MUSIC GENRE CLASSIFICATION

Surya Samarth Jagadish
College of Engineering- Drexel University
3141 Chestnut St, Philadelphia,PA-19104

E-mail: sj3244@drexel.edu

ABSTRACT

What distinguishes one piece of music from another has been a subject of debate for a while in recent years. Previous methodology primarily relied on classifying a musical track as a whole into particular genres based on the type of melodies and harmonies that were being used in a song. Although it is not necessarily a flaw, classification of music based on individual portions of a complete track instead of the entire song as a whole helps to drastically improve the accuracy of categorizing the musical melodies into not just one but multiple genres which could make a significant difference. People in the music industry such as Audio Engineers, Music Composers and Producers as well as major streaming platforms such as Apple Music, Spotify, Soundcloud and YouTube music would be captivated on this idea. Contrary to the current methods in practice to classify music, the rise of interest in audio engineering has led to the birth of advanced libraries in coding platforms which help in dissecting and analyzing music with high accuracy for better utilization . By implementing the functions of dedicated python libraries, audio files are converted into Mel Spectrograms-visual representation of audio. On acquiring the spectrograms, Harmony-Percussion segregation is done using librosa functions. Boxplots are created using BPM(tempo) and harmony spectrograms are further subject to XGBoost. 100 files each having 30 seconds audio files belonging to 10 different genres acquired from the famous GTZAN dataset, the MNIST of sounds are used to train this model.

1. INTRODUCTION

In light of recent developments and active evolvement of Data Analytics in the recent years, our perceptions have altered towards every aspect in this world and has led us to question if a machine can perform the same tasks that a human being can perform. Since Machine learning models can be implied to particularly image formats of data, the audio signals as we know through concepts of physics can be typically represented in a two-dimensional visual format i.e. Frequency-Time distribution wave plots that yield amplitude of the audio signals. Further, these wave plots are converted to corresponding Spectrograms using the concept of Fourier transform or Constant-Q transform which are mathematical models that help in transforming the signals between two different domains. In simple terms Fourier transform basically helps us to take a wave signal and figure out the base frequencies that constitute that wave.

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. When applied to an audio signal, spectrograms are sometimes called sonographs, voiceprints or voicegrams. Spectrograms are widely used for multiple things in various fields such as Echocardiogram, Radiography, Sonography, Ultrasound etc. Now, the spectrograms alone are a tedious task in itself to analyze, hence we subject the obtained spectrograms to Mel scale. As the name suggests the Mel scale gets its initials from the term ‘Melody’ which is a perceptual scale of pitches judged by audio engineers to be equal in distance from one another. The reference point between this scale and normal frequency measurements is defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz, 40 dB above the listeners threshold capacity. By implementing the Mel scale, the wave plots are converted from Frequency-Time distribution to log-decibels distribution.

2. RELATED WORK

There are a few studies to directly address the music genre classification. Mingwen Dong from Rutgers University who comes from a Psychology background proposed a related work (2018) summarizing the same concept where he aims to achieve human level accuracy in music genre classification using Convolutional Neural Networks (CNN) over the exact same dataset that has been implemented in this project i.e. the famous GTZAN dataset. The approach involves training a simple Convolutional Neural Network (CNN) to classify a short segment of the music signal. Further, the genre of music is determined by splitting in into short segments and then combining CNN's predictions from all short segments. The study claims to achieve human-level (70%) accuracy. Primarily, the goal here is to achieve general overall classification of a song and assign them into particular genres.

