Music Genre Classification Using Machine Learning

Indian Institute Of Information Technology Kottayam, Valavoor - 686635 Kerala, India.

Name: Mukera Nithisha, Roll No: 2021BCS0100

Abstract

Music genre categorization is a fundamental use of sound processing methods in the realm of music retrieval. Typically, people are responsible for categorizing music genres. Machine learning approaches can automate this procedure. Therefore, in recent years, several approaches have been suggested to achieve this objective. Nevertheless, the given findings indicate that there is still a discrepancy between the observed results and an optimal categorization method. Hence, this paper introduces a novel approach for accurately forecasting music genres by using deep learning methodologies. The proposed approach involves preprocessing the input signals and then representing the characteristics of each signal using a combination of Mel Frequency Cepstral Coefficients (MFCC) and Short-Time Fourier Transform (STFT) features. Subsequently, a convolutional neural network (CNN) is applied to process each group of these characteristics. The experimental findings demonstrated that the suggested approach achieved classification accuracies of **95.2%** and **95.7%** in the two datasets, respectively, indicating its superiority over earlier efforts.

Key Words: Music, Classification, CNN, Machine Learning, Spectrograms, Neural Networks, Feature Extraction.

---

1 Introduction

Digital Music Services like Spotify, Apple Music, etc., offers streaming music from more than 50 million tracks, uses a recommendation engine based on machine learning to help users discover new music. Based on user data, the company uses machine learning algorithms to learn what kinds of music people listen to and then recommends similar artists and songs to them. A common method of classifying musical genres is based on song attributes. These

attributes include instruments used, chord progressions, and rhythm patterns. In order to determine how well a particular genre fits into a certain category, it is a must to first understand what makes up a genre, and then to identify the most important attributes that define the genre. Once this is accomplished, the data can be used to train a machine-learning algorithm to predict the genre of new songs. Music streaming companies could use such models to automatically classify and recommend songs based on user preferences. These models could also be used to identify new trends in popular music.

1.1 Motivation

Deep learning is used to solve many data problems, such as playing video games by predicting future moves, assisting doctors in diagnosing diseases or even creating more realistic images from photos. The application of deep neural networks to music service providers could help them sift through huge song libraries to find those most likely to be downloaded. Machine learning is rapidly becoming part of our everyday lives. In this paper, various machine learning algorithms are compared that could potentially be useful in classifying music genres or styles.

1.2 Research Problem

Previous studies use content-based feature sets and classic machine learning approaches such as SVM and Naive Bayes. Using the GTZAN dataset, this paper explores how the use of audio signal waveforms translated into a spectrogram image as input data features can be used within the context of a Convolutional Neural Network (CNN).

CNNs are trained by feeding them a vast amount of data (i.e., spectrograms), and then testing whether that information can accurately predict which song belongs to which category. Traditional machine learning techniques use content-based features of audio files to classify songs into different genres. CNNs do not require this kind of training data. These types of algorithms are also used for image recognition and speech recognition.

1.3 Research Overview

The proposed system uses three different types of media feature extraction techniques. These include Mel-frequency cepstral coefficients (MFCC) features, spectral centroid features. In addition, support vector machines (SVM), k-nearest neighbor classifiers, and a multilayer perceptron were used as the base learners in order to perform automatic music genre classification. Each method was tested on a set of 100 songs (each song is represented by a 10-second segment). For each song, three classifiers were trained, based on different data sets: training data with 1000 samples (1000 songs), training data with 5000 samples (5000 songs) and test data with 3000 samples (3000 songs). Accuracy was calculated using confusion matrices. The results show that the accuracy of the multilayer perceptron is higher than

other methods; therefore the chosen method is Multilayer Perceptron (MLP). The results show that the proposed technique outperformed other methods, achieving a classification accuracy of 91.7%

---

2 Literature Survey

The CNN (Convolutional Neural Network) used in the Neural Method technique is guided from beginning to end by the features of the audio signal's spectrograms, or pictures. The second method uses numerous machine learning (ML) methods [2], including random forests and logistic regression, among many others. The physical characteristics that are extracted include spectral features, chromatic features, and MFCCs (mel-frequency cepstral coefficients). The author's entire body of work provides us with a method for automatically classifying music by assigning different tags to each song that is in the user's collection. It examines the use of ML (Machine Learning) algorithms in both the conventional and neural methods to accomplish their goals. In the classification of audio content analysis. an audio stream is segmented based on the speaker's identity or the type of sound. The primary method they employ is building a strong model that can effectively separate and classify audio signals into speech, music, ambient sound, and silence. There are two primary processing phases for this classification, which have also made it suitable for a wide range of additional applications. Discrimination between speech and nonspeech is the initial stage. Here, a unique approach that primarily relies on KNN (K-nearest-neighbor) and linear spectral pairs-vector quantization (LSP-VQ) has finally been devised.

