Music Genre Classification

Manav Desai
(202118005)
MSc DS, DAIICT
Gandhinagar, Gujarat
manavd522@gmail.com

Kashish Kothari
(202118011)
MSc DS, DAIICT
Gandhinagar, Gujarat
kashishk9.9.00@gmail.com

Dhvani Golani
(202118020)
MSc DS, DAIICT
Gandhinagar, Gujarat
dhvanigolani2000@gmail.com

Himanshu Verma
(202118029)
MSc DS, DAIICT
Gandhinagar, Gujarat
vermahimanshu1720@gmail.com

***Abstract*—It is safe to assume that music nowadays is as important in our lives as food and water. Better experience of music is possible when every listener is catered to his/her preferred choice of music. Most of the current Music Genre Classification projects work with a variety of classes/genres listened to all over the globe. In this project, we are focusing on genres that cater to the Indian audience by using a data set that contains music genres prevalent in the Indian Subcontinent, followed by their conversion to Mel-Spectrograms to train Machine Learning and Deep Learning models.**

***Index Terms*—Convolutional Neural Network, Music Genre Classification, Supervised Deep Learning, K-Nearest Neighbour Algorithm**

I. Introduction

Music has been the oldest and an integral element in people's life. While travelling or working, in morning or evening , in happy or sad moments music always walks along as a friend. Music has been downloaded and bought from different app and online music collection sites at a very high rates such that it has now become a daily life for major population.When it comes to bringing people together, the type of music should also be kept in mind for people with varying tastes.

Now the major question is what is genre? A genre of music is how it is classified. In our case the classifications are ghazal, sufi, semi-classical etc. Music genre have evolved over the time, for instance, Alternative/Rock, LoFi, etc. are genres which weren't prevalent earlier, unlike today. In today's time, music applications we find, have sorted music for us on the basis of the artist's name or the album. All these factors makes it difficult to find music of ones own choice. Since categorization is not possible manually, we take help of Machine Learning Algorithms.

Unlike the majority of Music Genre Classification models, we are working with a data set that contains Bollywood/Indian Subcontinental music, hence making our model more reliable to the Indian audience. Our main goal is not just to build a classifier, but the kind which can actually be implemented by our target audience and could satisfy their music needs.

II. Problem Statement

Music Classification is considered a very difficult task due to the unavailability of data set with appropriate tags along with the audio. The first step consists of visualising audio samples as waveform representation. Then follows the classification based on the extracted features. The main issues faced during Music Genre Classification are Feature Extraction from the given audio sample without losing any important information in the process. To make our Classifier more niche, the data set selected is made up of Indian Music Genres.

III. Data

A. Data Description

In this project, we will use an Indian Music Genre Data set from Kaggle. This is a freely available collection of audio samples from various Indian music genres. It is a balanced data set of 500 audio files divided into 5 classes/genres, each containing 100 high-quality .mp3 audio files of 45 seconds duration. viz.

Genre	Number of tracks
BOLLYPOP	100
GHAZAL	100
SEMI CLASSICAL	100
CARNATIC	100
SUFI	100

B. Data Pre-Processing

Our data mainly consists of audio files based on Indian Genre.Hence to extract features and to study the data in a better manner we convert the audio files into mel-spectrograms and MFCCs.

There are 2 ways to represent sound (music):

1. Time Domain - show the amplitude (loudness) of sound waves with respect to time which is not very informative and hence to better understand the sound we move to frequency domain

