Automatic Segmentation of Music Genres using Tensorflow CNN and Music Information Retrieval Technique

Prerana CH

Computer Science and Engineering

Vidya Jyothi Institute of Technology

Aziz Nagar, Telangana, India

preranacheguru@gmail.com

Sahiti Cheguru

Computer Science and Engineering

Gokaraju Rangaraju Institute of Engineering and Technology

Bachupally, Telangana, India

sahiticheguru2000@gmail.com

K. Tejasree

Computer Science and Engineering

Vidya Jyothi Institute of Technology

Aziz Nagar, Telangana, India

tejasreekomati05@gmail.com

Dr. Siddhartha Ghosh

Professor & Head of the Dept. of Artificial Intelligence (AI)
Head – Training and Placements

Vidya Jyothi Institute of Technology

Aziz Nagar, Telangana, India

siddhartha@vjit.ac.in

Tarine Deepthi

Computer Science and Engineering

Vidya Jyothi Institute of Technology

Aziz Nagar, Telangana, India

tarinedeepthi1998@gmail.com

ABSTRACT

Music is becoming increasingly easier to consume by way of apps on the internet and songs. Perhaps the most popular feature of the music is the musical form. A complex task in the field is the grouping of music tracks according to their criteria. For the structured arrangement of audio files and accordingly for the increasing interest in genre grouping in automated songs. Additionally a critical aspect of the identification and aggregation of music in related genres is the recommended method for song and album generator. Unconsciously, music training reflects the user's moment. It is an important undertaking over the respective field to describe and analyse these moments. We explore the effect of machine learning techniques on the implementation of models of predictive genre classification aimed at capturing distinctions between genres. The analysis of genres as aggregated bodies of musical publications such a notion includes analogous inductive models as per arbitrary and local parameters. This process can thus be modeled as an example-driven process for learning.

KEYWORDS

GTZAN, Genre, Machine Learning, Information Retrieval, Mel-frequency Spectrogram.

1. INTRODUCTION

The playlists mostly arranged by genre [1] are one main aspect of these facilities. This knowledge might come from the manual marking by the persons that publish the tracks. But this does not scale well and artists who want to capitalize on the success of a particular genre [2] could play games. A better choice is to focus on the automatic segmentation of music genres. Machine learning methods used as induction models to determine data segmentation. Although the study published here is our attempt to apply an inductive approach to genre [3] classification by using GTZAN [1] data our current work also explores audio properties for the same collection of tracks the features of the relevant domain of the function. We developed various models depending on the form of inputs. Each example is an audio sample of the 30s or roughly 1.3 million data points of raw audio. At a certain point in time, these floating-point values positive or negative reflect the displacement of the wave. Less than 1 percent of the data can be used to handle computing resources.

The reality that significance and relation are not localized or impartial material resources in information retrieval global concepts that arise from the whole text through the foundation are known well. In Information Retrieval [1], any quantitative model relies on a large number of parameters. We are going to try to build a classifier to categorize songs into various genres [4]. Let us imagine a situation in which for whatever reason we find on our hard drive a bunch of arbitrarily called mp3 files that are supposed to contain songs. Our goal is to organize them into various directories such as jazz, classical, country, rock, and metal according to the music genre.

1.1. Related work

The core concept of extracting information from music, which is Music Information Retrieval (MIR) [2], began a long time ago at the beginning of 2002 George Tzanetakis and Perry Cook published a paper in which they acquired the essence of interpreting music data and processing auditory data. It also launched a GTZAN [2] dataset named after George Tzanetakis, consisting of 1,000 songs and 30 seconds each. They began to identify the music genre as a challenge for pattern recognition.

A lot of study followed, such as Bertin-Mahieux, introduced in 2011 a dataset consisting of millions of songs along with their audio features and metadata. But this work posed problems in the Music Information Retrieval [3], such as the identification of cover songs the song year prediction is one of them. After that work was done in genre [5] recognition with the aid of lyrics in one of the papers by Alex Tsaptsinos published in ISMIR 2017. By using the Hierarchical Attention Network (HAN), she made a fair classification of twenty song genres [6]. In 2011, Sander Dieleman and two others published a paper using a pre-trained neural neural network for genre [7] and artist recognition.

