Genre Classification of Bengali Music Based on Audio Features

A project Report
CSE 4238 - Soft Computing

Submitted by

Anika Tanzim	**160204072**
Email:	**160204072@aust.edu**
Marufa Kamal	**160204073**
Email:	**160204073@aust.edu**

Submitted to

Mr. Mir Tafseer Nayeem

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Music is a cross-cultural universal, a global activity found in every known society man lives in. It is present everywhere we go, from airports to malls, restaurants to office parties, and even in the safe abode of one's home. With the upsurge of music currently in the world in a study, it has been observed that Americans spend more on music than they do on prescription drugs [1]. For the average American, background and deliberate music listening add up to more than 5 hours of music viewing a day [2]. This research demonstrates the popularity of music and how it affects the lives of ordinary people.

1.1 Origin of Bengali Music

Reflecting on the long history of Bangladesh, it has a culture that encompasses music with a sense of its heritage and origin. The start of music in Bangladesh starts long back by the Buddhist monks or saints between the 7th to 12th centuries composing a music genre named Charyacharyavinishcha, generally referred to as charyagiti [3]. Starting from traditional classical music, Bangladeshi music can be classified into a variety of genres. Humans construct and use musical genre labels to categorize and describe the vast universe of music. In recent times Bangladeshi music has become diverse and distinct with the influence of western culture and new trends. The primary genres in Bangladesh are Classical music, Rabindra sangeet, Nazrulgeeti, folk songs, Polligeeti, and Adhunik Gaan(modern music). These musical genres have no definite rules or limits since they originate from the dynamic interaction of popular understanding, marketing, history, and culture. Despite having many genres, the members of a particular genre share similarities between them. These characteristics mostly relate to instrumentation, rhythmic structure, pitch, and energy of the particular music.

1.2 Why Focus on Bengali Music Genre Classification?

Extracting music information and categorizing the music genres are popular nowadays with the upsurge of free music files on the internet whether it is an old song or a new one, the internet has it all. Automatic Music genre classification is widely used to create a playlist of the same genre and classify songs. Renowned applications such as **Napster**, **Spotify** , **Soundcloud** use genre classification to recommend music to their users. With a large number of songs in a database, it becomes difficult to manually annotate the genre to the songs and recommend them to the users. This is why they lean more toward music genre classification. Bengali music has a wide range of tones and diversities which makes the genre classification more interesting. When it comes to musical data, it is difficult to categorize music solely based on its title as the mood of the song can be quite different from the delivery it provides through the tunes of the audio. The audio features of the same data can be similar for example the tempo can hold the same number of beats, the frequency of the beats can have similar amplitudes, and such. Although there has been significant work in the development of features for English songs and their Music information retrieval(MIR), there is a lot of scope for Bengali songs and their genre classification techniques.

For our project we aim to classify 6 different genres of Bengali music namely: Polligeeti, Deshattobodhok, Cholochitro, Nazrulgeeti, RabindraSangeet and Adhunik Bangla from a dataset of Bengali songs. We will will extract audio features [4] [5] from the audio datas. Here, we are proposing a deep learning neural network model to analyze the audio data and find similarities in between the same genres of music.

Chapter 2

Related Work

Over the time, Many researchers have been working on music genre classification and have applied different models of Machine Learning and Neural Network. There are lots of works that can be found with machine learning and the Neural Network in English Music genres. But there are a few works that have been found on Bangla Music genres classification. Most of them have used Machine learning to implement their proposed model. Some of our related works are listed here:

2.1 English Music Genre Classification with Machine Learning

In 2017, Asim Ali and zain Ahmed in their research [6] proposed a model with accuracy 77% that has been trained with Machine learning algorithm "KNN" and "SVM". They have classified English Music into 10 genres: blues, classical, rock, jazz, reggae, metal, country, pop, disco, and hip-hop. From results, they have found that without the dimensionality reduction both k-nearest neighbor and Support Vector Machine (SVM) gave more accurate results compared to the results with dimensionality reduction. Overall the Support Vector Machine (SVM) is a much more effective classifier for the classification of music genres. It gave an overall accuracy of 77%. In this paper, Mel Frequency Cepstral Coefficients (MFCC) is used to extract information from our data. They have used the GTZAN dataset which has a collection of thousand sound files. Each of the files is thirty seconds in length. Their input dimension was 156 as they have computed the mean, std, min, and max of MFCCs with 39 feature vectors.

In their work, they have used only one feature i.e., MFCC. Using more features, they could achieve a better model.