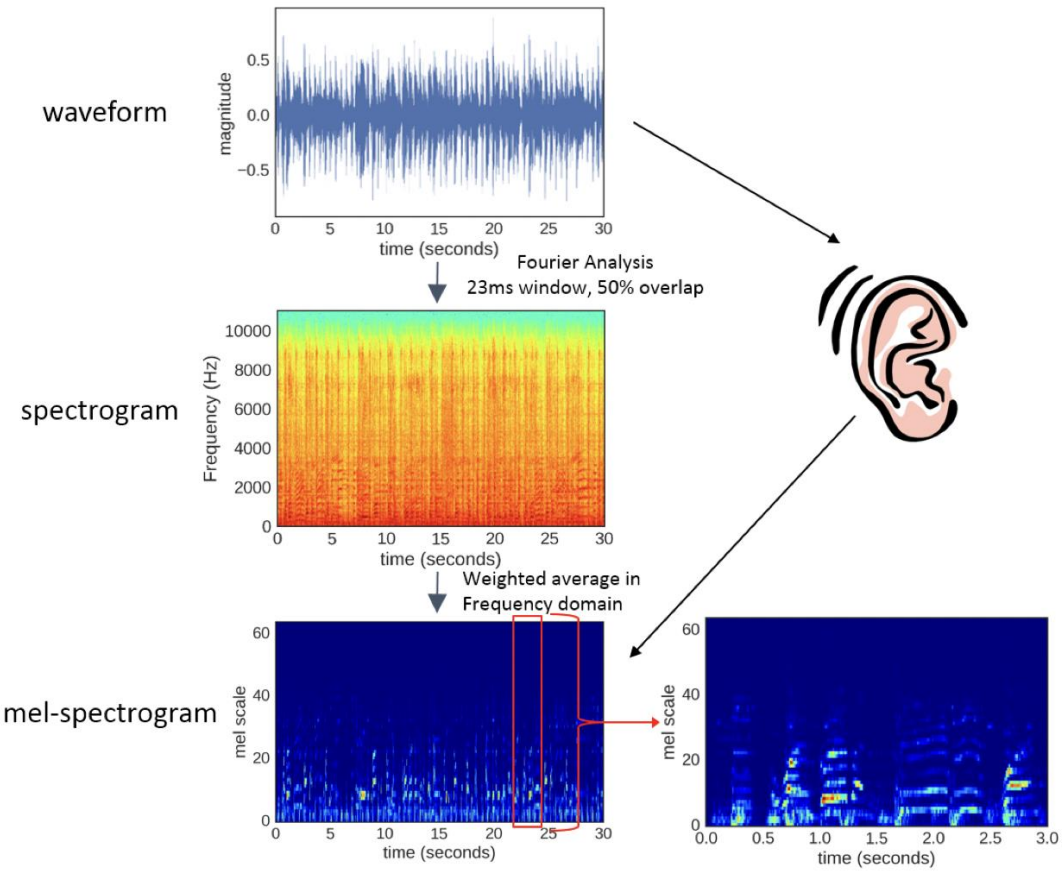

Fig 1.Image from Mingwen Dong Paper (2018) -Explains how the conversion takes place from frequency-time distribution waveplots to Mel Spectograms. Mel Spectogram mimics how human ear works.

Another study mainly focuses on how texture selection benefits music genre classification . Juliano H.Foleiss and Tiago F.Tavares (2019) introduced the concept of texture down sampling and selection. They state that, generally, the total number of sound textures per track is usually too high, and texture downsampling is necessary to make training tractable. Although previous work had solved this by linear downsampling, no extensive work had been done to evaluate how texture selection benefits genre classification in the context of the bag of frames track descriptions. The main objective was to evaluate the impact of frame selection on automatic music genre classification in a bag of frames

scenario. The results were achieved using novel texture selector based on K-means machine learning model to identify diverse sound textures within each track. Primarily, this journal also aims to solve the issue of music genre classification but takes a different approach involving full length tracks instead of 30 second samples from the GTZAN dataset. Ultimately, the goal varies as to classifying the entire song as a whole versus classifying individual sections of the song.