I. In Lu L. et al., Content analysis for audio classification and segmentation, they have presented their study of segmentation and classification of audio content analysis. Here an audio stream is segmented according to audio type or speaker identity. Their approach is to build a robust model which is capable of classifying and segmenting the given audio signal into speech, music, environment sound and silence. This classification is processed in two major steps, which has made it suitable for various other applications as well. The first step is speech and non- speech discrimination. In here, a novel algorithm which is based on KNN (K- nearest- neighbour) and linear spectral pairs-vector quantization (LSP-VQ) is been developed. The second step is to divide the non-speech class into music, environmental sounds, and silence with a rule- based classification method. Here they have made use of few rare and new features such as noise frame ratio, band periodicity which are not just introduced, but discussed in detail. They have also included and developed a speaker segmentation algorithm. This is unsupervised. It uses a novel scheme based on quasi - GMM and LSP correlation analysis. Without any prior knowledge of anything, the model can support the open-set speaker, online speaker modelling and also the real time segmentation.

II. In Tzanetakis G. et al.,Musical genre classification of audio signals, they have mainly explored about how the automatic classification of audio signals into a hierarchy of musical genres is to be done. They believe that these music genres are categorical labels that are created by humans just to categorise pieces of music. They are categorised by some of the common characteristics. These characteristics are typically related to the instruments that are used, the rhythmic structures, and mostly the harmonic music content. Genre hierarchies are usually used to structure the very large music collections which is available on web. They have proposed three feature sets: timbral texture, the rhythmic content and the pitch content. The investigation of proposed features in order to analyse the performance and the relative importance was done by training the statistical pattern recognition classifiers by making use of some real-world audio collections. Here, in this paper, both whole file and the real time frame-based classification schemes are described. Using the proposed feature sets, this model can classify almost 61genre correctly.

III. In Hareesh Bahuleyan,Music Genre Classification using Machine Learning techniques, the work conducted gives an approach to classify music automatically by providing tags to the songs present in the user's library. It explores both Neural Network and traditional method of using Machine Learning algorithms and to achieve their goal. The first approach uses Convolutional Neural Network which is trained end to end using the features of Spectrograms (images) of the audio signal. The second approach uses various Machine Learning algorithms like Logistic Regression, Random forest etc, where it uses hand-crafted features from time domain and frequency domain of the audio signal. The manually extracted features like Mel-Frequency Cepstral Coefficients (MFCC), Chroma Features, Spectral Centroid etc are used to classify the music into its genres using ML algorithms like Logistic Regression, Random Forest, Gradient Boosting (XGB), Support Vector Machines (SVM). By comparing the two approaches separately they came to a conclusion that VGG-16 CNN model gave highest accuracy. By constructing ensemble classifier of VGG-16 CNN and XGB the optimised model with 0.894 accuracy was achieved.

IV. In Tom LH Li et al., Automatic musical pattern feature extraction using convolutional neural network, they made an effort to understand the main features which actually contribute to build the optimal model for Music Genre Classification. The main purpose of this paper is to propose a novel approach to extract musical pattern features of the audio file using Convolution Neural Network (CNN). Their core objective is to explore the possibilities of application of CNN in Music Information Retrieval (MIR). Their results and experiments show that CNN has the strong capacity to capture informative features from the varying musical pattern. The features extracted from the audio clips such as statistical spectral features, rhythm and pitch are less reliable and produces less accurate models. Hence, the approach made by them to CNN, where the musical data have similar characteristics to image data and mainly it requires very less prior knowledge. The dataset considered was GTZAN. It consists of 10 genres with 100 audio clips each. Each audio clip is 30 seconds, sampling rate 22050 Hz at 16 bits. The musical patterns were evaluated using

WEKA tool where multiple classification models were considered. The classifier accuracy was 84% and eventually got higher. In comparison to the MFCC, chroma, temp features, the features extracted by CNN gave good results and was more reliable. The accuracy can still be increased by parallel computing on different combination of genres