2.2 A Comparative Study on Content-Based Music Genre Classification

In this research based on the comparative study of genre classification tehniques [7], the writers used the regular feature vectors along with DWCH to extract features from the 30-second audio files. They used MARSYAS, a public software framework for computer audition applications, for extracting the features such as MFCCs, FFT, Beat and Pitch. They also used the DWCH method to extract feature sets containing four features for each of seven frequency subbands along with nineteen traditional timbral features, but for better performance, they discarded the sets which contained less information remaining with 35 feature vectors. we use three different reduction methods to extend SVM for multi-class: pairwise, one against-the-rest, and multi-class objective functions. For experiments involving SVMs, they tested them with linear, polynomial, and radius-based kernels. For Gaussian Mixture Models, they used three Gaussian mixtures to model each music genre. The performance with FFT and MFCC was significantly higher than that with Beat or Pitch. In the case of the models used they have shown a comparison with different features and modeling techniques. The accuracy of the one-versus-the-rest SVM is 78.5% on average in the ten-fold cross-validation for DWCH features.

2.3 Bangla Music Genre Classification using Machine Learning

In the paper [8] Abhijit and Ehtesham in 2019, proposed a machine learning model in Bangla Music genre classification where they have compared SVM (Support Vector Machine), J48 (C4.5 Pruned / Unpruned Decision Tree), and NB (Naive Bayes Classifier). They have demonstrated the efficiency of the proposed method by extracting features from a dataset of 1200 containing Bangla music pieces and testing the automatic classification decisions. They applied a binary classification scheme. They have said that Support Vector Machine gives the best classification regarding music genre differentiation and Naive Bayes gives the worst. They also introduced a new feature, which was called the continuity feature. With their introduced continuity feature, they have achieved 86% with SVM. They have extracted some features like- Strongest Frequency FFT Maximum, Spectral Centroid, Zero Crossings, RMS, compactness, etc. But they have missed some important features by which they could achieve more accuracy. Again, they did not normalize the dataset which is important in machine learning to get better results.

2.4 Bangla Music Genre Classification Using Neural Network

In the paper [9] published in 2019, they made an approach for using neural networks to classify music genres. In this paper, they used 8 feature vectors: ZCR, Tempo, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, MFCC, Chroma Frequency, RMSE in order to determine the audio signal characteristics. They prepared a dataset with 6 genres of music and a total of 1742 songs and extracted all the features from these songs and trained the neural network models with those features. They trimmed each song to 120 seconds to keep the computation cost lower. They used different models to train the dataset, for instance, SVM, logistic regression, linear regression, KNN, and neural network model. They got an accuracy of 74% using the proposed 5layer NN model. They have used a large neural network model to train the data which could have been reduced for faster computing.

Chapter 3

Project Objective

Music has a great significance and is a constant part of people's daily life. When it comes to Bangla music, there are so many types and styles which can be classified as genres. Music classification plays a very fundamental role for the retrieval of music information and for music recommendation. Though Bangla music is very rich and traverses a great mixture of styles, there is no significant work on classifying genres of Bangla music based on their audios.

In ths study, we have proposed a system to do a multiclass classification of Bangla music genres which focuses on extracting symbolic features that have been achived by a deep neural network model.

Our work is divided into three main subtasks in a sequential manner. From gathering raw MP3 files to classify the genre, the procedures are listed here. Dividing them into different parts made it easier to undertake actions easily.We aim to target different important segments which are discussed in the following sections.

3.1 Preparing dataset

To build a model, a dataset is a first and foremost property that is unavoidable. In our specific work, resources as a prepared dataset of Bangla music are not available in such a manner we need. For having the low resources, we have to collect data which are in the form of MP3 files at the beginning. We are going to collect our desired data from different online sources like-music.com.bd and mp3haat.com. After collecting the audio files, we have to remove the files which are not usable i.e., bad data. We have to label the files as per their genres in different directories. The genres are already determined by the online sources from where we collect the dataset. Then we have to extract the segments of a specific offset that are going to use further procedures from the raw data diles. Then,

the segments of audio files will be ready to be used as our dataset.

3.2 Extracting Features from dataset

We cannot give the raw audio segments to a Neural network as input because it does not understand anything but the numbers. We need to convert audio into some numerical values to feed to Neural Network as input. For this, we have to extract various features.

In general, there are two types of audio features:

1. **Physical features:** Physical features refer to mathematical measurements computed directly from the sound wave, such as the energy function, the spectrum, the cepstral coefficients, the fundamental frequency, and so on.
2. **Perceptual features:** Perceptual features are subjective terms that are related to the perception of sounds by human beings, including loudness, brightness, pitch, timbre, rhythm, etc.