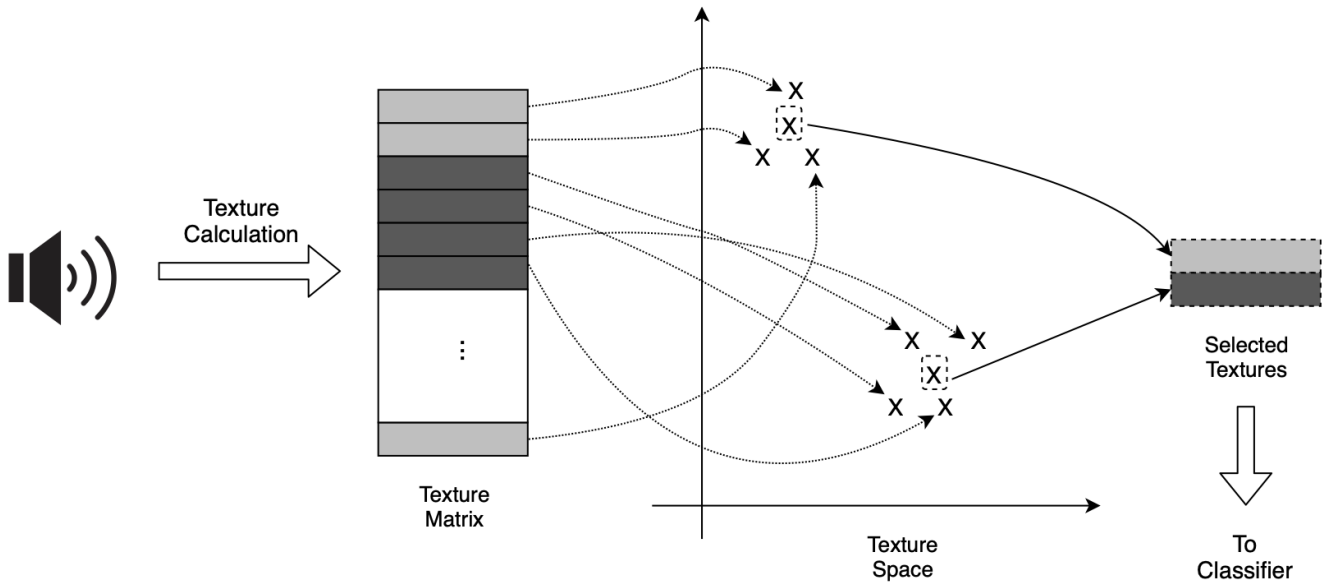

Fig 2. Image from Juliano H.Foleiss and Tiago F.Tavares paper (2019)- Explains the K-Means Texture Selection from full length tracks.

There exists other proposed methods as well among which one particular study stood out. Sergio Oramas, Oriol Nieto, Francesco Barbieri, Xavier Serra (2017) brought in the concept of classification based on common characteristics. Their work aims to expand the task by categorizing musical items into multiple and fine-grained labels, using three different modalities namely audio, text and images. The dataset used is MuMu, a new dataset consisting of more than 31000 albums classified into 250 genre classes. For every album, the related cover image, text reviews and audio tracks have been extracted as well. The classification takes place based on the combination of feature embeddings via deep learning methodologies. In theory, the concept is logically true i.e. it is possible to find certain trends when it comes to identifying music. This not only implies to the attributes of the music itself, but it also pertains to other aspects such as album art, the type of lyrics used, text reviews and various other features which act as a significant contribution towards identifying a particular genre.

Various other methods have been proposed for classification of music into particular genres. Some other notable mentions are- Codified language Modelling (Rodrigo Castellon, Chris Donahue, Percy Liang- 2021), Bottom-up Broadcast Neural Network for Music genre Classification (Caifeng Liu, Lin Feng, Guochao Liu, Huibing Wang, Shenglan Liu-2019), Music Genre Classification with Paralleling Recurrent CNN (Lin Feng, Shenlan Liu, Jianing Ya- 2017).

The most relevant works include the first three works mentioned above in detail which inspired this particular project which deals with quite a similar approach but attempts to solve a different issue. The concept I am particularly interested in mainly deals with identifying a song as a part of multiple genres and not just one.

❖ HOW IS MY WORK DIFFERENT?

A common experience most music enthusiasts face, i.e., there exist songs where in the intro tends to sound entirely different from the chorus. There exist exquisite pieces of music which include a mellow beginning or a soft intro with minimal instruments or acoustics, synth or almost no instruments at all and transition into a high voltage chorus such as rock or include various different instruments and a completely different beat pattern which brings us to our interests, i.e to classify a song that includes traces of different genres among different sections such as intro, verse, bridge, pre-chorus, chorus and outro of the song.