Genre [8] identification was also presented using spectrograms by Yandre Costa and the team in a paper published in 2013, which examined the extraction of features from spectrograms using local binary patterns, and predicted using a complex collection of classifiers. Their model was tested against Latin Music, which resulted in 83% accuracy. In 2014, Grzegorz Gwardys and Grzywczak used CNN on spectrograms and were able to achieve a 70 percent accuracy rate on the GTZAN [3] dataset. This concept of using spectrograms and CNN model motivated us to create a CNN model to achieve more precision and to identify songs.

1.2. Existing System

Current statistics show that gender classification is accomplished by using four or more machine learning [1] algorithms with the support of some handcrafted features. After classical Machine learning [2] algorithms, experiments are carried out using neural networks based primarily on CNN with XGBoost to achieve better accuracy. The extraction of audio features has become a big problem when it comes to classifying songs. GTZAN [4] used audio signals to distinguish trends between different genres of music, achieving a respectable 61 percent accuracy in classifying genres [8].

Many algorithms and neural nets are used for the GTZAN[5] dataset by extracting various features such as MFCC[1], chromium, spectral roll-off, spectral centroid, etc. In 2015, Ritesh, Klein, and Rosman used Logistic Regression on the same dataset and achieved 81% accuracy on GTZAN [6] 10 genres [9]. Basically, as in previous documents, the degree of correctness is not obtained by the use of current machine learning [3] models or any neural network models.

Du, Lin and other participants have developed a new hierarchical analysis to extract features from audio signal spectrograms. From their paper, we got the idea to extract the features using the spectrograms and the Mel scale readings. Our work aims to improve the precision of the classification of different GTZAN [7] genres [10].

2. PROPOSED SYSTEM

2.1. Data

Initially, we unsubstantiatedly searched for music knowledge accessible as an open-source resource. We've been trying to get structured data for easy pre-processing so that we can achieve good accuracy on our model. Few common datasets available for music classification are the Million Songs Dataset, the Free Music Archive, and the GTZAN [8] Genre [11] Categorized Dataset.

Between these common datasets, there are problems with Million Songs Data and FMA, as the measurement of information in their individual datasets is extremely difficult for the preparation and processing of the measurement of information. As Million songs dataset also does not include direct raw audio files as data it asks us to download from online streaming platforms. As a result, we used the GTZAN [9] dataset, which arranged ten genre [12] of song data to help make the classification process simpler.

Genre	Number of Samples	Length of each Sample
blues	100	30 seconds
classical	100	30 seconds
country	100	30 seconds
disco	100	30 seconds
hiphop	100	30 seconds
jazz	100	30 seconds
metal	100	30 seconds
pop	100	30 seconds
reggae	100	30 seconds

Table 1: GTZAN data glance. GTZAN dataset which has organized ten genres of song data that helps in making the classification process easy.

2.2. Feature Extraction

You can remove a lot of functions and functionality from any audio file. The simplest sound can be impersonated as a waveform, i.e. a 2D coordinate system, where the X-axis represents time and the Y-axis is amplitude. Typically, these waveforms are stored as 1D array or lists. The basic function of an audio waveform can be represented using spectrograms.

The spectrogram is essentially a frequency continuum for an audio signal over time. All of these simple transformations to spectrograms are based on the Fourier Transform (FT) formula.

$$\infty$$

$$(u) = \int_{-\infty} (y)^{-i2\pi y} \partial y \qquad (1)$$

Where z(y) is a function of time and y is a function of time. Z (u) depicts the transformed frequency function The FT's after-effect is only a pool of containers that depicts the frequency along with their magnitudes, which are represented as 1D array data.