2. Frequency Domain- tells us about different frequencies present in the signal.

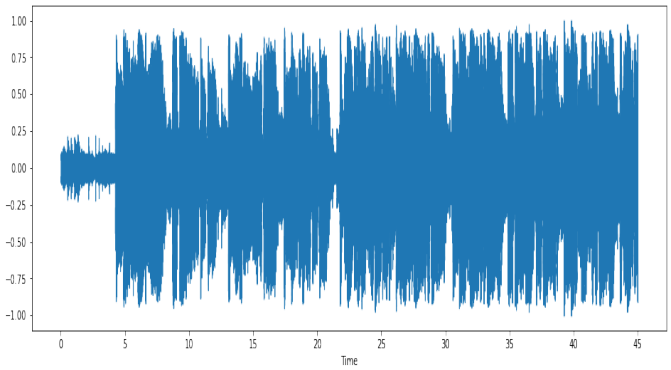

Fig. 1. Audio Waveform

Further, we move on to creating a spectrogram. A Spectrogram is a graphical or a visual representation of frequency of a signal with respect to time. where the x-axis represents time and y axis represents the frequency and the colour density depicts the amplitude of a frequency at a particular time.

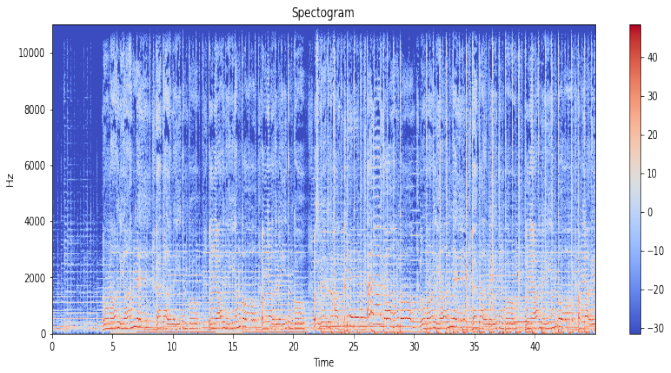

Fig. 2. Spectogram

Moving on, we plot mel-spectrogram. A mel-spectrogram is again a visual representation where frequency is plotted on mel- scale. Then after, in the final plot, x-axis represents time and y-axis represents mel scale.

Mel- spectrogram uses linear spaced frequency scale and so we switch to Mel-frequency cepstral coefficients (MFCC). MFCC is derived from a type of cepstral of audio clip. In mel-frequency cepstrum (MFC), the frequency band is distributed uniformly on the mel scale, to which the auditory system of human responds closely.

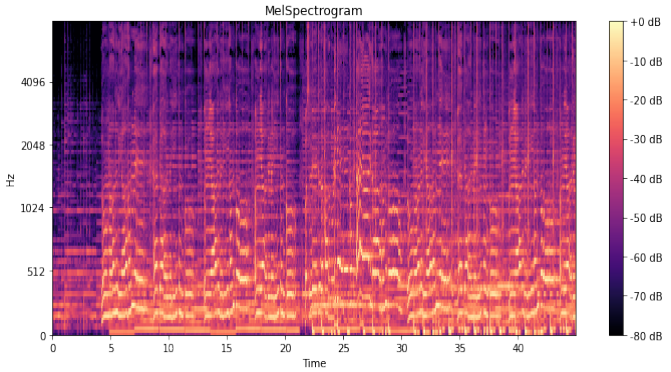

Fig. 3. Mel-spectogram

IV. Methodology

With the help of following methodologies we are able to achieve the following desired results :

A. Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of deep neural network which can take an image as an input and differentiate it from others after assigning learnable weights and biases to various aspects of the image. The main advantage of using CNN is its little dependence on pre-processing, i.e. it requires very less pre-processing and human intervention while developing the functionalities.
A CNN architecture consists of convolutional layers along with fully connected layers at the end to create a multilayered neural network. Every layer consists of number of neurons which take input from the previous layer or as an initial input and after performing dot product with the weights, the result is passed onto the next layer or passed as the output.

