean accuracy – for a simplified overall idea of a model's performance;
- F1 score – takes into account precision and recall;
- Confusion matrix across genres – to identify which pairs are most challenging for each model and use this information to improve a collective decision;

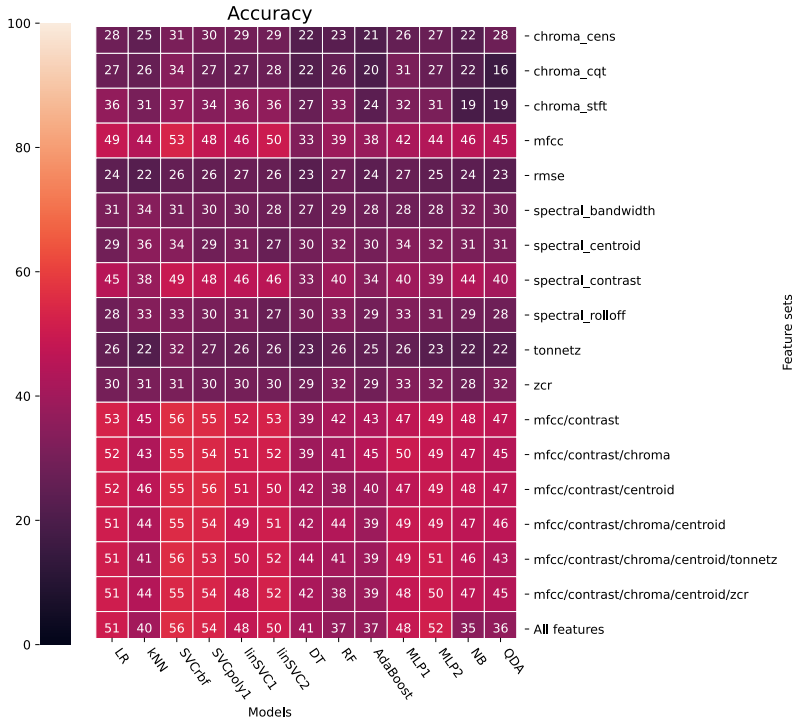

Figure 2: Models vs. Features Sets and their Accuracy-measures on the validation set [0-100%].

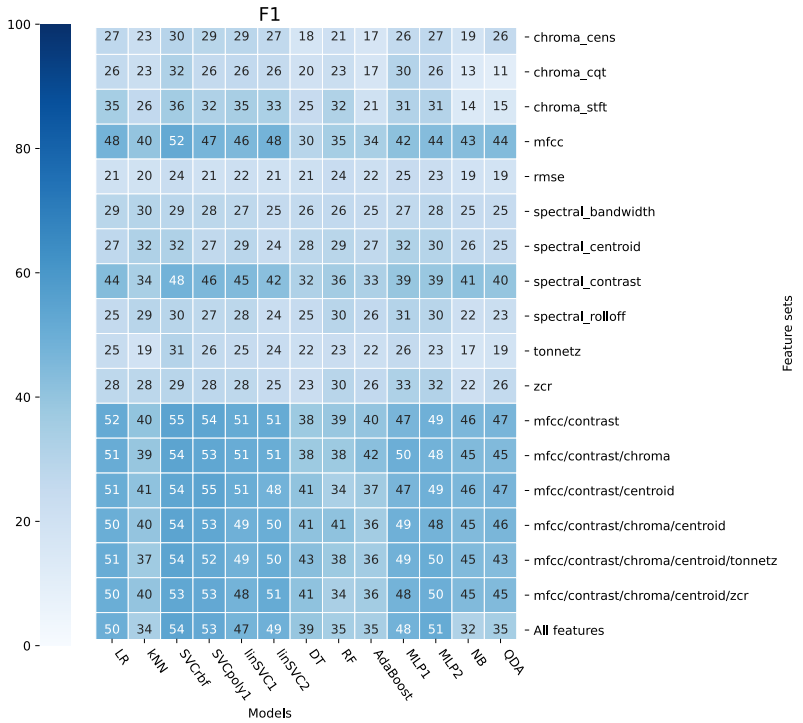

Figure 3: Models vs. Features Sets and their F1-measures on the validation set [0-100%].