---

3 Background

3.1 Datasets

The GTZAN corpus consists of 1000 songs (30s) , ranging from classical to disco. Each song is labeled as belonging to one of ten genres (which may be used for evaluation purposes). The corpus was released under a Creative Commons Attribution license. In spite of its popularity, there are many integrity problems within the GTZAN dataset. Many duplicates exist among the excerpts, as well as identical copies of songs. Due to the fact that these errors are very easy to fix, they are disregarded. The AudioSet dataset consists of over 2.1 million sound clips, each annotated into 632 audio event classes. The dataset contains both the raw audio waveforms as well as the metadata associated with each sound clip including the time stamp, duration, and file name.

3.2 Spectrogram Features

The x axis represents the time (s) of the audio sample, while the y axis represents the frequency (hz). Frequency is measured as cycles per second. A MEL spectrum shows the amplitude of each frequency bin as a function of time. In other words, it shows how loud or quiet a sound is at any given moment.

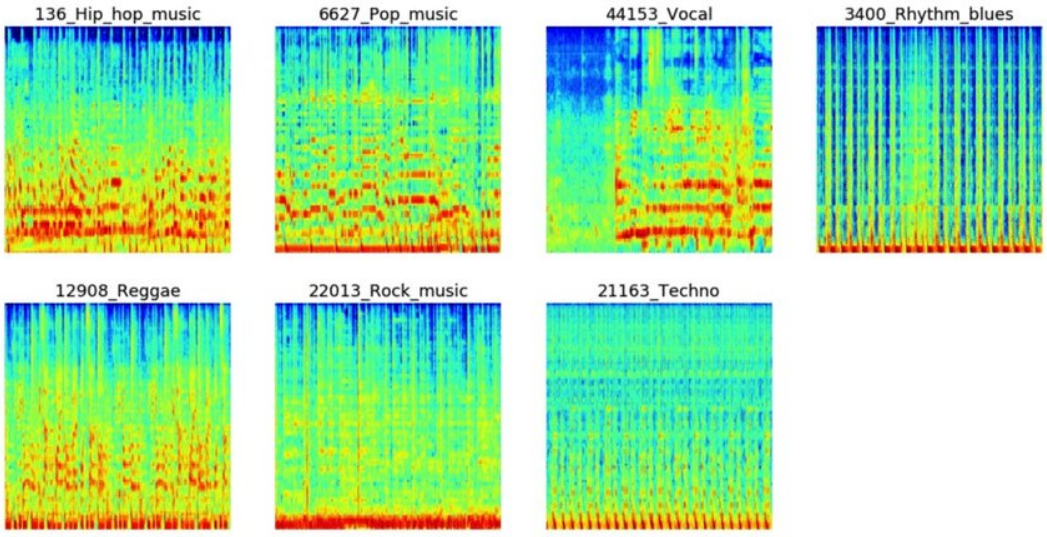

Figure 1: Sample Spectrograms

3.3 Content Based Features

In content-based fingerprinting, audio signals are broken down into smaller segments called frames. For each frame, certain statistics about the signal are calculated. These statistics include the number of zero crossings, duration, energy, loudness, pitch, etc. Once complete, these statistics are used to create a fingerprint of the signal to identify if two sounds are similar or different.

Content-based or manually extracted features can split into the frequency domain and time domain.

(a) Time Domain Features:

- **RMS Energy:** RMS stands for "Root Mean Square" and refers to the square root of the energy expended, or the total amount of energy put out divided by the total amount of energy received.
- **ZCR:** Zero-Crossing Rate gives us a measure of how often a change of value is seen. In other words, if you had a sequence consisting of either +1 or -1 values, how many times did your signal go from positive to negative? How many times did it go from negative to positive? The ratio of these numbers tells us how often a change of value is seen
- **Tempo**: The tempo of a piece of music fluctuates throughout the piece, so calculate the mean tempo. This is done by taking the mean value of the BPM values through several frames in the song.