As mentioned above, the approach used to get to a certain extent in the project was similar to other important mentions above. The initial process includes converting audio songs which are in .wav format to Wave plots representing the sound waves in the form of frequency-time distribution. The next step includes converting the wave plots to Spectrograms. Now for betterment of accuracy, the spectrograms are plotted again with respect to logarithmic values which give a distinct difference between the harmonies and intensities of a particular piece of music. We then use a dedicated Python library known as Librosa.

LIBROSA- Created by Brian McFee- Assistant Professor of Music Technology and Data Science at NYU Steinhardt and the NYU center for Data Science. It is a recently developed package for python, MATLAB which particularly helps for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. It helps to perform many tasks over audio files such as audio segregation, tempo detection, plotting spectrograms, intensity analysis and various other functions.

Furthermore, we step into the next aspect of the project wherein we make use of a rather unique Machine learning Model known as XGBoost in order to get the best accuracy for our result unlike the relevant works mentioned above where mostly CNN is being used.

XGBOOST- In basic ML terms, there exists two concepts known as Bagging and Boosting. More specifically in Boosting technique, there exists something known as XGBoost which stands for Extreme Gradient Boosting which is a scalable, distributed gradient-boosting decision tree (GBDT) Machine Learning library. It provides paralell tree boosting where in it computes both linear model as well as tree algorithms and is a leading library for regression, classification, and ranking problems.

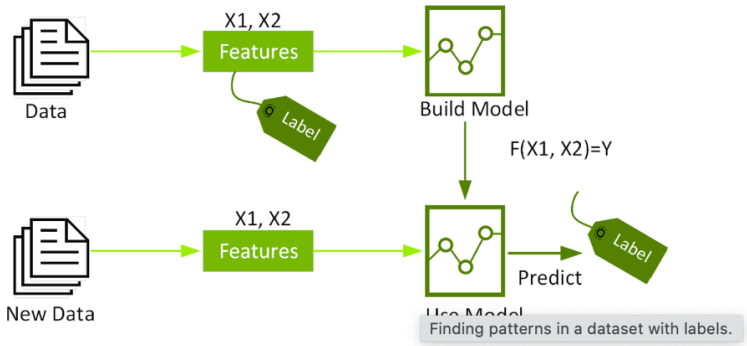

Fig 3. Image obtained from reference[4]- shows the procedure of XGBoost

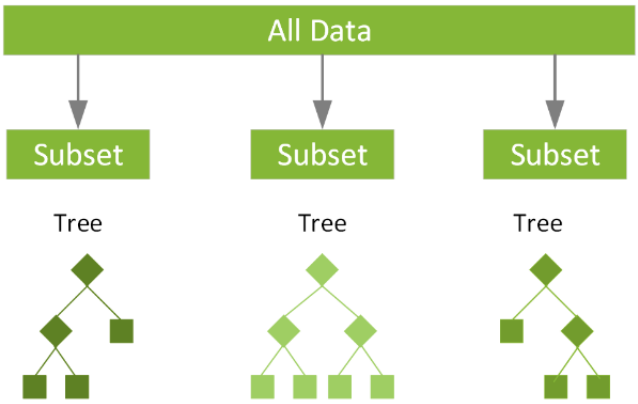

Fig 4. Image obtained from reference[4]- shows the procedure of XGBoost

Moving on to the next step which particularly holds as unique to this project i.e. separation of Harmony from percussions i.e Drum beats, Electronic beats which includes different types of Kicks, Toms, Snares, Hi-Hats, Crash etc. This is achieved through Librosa which includes something known as Spectral features that help to analyze timbre, harmonies which are usually the product of spectrogram and a filter bank. This leads us to the next step i.e. to separate harmony from percussions and give us different audio file representation. Using the 'tempo' command of Librosa we can obtain the BPM (Beats Per Minute) scales of different beats and create relevant boxplot of the type of beats. Now it is important to note that percussions are relatively easier to analyze since they are recursive in the exact same pattern and are mostly consistent throughout the song, whereas the Harmonies vary with intensity of instruments that are being used. Hence in order to get a better output to perform our analysis we use a feature known as Chroma feature from the Librosa library. The Chroma feature basically helps to measure the different energies in each pitch class and sort them into 12 different tones. Basically, the Chroma feature helps to compute a Chromogram from a waveform or Power Spectrogram.