7.2 Experiment Setup

The FMA dataset has tracks with more than one genre labelled, thus, we could attempt a multi-class genre classification. However, since we have a total of only 8 classes in the FMA-Small variant, we preferred to use the most predominant class as our ground-truth instead: we select the database attribute named `genre-top` as our genre label ground truth. FMA-Small contains 8,000 tracks in 8 different genre categories: Hip-Hop, Pop, Folk, Experimental, Rock, International, Electronic, and Instrumental.

7.3 Ensemble Performance

By running 5 folds, we get each individual classifier performance and our final ensemble performance with validation data. After training with our training data, Figure 4 Ensemble One has F-1 around 56%, and in Figure 5, Ensemble Two shows a similar F-1 of 59%. However, when we test with testing data, we have around 46% on F-1 with Ensemble One and Two. When we look at the numbers carefully, both of Ensemble One and Two F-1 performances are mostly better than the individual members' performances. Out of our surprise, we found out that one of the individual model, RBF SVC, has outperformed our two ensemble versions on both validation and test data.

7.4 Confusion Matrices

Visually, the two confusion matrices in Figures 6 and 7 are very similar. In both we notice great confusion for Folk music classification, which is more often misclassified as Experimental and International than correctly predicted. Considering the classes have a uniform distribution, we see that for both ensembles, the easiest i.e. most often correctly classified genres are: Hip-Hop, Electronic, and Rock.

There is also a high error rate of Pop songs being classified as Electronic, but not the other way around. This might indicate a tendency of pop songs to have electronic components (at least in this particular dataset), while still being seen as pop. Pop music may be more determined by their wide adoption than by a particular set of instruments, tempos, or melody, as the name suggests. This confusion might be an indicator that the set of extracted features lack a "popularity" metric to confidently label a track as "Pop".

The high imbalance of correct prediction across classes suggests that there is a lower variance of styles in these easier classes in this dataset, while the seemingly higher variance of Folk, Pop, and Experimental might be counteracted by an expanded set of features, perhaps towards the track's origins and receptivity instead of technical analysis, since the other features extracted do not seem to improve the overall performance significantly, as showed in Figure 3. Perhaps features such as geographical location and a popularity score would help in discerning the targeted public and indirectly assisting the genre classification.

8 CONCLUSION

Model Selection. With the obtained results, our RBF SVC seems to consistently outperform the other individual methods and be slightly better than both ensemble models created in the validation and testing sets. The usage of an ensemble model with greater diversity such as Ensemble Two did not present any clear advantage, although the choice of an ensemble as opposed to a single classifier might still generalize better for a different dataset. This remains to be verified.

Feature Selection. Regarding the choice of features, **MFCC** and **Spectral Contrast** presented a clear advantage over other features in most of the classifiers tested, proven to be good predictors for genre. The combination of these two features resulted in an even superior accuracy across all classifiers. Moreover, the addition of the other features extracted did not result in any clear advantage – the addition of all features even performed worse for most individual classifiers.

9 FUTURE WORK

Neural networks are likely to perform better for a same dataset, but it is hard to assert their advantage across datasets. Because of the learned feature extraction, they may overfit to their training data and have stability issues across datasets with different properties. With these potential weaknesses enlightened, however, it would be interesting to see the comparison of a neural-network approach against an ensemble on the same and different datasets in a leave-one-out cross validation.

A second possible follow-up is the usage of further metadata for genre classification, exploring the tracks not only for what they are (as a time series), but exploring beyond that and add information on how they are perceived by the public (e.g. popularity, demographics of listeners) and the creators (e.g. title, album, release date, lyrics). Works exploring this kind of data can go beyond genre classification and even serve as predictors of when, where, and for whom a given track is more likely to succeed.