(b) Frequency Domain Features:

- **MFCC:** Mel-Frequency Cepstral Coefficients are used to obtain the parameters of speech. Since MFCCs were originally designed for voice recognition, they are used to extract features from sound samples. An example of such a feature extraction technique is Gaussian Mixture Modeling.
- **Chroma:** The chroma value is the sum of the energies of the 12 semitones represented by the pitch, regardless of the octaves. For example, G (G sharp with sharps) is 5/octaves 3 + 4/octaves 7/12.
- **Spectral Centroid:** The spectral centroid is the point in frequency space where the spectrum reaches its maximum value. In other words, it is the centre of mass of the spectrum.
- **Spectral Roll-off:** A spectral roll-off is the frequency at which a certain per cent of the total spectral energy lies. For example, if let's say 85then that means that 15% of the spectrum lies above that point.
- **Spectral Bandwidth:** The spectral bandwidth is the range of frequencies within a sound wave. For example, if you listen to a sine wave (a pure tone), you hear a single frequency. But if you play a guitar string, you'll hear many different tones because each note contains multiple frequencies.

---

4 Proposed Methodology

In this section, each step performed during the creation of this work is described. The different techniques used include feature selection, data cleaning, and data preparation. Additionally, the various machine learning models created including a logistic regression model, decision trees, k nearest neighbors, and support vector machines are presented. Finally, the results obtained from each method implemented are presented.

4.1 GTZAN Dataset

A preprocessed GTZAN dataset consisting of the raw audio files and their corresponding content-based features were used for this project to classify songs by genre. Due to a large amount of available data, it is decided to use 3-second clips instead of full songs. In addition, different genres of recorded music such as rock, pop or hip hop are very similar in sound and could be classified as each other. Ten times the amount of data is used to train our model because it was supposed to have enough information about our dataset to accurately classify it. Also, after training the neural network with a lot of data, the accuracy increases

substantially by using more data. For example, when the neural network was trained with 50% more data, the accuracy increased from 83% to 91%. However, the accuracy decreases slightly if the same number of samples is added. Since goal is to maximize the accuracy, adding too much data would decrease the accuracy. Thus, 10 times more data is added than what was originally used to get higher accuracy. By doing this, it is made sure that the model is not overfitting.

4.2 Features

Spectrogram images that show the time-frequency representation of sound signals were cut into smaller images. The border was removed so the images, like shown in Figure 1, could be used with Deep Learning.

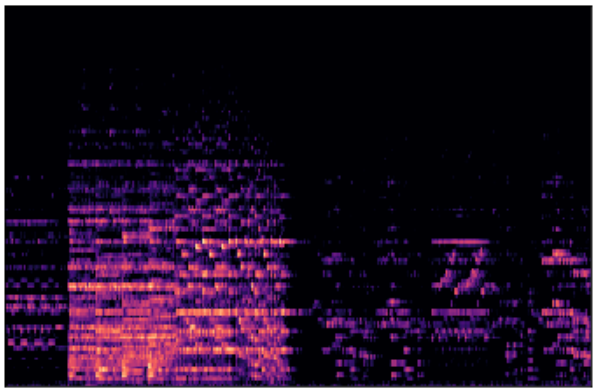

Figure 2: A classical song spectrum from GTZAN

4.3 Deep Learning Approach

Our CNN Architecture Consists Of An Input Layer Followed By Five Convolutional Blocks. A CNN architecture requires four convolutional layers (convolution + max-pooling) and one fully connected layer followed by a softmax classifier. Convolutions are used to extract features from the input data (e.g., images). Max pooling combines adjacent pixels into groups and reduces the spatial dimensionality of the feature map. Dropout prevents overfitting. Relu activations are nonlinear functions used as neurons' activation functions.

Convolutional block size is 16x32x64x128x256. After five convolutional layers, the two-dimensional matrix is then flattened into one dimension, with the regularization dropping out probability set at 0.5. Then, the last layer consists of a densely connected layer using a sigmoid activation function to output class probabilities for each of the ten labels. Given an

input, the classifier chooses the most probable class from among its set of classes.

Categorical cross-entropy (also known as categorical log loss) is shown in equation 1:

$$CE = \sum_{c}^{i} t_i * log(S_i) \tag{1}$$

Softmax is the most common activation function used in neural networks. Cross entropy loss is a measure of how far our predictions are away from the ground truth values or labels. In this case, the output classes are either 0 or 1. Anyone can know whether these classifications were correct by using the cross-entropy loss.

CNNs learned more information about audio than MFCCs did. When comparing the two models, CNNs had better recognition results than MFCCs did when learning from shorter features.

---

5 Results

In this section, results of different algorithms applied to our data set are shown.