DATASET- The Data gathered consists of a collection of 10 different genres including 100 files each in every genre. The audio files are categorized into 2 csv files with one having 30 seconds patches and another having 3 second patches (mainly to maximize the data input). This dataset is obtained from the famous GTZAN dataset, the MNIST of sounds used for MGR (Music Genre Recognition). The audio data is obtained in the form of .wav files as seen in the figures below.

features 30 seconds: a csv with a mean and average value of different extracted features from the audio files

filename	length	chroma_stft_mean	chroma_stft_var	rms_mea
1000 unique values	660k 676k	0.17 0.66	0.04 0.11	0.01
blues.00000.wav	661794	0.3500881195068359 4	0.0887565687298774 7	0.1302279
blues.00001.wav	661794	0.3409135937690735	0.0949802547693252 6	0.0959478
blues.00002.wav	661794	0.3636371791362762 5	0.0852751955389976 5	0.1755704
blues.00003.wav	661794	0.4047847092151642	0.0939990356564521 8	0.1410938

rms_mean	rms_var	spectral_centroid...	spectral_centroid...	spectra
0.01 0.4	0 0.03	570 4.44k	7.91k 3.04m	898
0.1302279233932495	0.0028266964945942163	1784.165849538755	129774.06452515082	2002.449
0.09594780951738358	0.002372739138081 67	1530.1766787460795	375850.07364868594	2039.036
0.17557041347026825	0.0027459163684397936	1552.8118647610036	156467.64336831577	1747.702
0.141093000769615 17	0.0063463463447988 03	1070.1066149971282	184355.94241695415	1596.412

Fig 5,6- Dataset obtained from Reference[]- Gives the different mean, variance and other values for different genres

	# spectral_centroid...	# spectral_centroid...	# spectral_bandwid...	# spectral_bandwid...
	0.03 570 4.44k	7.91k 3.04m	898 3.51k	10.8k 695k
4594216	1784.165849538755	129774.06452515082	2002.4490601176963	85882.76131549841
808167	1530.1766787460795	375850.07364868594	2039.0365160924302	213843.75549673408
8439793	1552.8118647610036	156467.64336831577	1747.702311667602	76254.19225682177
4798803	1070.1066149971282	184355.94241695415	1596.4128721770246	166441.49476892402

Fig 7. Dataset obtained from Reference[] - Gives the different mean, variance and other values for different genres

3. EXPERIMENT

The main procedure includes, taking the input of a particular song and breaking it down into patches of 30 seconds or lesser and processing it to obtain Mel Spectrograms. Once the harmonies are separated, the analysis is done where in the obtained patches of data is compared with the existing dataset of different musical pieces. On recognizing the type of genre of one particular patch the system iterates onto the next set of patches of the song. This procedure continues until the entire song has been analyzed . Once the entire song has been compared to the existing dataset, the final result consisting of different genres is taken into consideration and repetitive genres are concluded as one single entity. The final output gives us a result of all possible genres that the individual sections of the song was resembling based on the training.