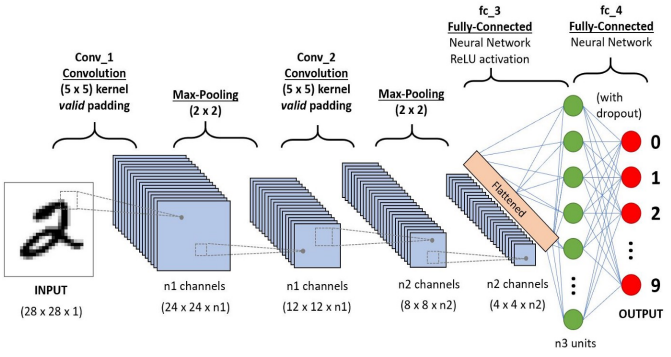

Fig. 4. A CNN sequential model architecture
[4]

We generally use three types of layers to build a Convolutional Neural Network : Convolutional Layer, Pooling Layer, Fully Connected Layer. MaxPooling is a common layer used in CNN which is similar to downsampling the data. It basically reduces the dimensionality of the images by reducing the pixels from the output of the previous layer.

B. K-Nearest Neighbour

The K-Nearest Neighbour algorithm is a supervised machine learning technique that can address both classification and regression problems. KNN estimates the distance between all points in close vicinity to the unknown data and eliminates those with the smallest distances. As a result, it's sometimes called a distance-based algorithm. The KNN algorithm tends to use a majority voting mechanism. It gathers data from a training set and utilises it to create predictions for new records later on. It makes no assumption of data as it is a non-parametric algorithm. The count of the nearest neighbours is indicated by K-value. Distances between test points and training labels points must be computed. KNN is a slow learning algorithm since updating distance metrics with each iteration is computationally expensive.

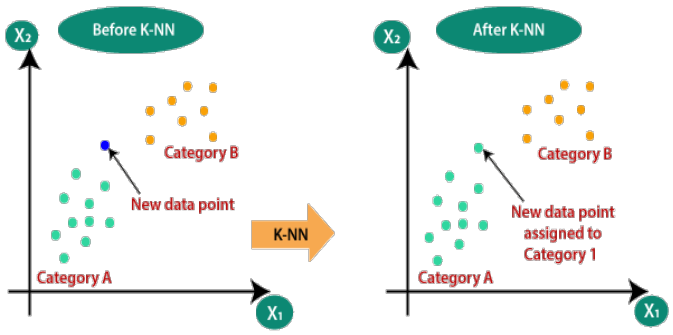

Fig. 5. KNN model
[3]

C. Random Forest Classifier

Random forest is a supervised machine learning algorithm used for classification and regression problems. It creates decision trees using various samples, by the majority vote for classification and the average for regression.

V. Experimental Results

In the proposed model, we use supervised learning approaches like K-Nearest Neighbour, Random Forest along with a deep learning approach like Convolutional Neural Network. The audio samples first go through the pre-processing stage where they are converted to MFCCs; these are our features. These coefficients are then shaped into 2-dimensional arrays which are given as input to the KNN and Random Forest models. The accuracy of the model is calculated using the

$$Accuracy = \frac{Numberofsongscorrectlyclassified*100}{Totalnumberofsongs}$$

A KNN model is trained on the dataset resulting in accuracy of 56%. In comparison, Random Forest Classifiers are also part of the experiment.To find better hyperparameters for the Random Forest Classifier, we used RandomizedSearchCV provided by sklearn.ensemble library. The search is carried out against 10 different variations. The variations are selected randomly from the following grid.

Hyperparameters	Values
n_estimators	[10,100,200,500,1000,1200]
max_depth	[None,5,10,20,30]
max_features	["auto","sqrt"]
min_samples_split	[2,4,6]
min_samples_leaf	[1,2,4]

Even after tuning the hyperparameters, the model performed poorly with an accuracy of 25%. Following are the best hyperparameters found after RandomizedSearch.

Hyperparameters	Values
max_depth	20
max_features	'auto'
min_samples_leaf	2
min_samples_split	2
n_estimators	10

For CNN, the features are transformed into 4-dimensional array of shape (batch size, 40, 646, 1). A batch size of 64 and an initial learning rate of 3e-4 is maintained throughout the training process. The model is trained for 100 epochs and the learning rate is dynamically reduced by a factor of 0.2 if the validation loss is unaffected for 5 epochs until the minimum learning rate is 1e-4. The CNN model consists of 5 consecutive Convolution layers with ReLU, each of which is followed by Max Pooling and Batch Normalization as described below. The resultant features from the layers above is flattened and fed as an input to a feed-forward layer with 64 and a dropout rate of 0.3 with ReLU Activation. The output layer is a dense layer of 5 neurons with softmax activation and the resultant vector will represent the probability of the audio belonging to a genre. Following is our CNN architecture in detail:

