Music Genre Classification

Project Report
IIITD

Dheeraj (2017044) **Rameshwar Mishra (2017180)** **Sachin Nandal (2017185)**

Code files: Drive

Abstract

Music Genre classification is essential in today's world due to the rapid growth in music tracks. Ease in sharing data over the internet and other modes motivates many new tracks every day; with such a vast collection of songs, the need to group them into few semantic categories arises. To have better access to these, we need to index them accordingly. In the past, it has been seen that a few characteristics of music affect the way users perceive music, and those can be used to categorize them into different human-made classes called genres. In this project, we have implemented a system to classify the music genre from an audio file. We will also analyze and compare a few Machine learning technique's performances to get the best model for the genre classification task.

1 Introduction

Downloading, sharing, purchasing music is part of daily life now, and it motivates the addition of thousands of songs every month. Users tend to use genres like classical, pop, jazz to formulate their preferences. Music genres are categorical labels invented by humans to group or characterize a part of the music—members of a particular genre share some standard features that distinguish them. The characterization generally happens based on rhythmic structure, frequency analysis, instrumentation, and the audio signal's harmonic content. Given the massive size of the audio collection, genre classification plays a vital role in the search, retrieval, and music recommendation.

We have implemented and compared the performance of two different classes of machine learning models. First is convention linear classifiers KNN and SVM with rbf kernel, these models will be trained on the features extracted from audio. Second is deep learning-based CNN model which will be trained end to end. For CNN input will be MEL-spectrogram of audio signal. The challenging part of any audio-related classification task is to identify the features to consider for learning. We will consider standard audio features used to characterize an audio signal like Mel Spectrograms, Chroma Frequencies, Mel-Frequency Cepstral Coefficients (MFCCs), Spectral centroid, Zero crossing.

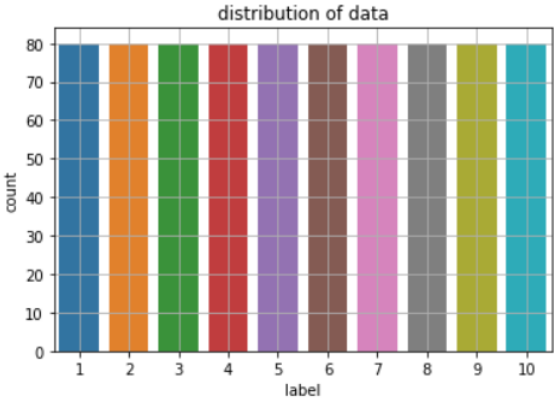
2 Related Work

Many researchers have experimented different techniques and feature extraction rules over the years to classify music into different genres effectively. The most common approach is to use audio signal processing concepts for features extraction with powerful machine learning techniques to build a stable and effective classifier. [1] In the early stage timbral texture, rhythmic and pitch content were used as features with k-nearest neighbors to achieve an accuracy of 61%. Tzanetakis and Cook were one of the first peoples who contributed to this field by creating GTZAN dataset which is to date considered as a standard for genre classification and we will also be using the same in this project.[5] T. Lidy and A. Rauber in 2005 studied the effectiveness of psycho-acoustic features in defining and modeling rhythmic content of music, and with introduction of two new features they were able to improve the performance of music genre classification by up-to 9.33%. With advancement in the area of speech and audio analysis, MFCCs features were introduced to the world, these features are widely used in speech recognition and they provide a compact representation of the spectral envelope of an audio signal. Almost all the research afterwards included MFCCs in their feature space. [6] Yannis Panagakis, Gonzalo R. Arce experimented with sparse representation-based classifiers with above mentioned auditory features. With some assumption on dataset and complex modelling they were able to get 60% accuracy. Earlier stages of research used ML techniques like SVM, K-nearest neighbors and some other linear classifiers but recently in past few years Deep Learning techniques are showing real growth.

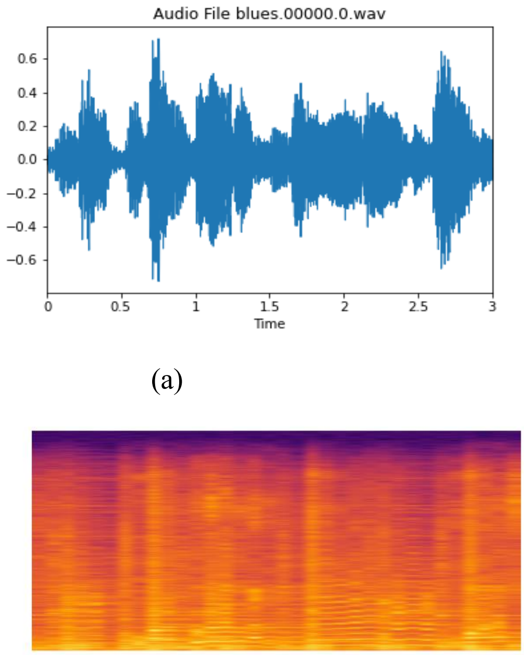
Convolutional neural networks with MEL-spectrogram produce a very high accuracy, up to 75-80% (On specific dataset) with some complex modification and pre-processing, which is almost 20% more than the average human accuracy to classify the genre. Many researchers are working on different configuration of CNNs like numbers of layers, size of filter etc. to learn the pattern more precisely.