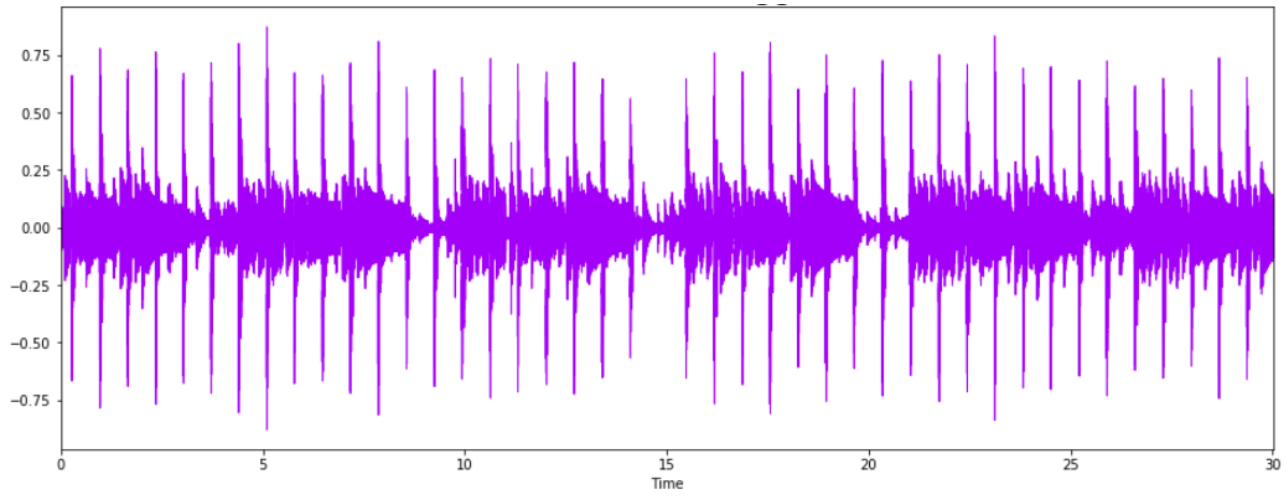

Fig 8. Image compiled by self- Wave plot of Hip-Hop

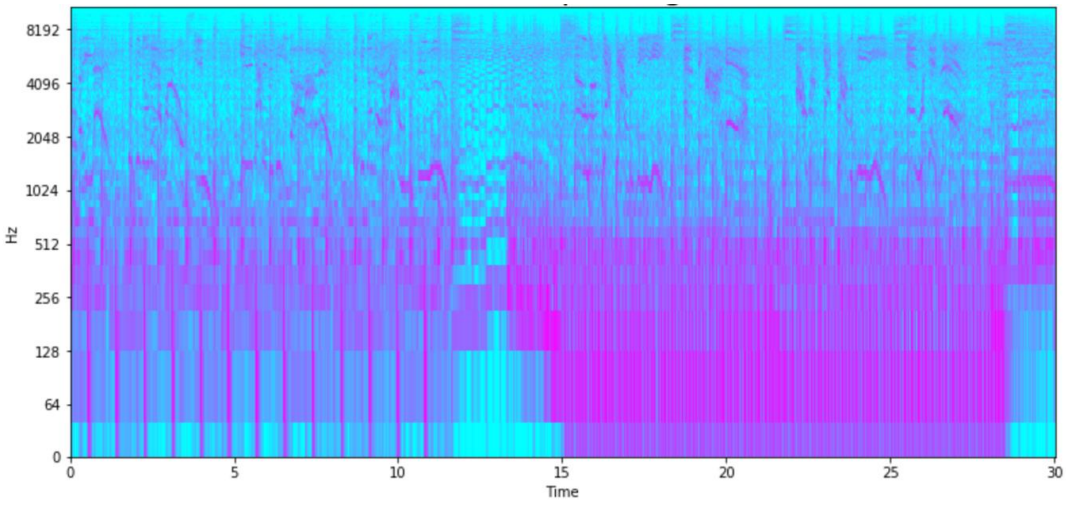

Fig 9. Image compiled by self- Mel Spectrogram example of Rock. It can be noticed that the intense areas are in purple while blue areas represent the noise/stagnant region.