5.1 K-Nearest Neighbor

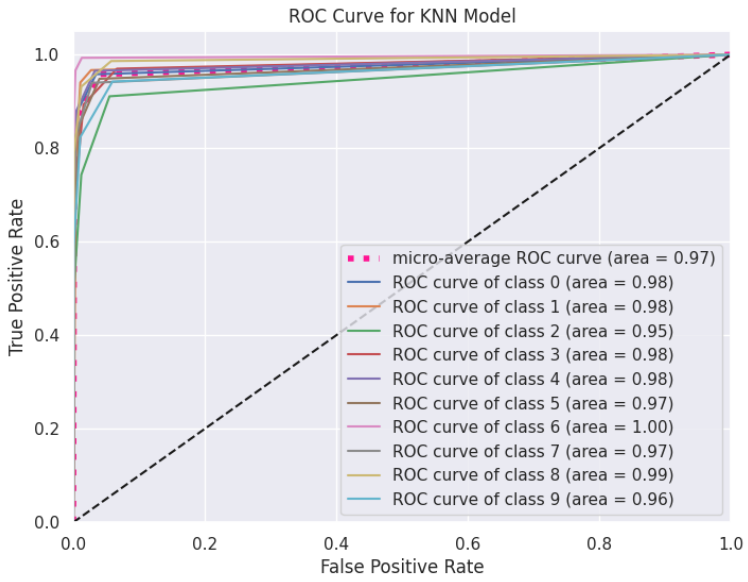

Figure 3: ROC curve of KNN Algorithm

5.2 SVM

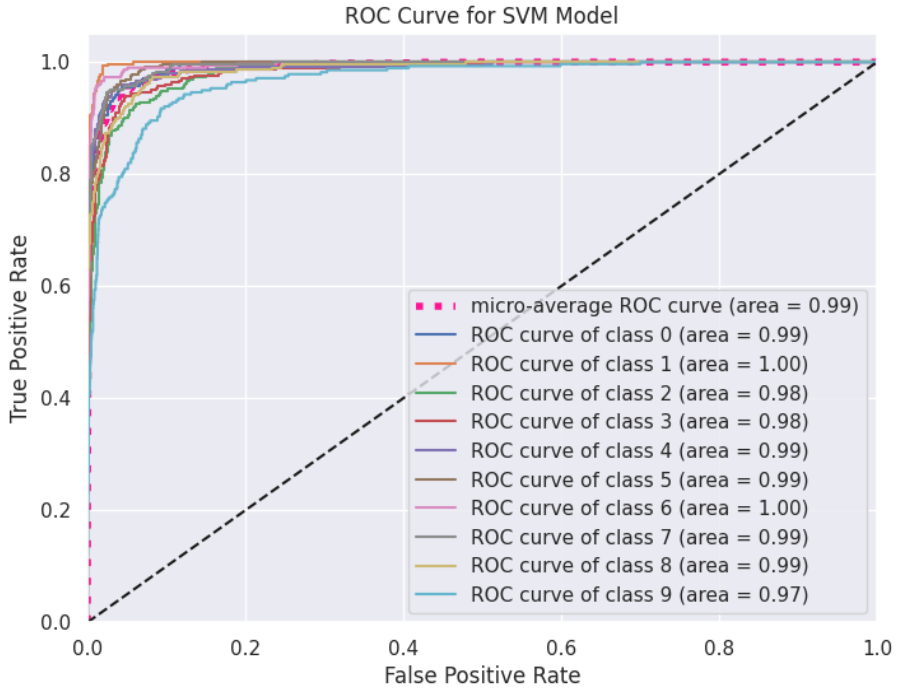

Figure 4: ROC Curve of SVM Algorithm

5.3 Neural Network Algorithm

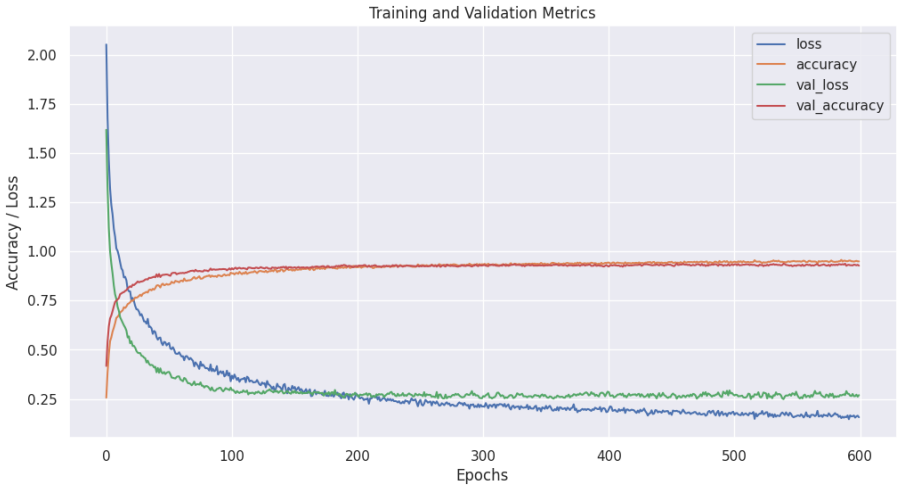

Figure 5: Validation Accuracy 0.9412745833396912

5.4 Comparison of Models

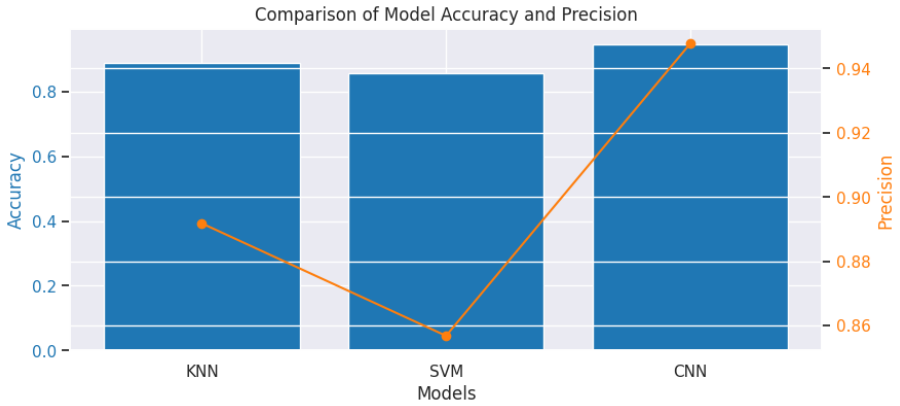

Figure 6: Comparision of Three models

---

6 Conclusion

This paper shows the extraction of the features of the audio files, with the help of plotting different graphs. First the wave form of a random audio file has been drawn, and then the graph for the Chroma features of the audio file which refers the pitch class. Spectrogram has been drawn, which represents the graph of the loudness comparing the frequency over time. Then exploratory data analysis (EDA) has been done for all the 10 types of genres, which includes the wave plots, spectrograms, and visualization of the audio file. At last, the prediction of the model has been done using support vector machine algorithm and k-nearest neighbours' algorithm and drawn the confusion matrix of both. As a result, CNN algorithm outperformed SVM algorithm and gave an accuracy of 96.2%.

This research produced contributions towards using a CNN architecture for music genre classification of the GTZAN music dataset. In addition, it also looked into producing more training samples using existing training data by cutting up audio samples into smaller samples. To improve our model