3 Dataset analysis and Feature Extraction

This project will use a publicly available GTZAN dataset that contains music files grouped in 10 genres- blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock. There are 1000 samples in total, with 100 samples for each genre.



The duration of each audio clip is 30 seconds. The sampling rate is 22050. The GTZAN dataset is the most-used public dataset for evaluation in machine listening research for music genre recognition. The dataset is available in Kaggle and comes with audio and its image-based representation in form of MEL-spectrogram.



(a)

(b)

Fig 1: Sample Amplitude envelope(a), spectrogram (b), from blues genre.

3.1 Pre-processing for raw audio

a) The size of our dataset is 1000 and to increase it we will perform data augmentation by taking random clips of 3sec from 30 sec clips, it will increase the size of data by almost 10 times.

3.2 Major Feature Extractions

We will mainly consider the following features of music in our model–

a) Zero Crossing Rate: It is the rate of sign changes along the signal.

b) Spectral centroid: It indicates where the center of mass is located.

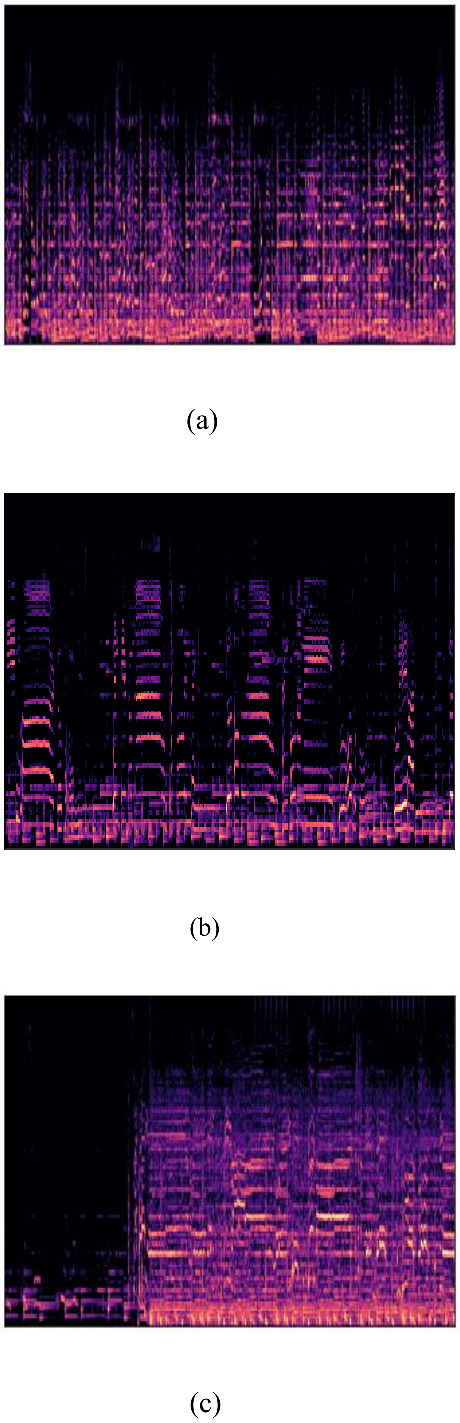
c) Spectral Rolloff: It measures where given amount of signal lies and helps in characterizing the signal.

d) Mel-spectrogram: It defines the overall shape of spectral envelope of a signal. First, we do STFT of the signal and then apply Mel scaling followed by discrete cosine transformation of logarithmic of Mel scaled bins.

Note: Mel-spectrogram will be used for CNN based model as it expects an image as an input.

e) MFCCs: Representation of power spectrum of a signal. They are the coefficients representing Mel spectrogram.

f) Chroma Features: It projects the audio into 12 different pitch classes.



(a)

(b)

(c)

Fig 2: Sample Mel-spectrogram from blues(a), country(b), hip-hop(c).

For all the above features, we will consider the mean and variance of overall feature vector for training model. Feature a, b, c is useful in case of any signal but d, e, f is particularly effective in case of audio and music.