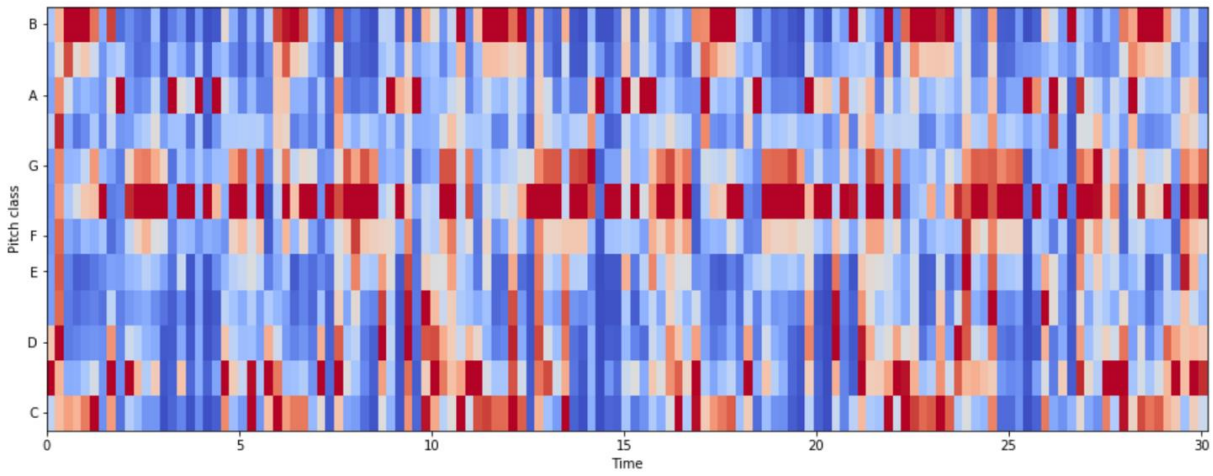

Fig 10. Image compiled by self- Example of how Chroma feature looks after 12 tone splitting

4. CONCLUSION

In this study, we addressed a particular aspect related to Music Genre Classification which cannot be particularly addressed as a flaw but can definitely make significant difference in the way music can be classified. Results acquired were comparatively lower due certain issues faced with it comes to the vagueness of music. Better results could be acquired using the concept utilized by one of the papers mentioned in the relevant work section (Sergio Oramas, Oriol Nieto, Francesco Barbieri, Xavier Serra (2017)) which could enhance the classification and reduce the vagueness by making use of multi-models apart from audio alone namely text and album art/ cover image. The model nevertheless acquired the skill to compute and categorize the patches in multiple genres thereby doing justice to the core intent of this project. This concept would be insightful for audio engineers since they create and update music softwares such as FL Studio, Ableton etc where the music library keeps getting updated with different samples of music being added to the library on a regular basis. This concept would also improve the music suggestion aspect in mainstream platforms such as Spotify and Apple music based on multiple categories a song might belong to. YouTube could essentially

suggest video content based on the particular categories a video song might belong to. In terms of future work, this project holds good potential if developed using deeper knowledge.

5. REFERENCES

[1] Dong, Mingwen. "Convolutional Neural Network Achieves Human-Level Accuracy in Music Genre Classification." *ArXiv.org*, 27 Feb. 2018, https://arxiv.org/abs/1802.09697v1.

[2] Foleiss, Juliano H., and Tiago F. Tavares. "Texture Selection for Automatic Music Genre Classification." *ArXiv.org*, 28 May 2019, https://arxiv.org/abs/1905.11959v1.

[3] Oramas, S. *et al.* (2017) *Multi-label music genre classification from audio, text, and images using deep features*, *arXiv.org*. Available at: https://arxiv.org/abs/1707.04916v1 (Accessed: December 6, 2022).

[4] *What is XGBoost?* (no date) *NVIDIA Data Science Glossary*. Available at: https://www.nvidia.com/en-us/glossary/data-science/xgboost/ (Accessed: December 6, 2022).

[5] (no date) *Proceedings of the python in science conferences*. Available at: https://conference.scipy.org/proceedings/ (Accessed: December 6, 2022).

[6] EnthoughtMedia (2015) *Librosa audio and Music Signal Analysis in Python | scipy 2015 | Brian McFee*, *YouTube*. YouTube. Available at: https://www.youtube.com/watch?v=MhOdbtPhbLU (Accessed: December 6, 2022).

[7] *Music genre classification* (no date) *Papers With Code*. Available at: https://paperswithcode.com/task/music-genre-classification (Accessed: December 6, 2022).

[8] Olteanu, A. (2020) *GTZAN dataset - music genre classification*, *Kaggle*. Available at: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification (Accessed: December 6, 2022).