With sound details, particularly music, a well-known move up to the spectrogram is the use of a Mel-scale rather than a directly separated recurrence scale. A Mel-scale depends on the analysis of the pitch. As recurrence builds, comparable Mel-interims require larger and larger recurrence leaps. Steve's formula that converts f (Hertz) to m (mels)

$$m = 2595 \log10 \left(1 + \frac{f}{700}\right) \qquad (2)$$

This is the history of the Mel spectrograms [2], where the output can be interpreted as an image in a 2D frequency (Hertz) coordinate system on the X-axis and time on the Y-axis. The example of the Mel spectrogram [3] for a song is as follows:

2.3. Approach

Compare the classic approach to extract features and use a classifier (e.g. SVM) against the Deep Learning approach of using CNNs for audio representation (Mel spectrogram [4]) to extract and classify features. Resume of the deep learning approach:

Shuffle the input and split into train and test (70%/30%). Read the audios as melspectrograms, splitting then into 1.5s windows with 50% overlapping resulting in a dataset with shape (samples x time x frequency x channels). Train the CNN and test on test set using a Majority Voting approach.

As of now, after completing the data collection and extraction function parts, we have left the audio file image detail. In Neural Networks, all other deep learning techniques stand out when it comes to dealing with image data.

That's why we chose to feed our Mel spectrogram [5] as input for the CNN model. We've been putting picture info on our CNN model. The division of the training set with training and testing percentages of 70 and 30 gave good results for our model. First, the Mel spectrogram [6] image of each audio file is passed through five separate convolution layers to extract different features from the image. After each sheet, the model is sent through the Relu activation feature and maxpooling is finally completed. Considering 2D convolution, it was easy to identify time-related patterns.

Finally, after going through all five convolution layers, it is sent to the flattening layer where all values are flattened to the 1D array and transferred to a completely linked neural network, which returns the probabilities of all ten genres. So all genres are one hot encoded that we prefer to use a categorical cross-entropy loss metric.

We used Adam optimizer for optimization and proceeded to adjust the learning rate from time to time to produce better performance. We used the Softmax activation feature for the output layer, which worked well for our model.

We ran the model at 150 epochs each consisting of a batch size of 128. By changing the different parameters and the learning rate, the accuracy rate ranged from 75 to 82%. The highest accuracy we have achieved is 83 percent, which is embedded in the user interface that will be addressed in the results section below.