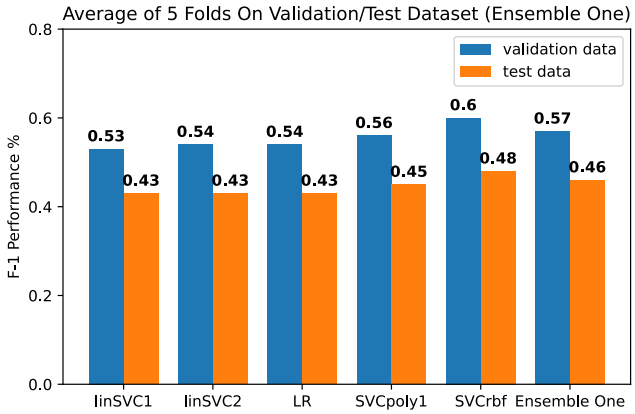

Figure 4: Ensemble One and its members' F1 performance.

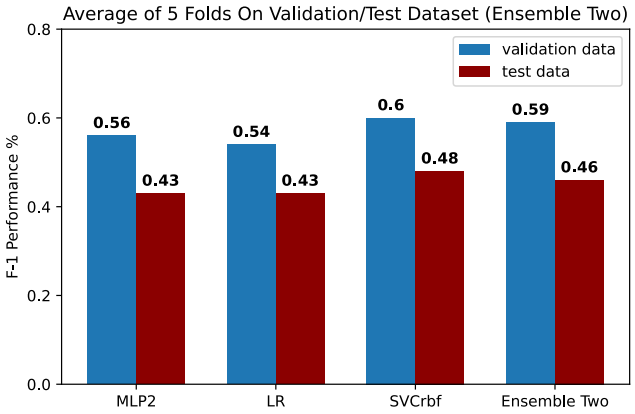

Figure 5: Ensemble Two and its members' F1 performance.

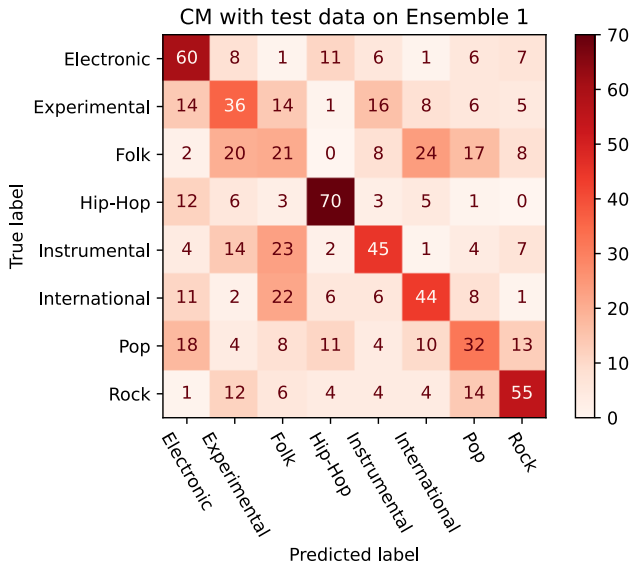

Figure 6: Confusion Matrix for Ensemble One.