Layer	Filters	Kernel	Activation
Conv2D + MP + BN	32	3x3	ReLU
Conv2D	64	3x3	ReLU
Conv2D + MP + BN	64	3x3	ReLU
Conv2D + MP	128	3x3	ReLU
Flatten	-	-	-
Layer	**Neurons**	**-**	**Activation**
Dense + Dropout	64	-	ReLU
Dense	5	-	Softmax

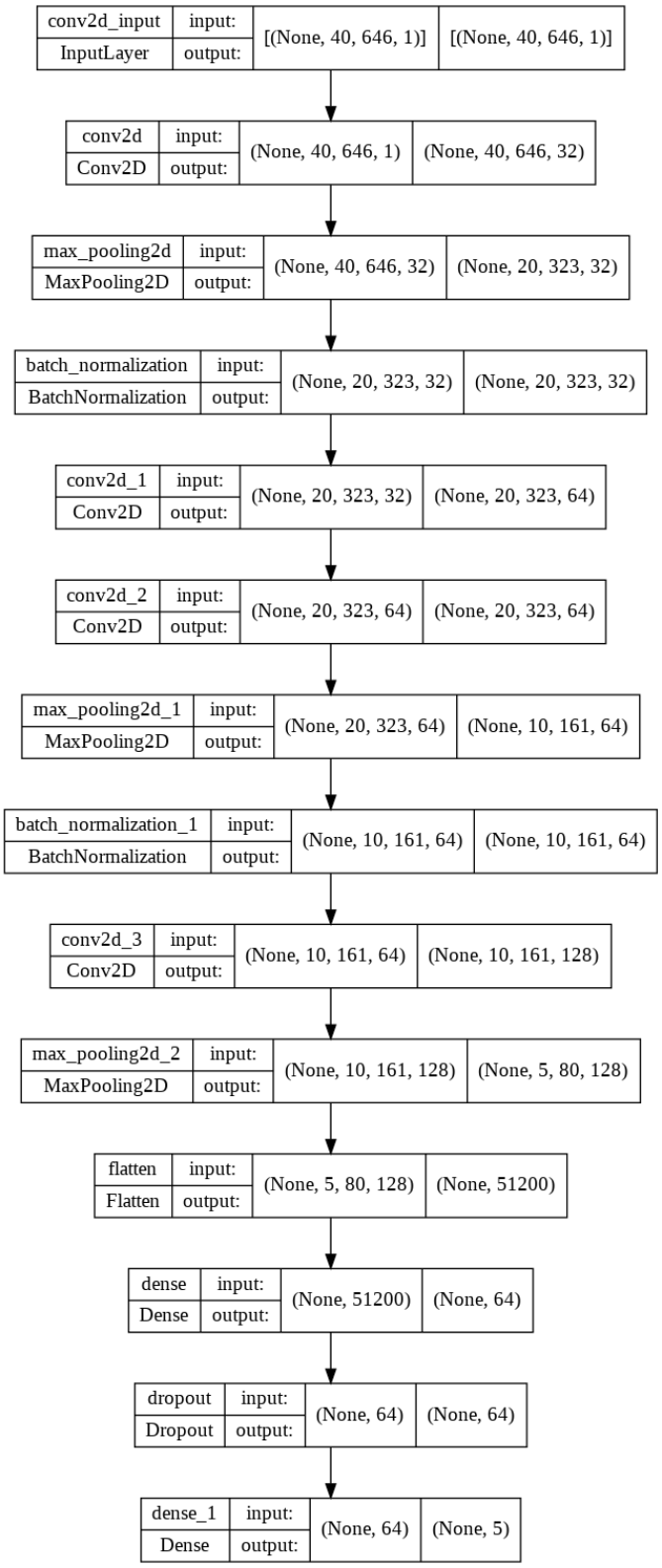

Fig. 6. Model Architecture

The trained CNN model performs decently with a test accuracy of 78% and training accuracy of 88% but can be improved using hyperparameter tuning given the resources.