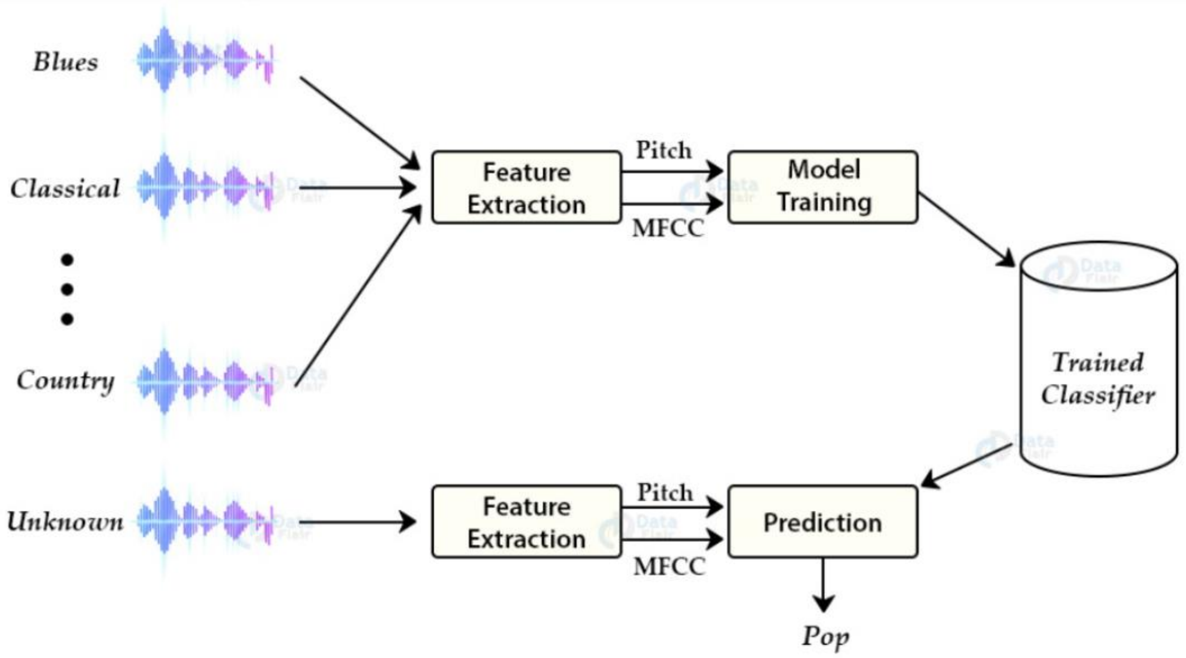

Figure 1: The figure explains the standardized classifying task system architecture. We need to gather information from audio samples such as spectrograms, MFCC, then using a model to categorize the genres of music.

As each audio track is in .wav format. The first step in the music genre classification project will be to separate modules from audio files it involves detecting and discarding noise from linguistic material there are several steps for certain features to be generated since the audio signals are varying continuously we first segment these signals into smaller frames each frame is estimated 20-40ms long. Then we attempt to describe various frequencies in each frame now distinct linguistic frequencies from noise frequencies it then requires a discrete cosine transform of these frequencies to discard the noise using discrete cosine transform we only retain a specific segment of frequencies with high data accuracy.

3. RESULTS

In order to compare the results across multiple architectures, we have taken two approaches to this problem: one using the classic approach of extracting features, and then using a classifier. The second method, which is implemented in the src file, is a Deep Learning technique that feeds CNN with a melspectrogram [7].

Model	**Acc**
Decision Tree	0.5160
Random Forest	0.6760
ElasticNet	0.6880
Logistic Regression	0.7640
SVM (RBF)	0.7880
CNN2D	0.832

Table 2: Current results obtained on the test set after extracting the features. The result obtained from the deep learning approach which we have tested with a simple custom architecture.

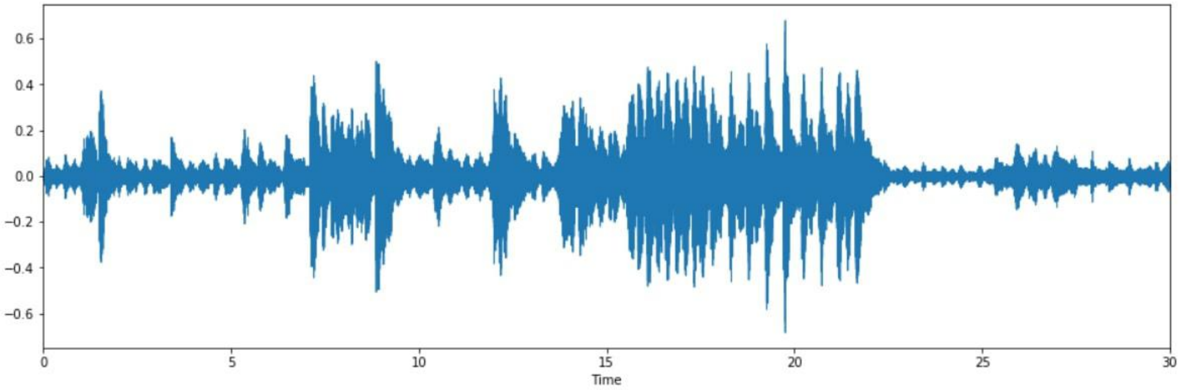

Figure 2: Visualizing Audio file as a Wave form. Data visualization of the audio wave form with the usage of matplotlib and librosa libraries