4 Methods

4.1 K-nearest neighbor

After extracting the features from audio, we applied the KNN algorithm. The concept behind KNN is to individually look at K neighbors of test points and predict the most popular label. Neighbors are selected based on their Euclidean distance. In our project, we have taken k=10 as it gave the best results.
The below equations show the weighted associated with each neighbor–

$$w_i \propto ||x - x^{(i)}||_2, \quad \sum_{i=1}^{1} 0 w_i = 1.$$

Prediction –

$$\arg\max_y \sum w_i \mathbb{1}(y = y^{(i)}),$$

4.2 Support Vector Machine

Similar to previous model, SVM with rbf kernel was applied on extracted features. SVM is a margin classifier which can learn complex relationships by finding

$$\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \quad \text{s.t. } y^{(i)}\left(\sum_j \alpha_j K(x^{(j)}, x^{(i)}) + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

The kernel used is

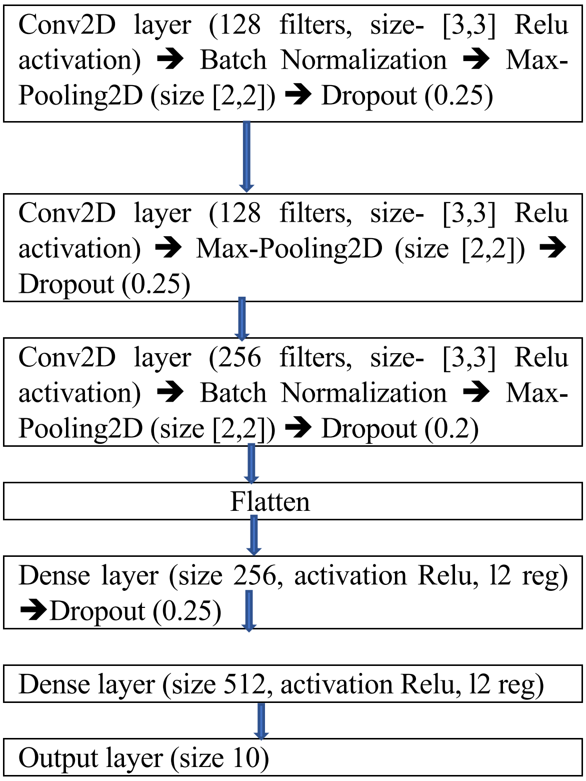
$$K(x^{(j)}, x^{(i)}) = \exp\left(-\frac{||x^{(j)} - x^{(i)}||_2^2}{2\sigma^2}\right).$$

Where w's are weights of the classifier and alpha is the coefficients of data points (non-zero only when functional margin =1) and epsilon is the model error. Kernelization helps in transforming the data into higher dimension so that they can be linearly separated.

4.3 Convolutional Neural Network (CNN)

This is deep learning based neural network which use convolutional filters in order to learn features of input. By literature we know that CNN works best with images and related classification tasks so we use Mel-spectrograms for training CNN.

Model



Conv2D layer (128 filters, size- [3,3] Relu activation) ➔ Batch Normalization ➔ Max-Pooling2D (size [2,2]) ➔ Dropout (0.25)

Conv2D layer (128 filters, size- [3,3] Relu activation) ➔ Max-Pooling2D (size [2,2]) ➔ Dropout (0.25)

Conv2D layer (256 filters, size- [3,3] Relu activation) ➔ Batch Normalization ➔ Max-Pooling2D (size [2,2]) ➔ Dropout (0.2)

Flatten

Dense layer (size 256, activation Relu, l2 reg) ➔Dropout (0.25)

Dense layer (size 512, activation Relu, l2 reg)

Output layer (size 10)

The above model was finalized after trying different parameters and choosing the best. Model contains 3 convolution layers with Relu activation. Max-pooling, batch normalization and dropouts are added in intermediate steps. After convolution layer flattening operation is performed followed by two fully connected layers with 0.5 dropout each and final layer corresponds to the output.

5 Result and Analysis

Metric we used to evaluate our models is accuracy, which in simple word is amount of data classified correctly.

Accuracy = (# points classified correctly)/ (#Total data points)

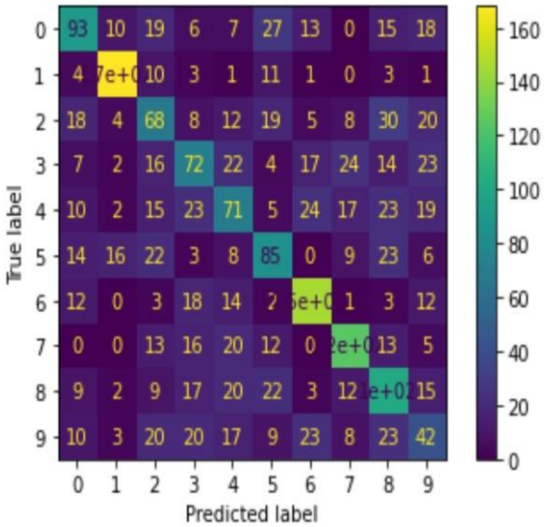
We selected all the hyperparameters based on empirical results and also by trying different values and checking model's performance. Some of our decisions in choosing parameters were motivated by our academic understanding of the course and prior knowledge in this field.