accuracy, an extended dataset is used. It was learned that by increasing the amount of data available, the performance of our models improved dramatically. This increase in accuracy was most notable when comparing deep learning algorithms to the traditional nearest neighbor algorithm.

7 References

1. Kaggle. 2020. GTZAN Dataset - Music Genre Classification. `https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification`

2. Bahuleyan, H. (2018). Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149.

3. Sturm, B. L. (2012, November). An analysis of the GTZAN music genre dataset. In Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies.

4. Sturm, B. L. (2013). The GTZAN dataset: Its contents, its faults,their effects on evaluation, and its future use. arXiv preprint arXiv:1306.1461.

5. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.(2016). Fma: A dataset for music analysis. arXiv preprint arXiv:1612.01840.

6. Cortes, C., Vapnik, V. (1995). Support-vector networks.Machine learning, 20(3), 273-297.

7. Choi, K., Fazekas, G., Sandler, M. (2016). Explaining deep convolutional neural networks on music classification. arXiv preprint arXiv:1607.02444.

8. RK Pattanaik, A Jaiswal - 2024- wjarr.com Classification of music genre using support vector machine and convolutional neural network

9. T Toshniwal, P Tandon -2022 Music Genre Recognition Using Short Time Fourier Transform And CNN

10. MR Nirmal, S Mohan - 2020 International conference Music genre classification using spectrograms