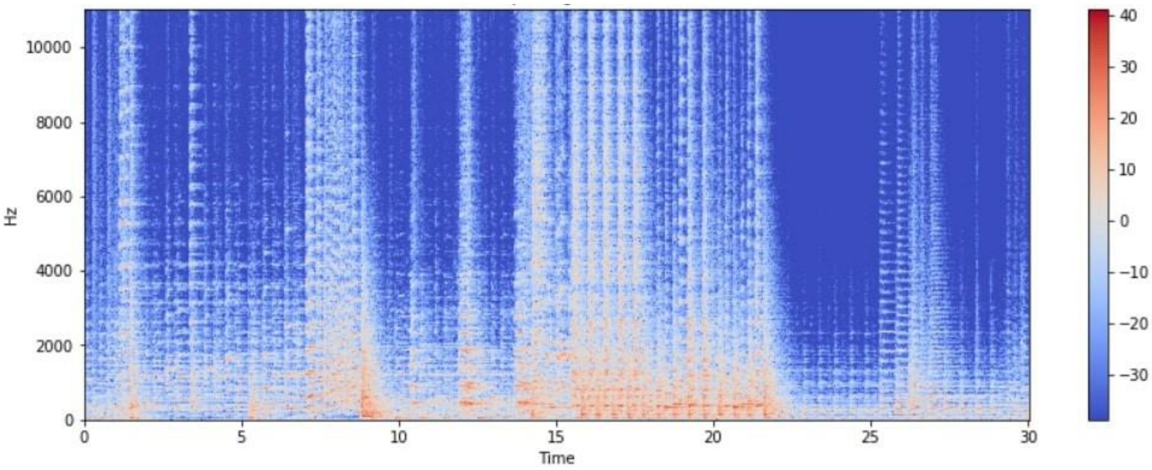

Figure 3: Visualizing Audio file as a Spectrogram. Data visualization of the audio wave form with the usage of matplotlib and librosa libraries

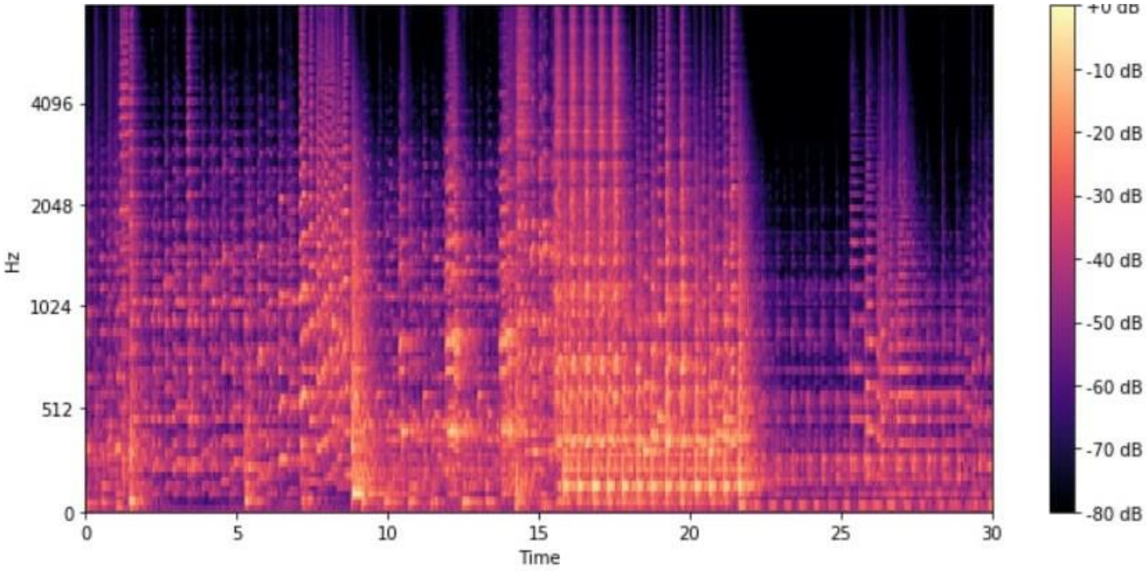

Figure 4: Visualizing Audio file as MelSpectrogram. Data visualization of the audio wave form with the usage of matplotlib and librosa libraries

When we checked our model against custom feedback like Ananya Birla Better.mp3, it turned out to be a type of hip-hop, which is pretty true for that album. Generally speaking, the model was defined and achieved the most notable accuracy of 83 per cent when measured against the GTZAN [10] dataset.