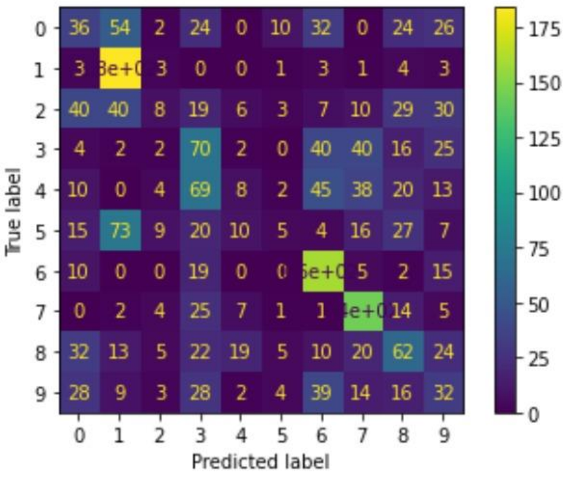
Model	Train acc	Test acc
KNN	1	48.65
SVM	57	35
CNN	88	67

Table 1: Performance of different models

For KNN we tried with (1,4,7,10,15,25) for the value of K and got best results for K=10, and for SVM we chose rbf kernel as it theoretically transforms the data into infinite dimension and the data, we considered was too complex to be classified in lower dimensions. A low accuracy for KNN somehow indicates that there is no spatial relation between audio signals of same class which can be manipulated by ML algorithms.

(a)





(b)

Fig 3: Confusion matrix for KNN(a),and (b)SVM

From both the confusion matrices we can say that music genre classification in not an easy task for conventional ml algorithms even with various features.

For CNN window sizes were chosen experimentally, we played with the values of filter size, dropout ratio, regularizer, max-pool layers/type in order to get the optimal model. We chose Adam optimizer since the image representation of time domain audio results in sparse gradients problem. Adam optimizer mitigates sparse gradient problem by taking varying learning rate and also reduces the effect of noise by taking momentum into account.

From accuracy of all models, we can see that all models suffer overfitting phenomena and the diff between train and test accuracy is quite significant in case of all the three models. From further analysis it can be seen that our CNN based model fails to classify rock more than 50 percent of the time and confuses it with country or blue.

6. Conclusion and Future work

Of all the classifiers considered, CNN with frequency-based Mel spectrogram works the best compared to feature-based linear classifiers. Working with raw audio gives an idea about how the signal's amplitude varies rather than its rhythmic content. In contrast, in this project, we explored different features to characterize the audio signals. Mel-spectrograms are a visual representation of audio signals, and CNN performs best on images, so using them with CNN resulted in the best performance. Failure of SVM and KNN in genre classification task indicates that the underlying hypothesis is way more complex to be solved by linear classifiers.

As we found that deep learning approach works better in case of genre classification, researchers can look into different deep learning techniques like RNN, LSTM etc. to achieve better results. We can also look into transfer learning for genre classification task. Future studies may also identify ways to pre-process data to eliminate noise up-to a good extent

7. Contribution

All the project work is completed in pretty collaborated manner and all the team members participated in all parts of the project but took responsibility of few micro tasks as follows –

Rameshwar (2017180) – Literature review, Model analysis, Data pre-processing, feature extraction.

Sachin Nandal (2017185) - Literature review, KNN and SVM training and prediction. Dataset analysis.

Dheeraj (2017044) – CNN training and testing, Literature review, Result analysis, Idea and proposal.

The task of project report and presentation was a team effort.

References

[1] George Tzanetakis, Perry Cook

Musical Genre Classification of Audio Signals

IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 10, NO. 5, JULY 2002

[2] VISHNUPRIYA S, K. MEENAKSHI

Automatic Music Genre Classification using Convolution Neural Network

International Conference on Computer Communication and Informatics (ICCCI -2017)

[3] Chandsheng Xu, Mc Maddage, Xi Shao, Fang Cao, and Qi Tan

Musical genre classification using support vector machines, IEEE Proceedings of International

Conference of Acoustics, Speech, and Signal Processing

[4]Snigdha Chillara, Kavitha A S, Shwetha A Neginhal, Shreya Haldia, Vidyullatha K S

Music Genre Classification using Machine Learning Algorithms: A comparison

International Research Journal of Engineering and Technology

[5] Thomas Lidy and Andreas Rauber. 2005.

Evaluation of feature extractors and psycho-acoustic transformations for music genre classification

Vienna University of Technology

[6] Yannis Panagakis, Gonzalo R. Arce

MUSIC GENRE CLASSIFICATION USING LOCALITY PRESERVING NON-NEGATIVE TENSOR FACTORIZATION AND SPARSE REPRESENTATIONS