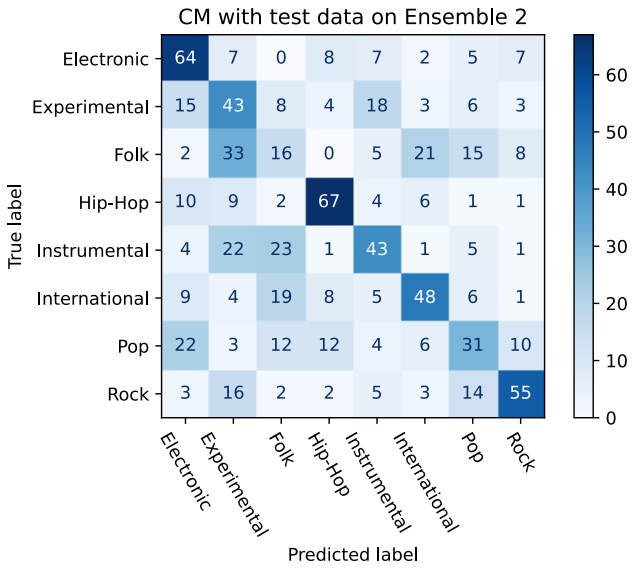

Figure 7: Confusion Matrix for Ensemble Two.

REFERENCES

[1] Hareesh Bahuleyan. 2018. Music Genre Classification with Machine Learning Techniques. (2018), 1–4. https://doi.org/10.1109/siu.2017.7960694 arXiv:arXiv:1804.01149v1

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011* (2011), 591–596.

[3] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A Dataset for Music Analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*. arXiv:arXiv:1612.01840v3

[4] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2017), 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[5] Christopher Harte, Mark Sandler, and Martin Gasser. 2006. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia - AMCMM '06*. ACM Press, New York, New York, USA, 21. https://doi.org/10.1145/1178723.1178727

[6] Piotr Hoffmann, Andrzej Kaczmarek, Paweł Spaleniak, and Bożena Kostek. 2016. Music Recommendation System. *Asian Journal of Information Technology* 15, 21 (2016), 4250–4254. https://doi.org/10.3923/ajit.2016.4250.4254

[7] Bozena Kostek, Adam Kupryjanow, Pawel Zwan, Wenxin Jiang, Zbigniew W Raś, Marcin Wojnarski, and Joanna Swietlicka. 2011. Report of the ISMIS 2011 Contest: Music Information Retrieval. In *Foundations of Intelligent Systems*, Marzena Kryszkiewicz, Henryk Rybinski, Andrzej Skowron, and Zbigniew W Raś (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 715–724.

[8] Tao Li and Mitsunori Ogihara. 2005. Music Genre Classification with Taxonomy. *In Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA* (2005), 197–200.

[9] Tao Li and Mitsunori Ogihara. 2006. Toward Intelligent Music Information Retrieval. *IEEE Transactions on Multimedia* 8, 3 (2006), 564–574. https://doi.org/10.1109/TMM.2006.870730

[10] Shin Cheol Lim, Jong Seol Lee, Sei Jin Jang, Soek Pil Lee, and Moo Young Kim. 2012. Music-Genre Classification System Based on Spectro-Temporal Features and Feature Selection. *IEEE Transactions on Consumer Electronics* 58, 4 (2012), 1262–1268. https://doi.org/10.1109/TCE.2012.6414994

[11] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference* Scipy (2015), 18–24. https://doi.org/10.25080/majora-7b98e3ed-003

[12] Loris Nanni, Yandre M.G. Costa, Alessandra Lumini, Moo Young Kim, and Seung Ryul Baek. 2016. Combining Visual and Acoustic Features for Music Genre Classification. *Expert Systems with Applications* 45 (2016), 108–117. https://doi.org/10.1016/j.eswa.2015.09.018

[13] Bob L. Sturm. 2013. The GTZAN Dataset: Its Contents, Its Faults, Their Effects on Evaluation, and Its Future Use. 11 (2013), 1–29. https://doi.org/10.1080/09298215.2014.894533 arXiv:1306.1461

[14] Ran Tao, Zhenyang Li, and Ye Ji. 2010. Music Genre Classification Using Temporal Information and Support Vector Machine. In *ASCI Conference, 2010.*

[15] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing. 2016. Improved Music Genre Classification with Convolutional Neural Networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-Sept (2016), 3304–3308. https://doi.org/10.21437/Interspeech.2016-1236