	Training	**Testing**
Accuracy	88%	78%
Loss	0.2365	0.6924

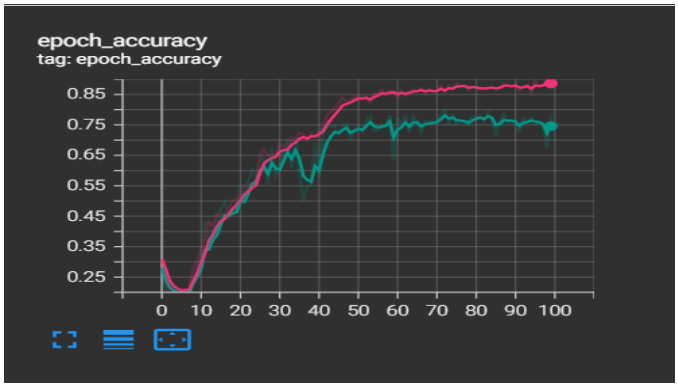

Fig. 7. Model Accuracy

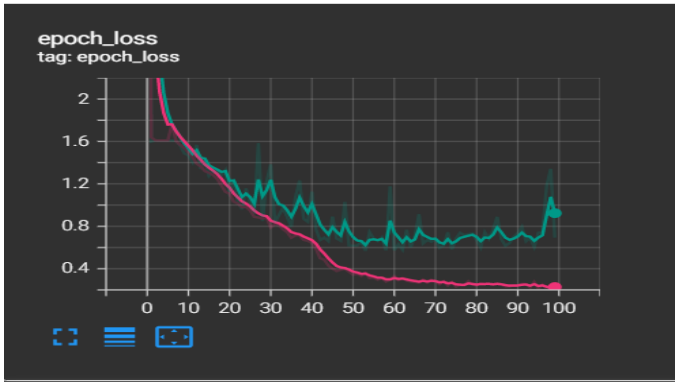

Fig. 8. Model Loss

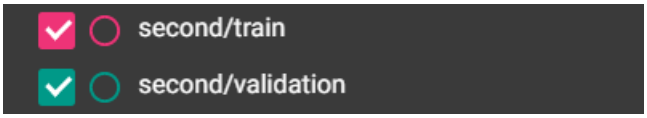

Fig. 9. Legend

Model	**Accuracy**
CNN	0.78
KNN	0.56
Random Forest	0.25

VI. Conclusion and Future Scope

In conclusion, CNN performed the best out of the three with the test accuracy of 78%. However, provided a larger dataset and more resources, this can be improved extensively.
The proposed deep learning model can be improved in future using several different techniques like Hyper-Parameter Tuning, Multimodal Features and Data Augmentation. Moreover, we can even explore the realms of Recurrent Neural Networks, specifically LSTMs. Recent publications have also enforced the use of transformers due to its comparative performance and less training time compared to LSTM.

References

[1] S. Vishnupriya and K. Meenakshi, "Automatic Music Genre Classification using Convolution Neural Network," 2018 International Conference on Computer Communication and Informatics (ICCCI), 2018, pp. 1-4, doi: 10.1109/ICCCI.2018.8441340.

[2] L. K. Puppala, S. S. R. Muvva, S. R. Chinige and P. S. Rajendran, "A Novel Music Genre Classification Using Convolutional Neural Network," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1246-1249, doi: 10.1109/ICCES51350.2021.9489022.

[3] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[4] https://www.kaggle.com/winchester19/indian-music-genre-dataset

[5] C. Chen and X. Steven, "Combined Transfer and Active Learning for High Accuracy Music Genre Classification Method," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 53-56, doi: 10.1109/ICBAIE52039.2021.9390062.

[6] https://en.wikipedia.org/wiki/Mel-frequency$_cepstrum$