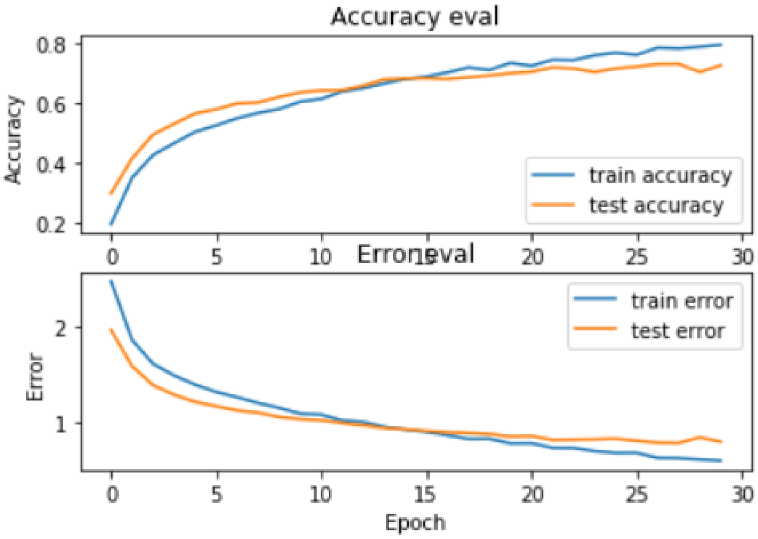

Figure 5: Model's Accuracy and Error graphs of Train and Test Accuracies and Errors. Learning Curves – used for model selection; Epoch 4 has the minimum validation loss and highest validation accuracy

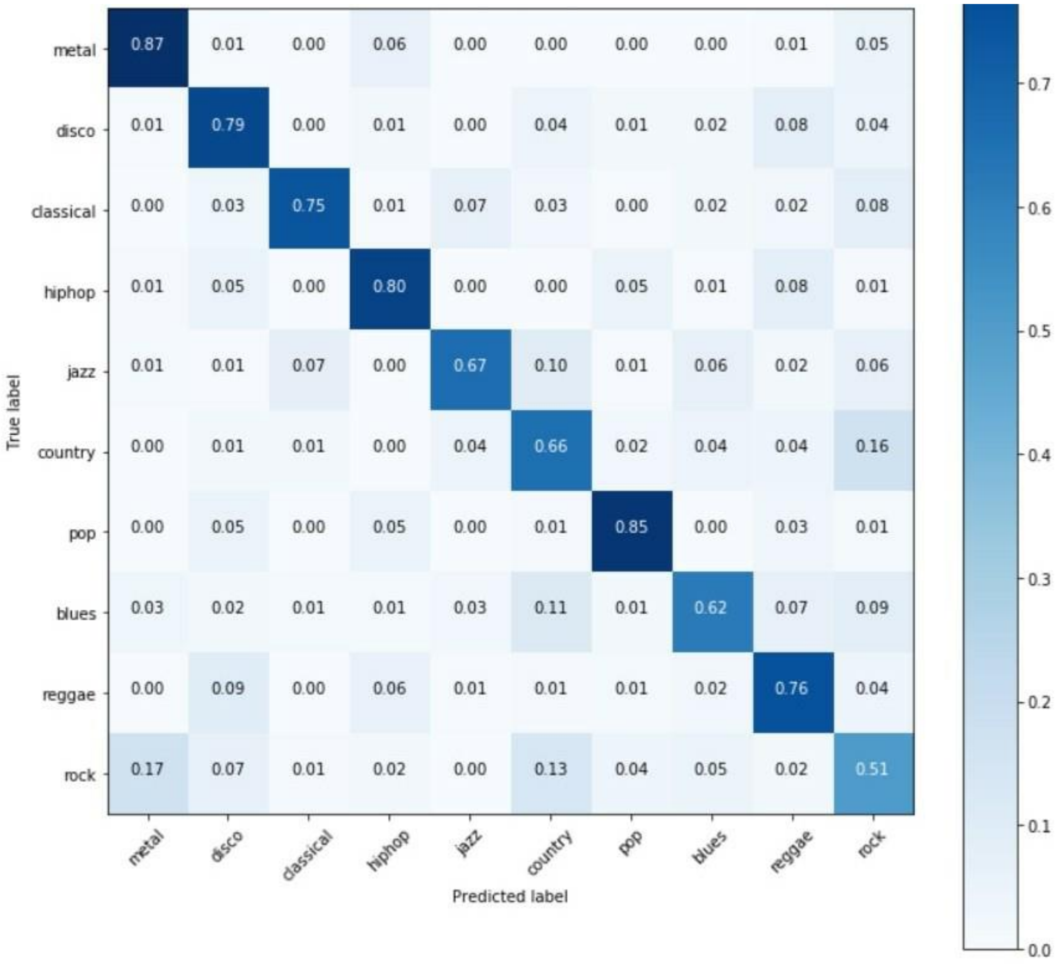

Figure 6: **The confusion matrix of the model, which speaks about the relation between the forms**

4. CONCLUSION

At the end of this research work, we are able to achieve 83% precision over our model. Model accuracy can be enhanced by adjusting some parameters by performing some hyper parameter optimization techniques and playing with the optimizer learning rate. We will continue to train our model and try to incorporate parallel pipelining with RNN and CNN so that data can be processed and trained efficiently for good performance. We may install this as a suggested application for a significant portion of the music platforms. Add on features such as sensing emotions and playing a song according to it, which can be used as a potential reach.

REFERENCES

[1] G. Tzanetakis and Cook. 2002. Musical genre classification of audio signals. Pages [1-3]

[2]https://towardsdatascience.com/using-cnns-and-rnns-for-music-genre-recognition

[3] Daniel Grzywczak and Grzegorz. 2014. Deep image features in music information retrieval. Volume 60

[4]https://www.researchgate.net/publication/324218667_Music_Genre_Classification_using_Machine_Learning_Techniques

[5]https://www.researchgate.net/publication/3333877_Musical_Genre_Classification_of_Audio_Signals

[6] Witten, I.H. Frank, E. Kaufmannm M. Data Mining: Practical machine learning tools with Java implementations. www.cs.waikato.ac.nz/˜ ML/weka/book.html, San Francisco, 2000.

[7] An Enhanced Automatic Song Genres Classification Using Deep Learning Technique, International Journal of Advanced Science and Technology Vol. 29, No. 03, (2020), pp. 5899- 5903

[8] Manaris, B. Sessions, V. Wilkinson J. “Searching for Beauty in Music-Applications of Zipf’s Law in MIDI-Encoded Music”, ISMIR 2001

[9] Boisen, S. Crystal, M. Schwartz, R. Stone, R. and Wischedel, R. “Annotating Resources for Information Extraction, LREC 2002 pp. 1211- 1214

[10] Fabbri, F. IL Suono in cui Viviamo. Arcana, Italy, 2002.

[11] J F Gemmeke, D PW Ellis, D Freedman, A Jansen, W Lawrence, R C Moore, M Plakal, and M Ritter. Audio set: An ontology and human-labeled dataset for audio events. In ICASSP, 2017.

[12] T Bertin-Mahieux, D PW Ellis, B Whitman, and P Lamere. The million song dataset. In ISMIR, 2011.

[13] A Porter, D Bogdanov, R Kaye, R Tsukanov, and X Serra. Acousticbrainz: a community platform for gathering music information obtained from audio. In ISMIR, 2015.

[14] S. Lippens, J.P Martens, T. De Mulder, G. Tzanetakis. A Comparison of Human and Automatic Musical Genre Classification. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. 1520- 6149, IEEE.

[15] Tao Li and George Tzanetakis, Factors in automatic musical genre classification, in Proc. Workshop on applications of signal processing to audio and acoustics WASPAA, New Paltz, NY, 2003, IEEE.