For our project, we are going to extract 9 important features from our dataset that are going to help to classify the genres. The features will be -

1. Zero-crossing rate
2. MFCC (Mel-Frequency Cepstral Coefficients):
3. Spectral-roll off
4. Spectral flux
5. Chroma features
6. Pitch
7. Spectral centroid
8. Spectral Bandwidth
9. tempo

These features as numbers will be fee to our model.

3.3 Training Model

After extracting the features, we are ready to feed our model. We propose a Sequential model with multiple layers in it. We try different layered models to figure out a suitable one for our purpose. To train a model, at first we need to go through some steps like- Compose transorms, make data iterable, Create the class, Instantiate model class and Instantiate optimizer class.

After go through all the steps, we have to train our model with different settings to see variation in outcomes. The models run should be saved for later use.

3.4 Project Objective Example

Name of the song	Input Feaures as numeric value	Output	Name of The Genre
Aamar Cokhe (music.com.bd).mp3	-35.36661911, 21.04997063, 1124.77304459, ..., 1.341507316, 249.4001617, 107.6660156]	1	Polligeeti (পল্লীগীতি)
amar-mon-mane -na-asha.mp3	[0.08877671886, 0.002441515999, 1614.522763 , ... , 133.423445674, 26.167,21.053567, 103.4567]	0	Adhunik Bangla (আধুনিক বাংলা)
jokhon-eshechile -srabani-sen.mp3	[0.45673, 0.97424798643, 1854.654235 , ... , 120.8764, 56.8764876, 67.9872345]	5	RabindraSangeet (রবীন্দ্র সংগীত)
Tumi-Sundor-Tai _Firoza-Begum.MP3	[0.8343722755, 0.6583533996655, 1234.522763 , ... , 3445.6789, 34.987632,453.984566677]	4	Nazrulgeeti (নজরুলগীতি)

Table 3.1: Example for Bangla Music Genre Classification

Chapter 4

Methodologies

This project ranges from gathering raw MP3 files to predicting the genre of a particular class. The entire process is broke down into subtasks. We followed three steps to complete our objectives. These measures are in a logical order.

1. Preparing our training dataset with Bengali songs from 6 genres.
2. Extracting features from the audio data
3. Training the Deep Neural Network model.

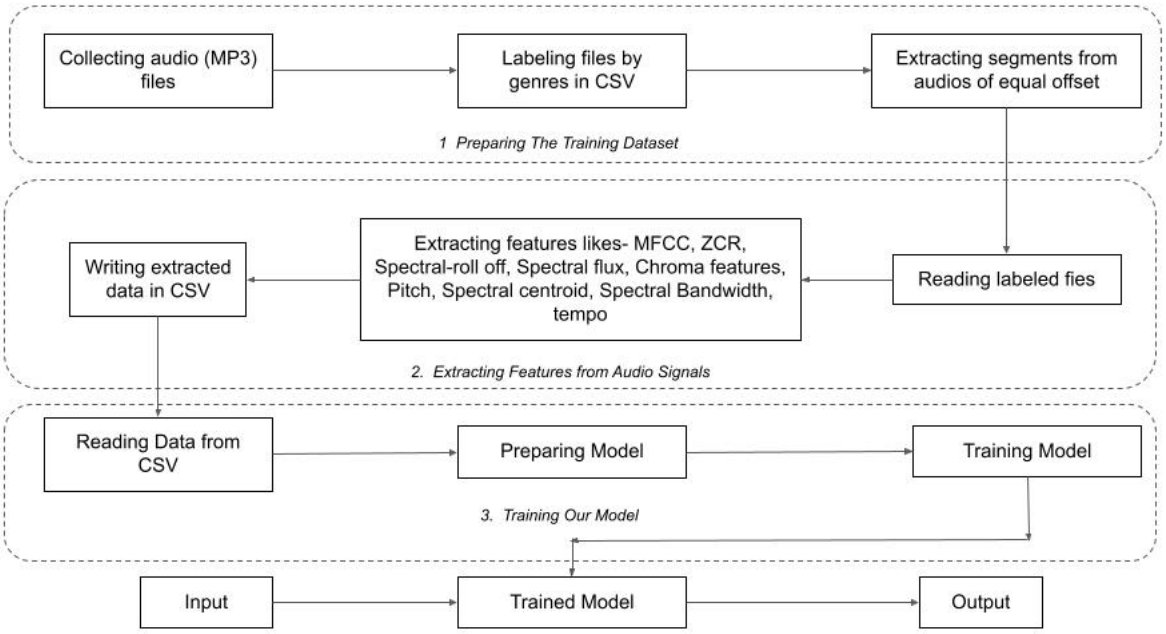

Figure 4.1: Architecture of Our Proposed Model

4.1 Preparing The Training Dataset

At the initial stage data collection was essential as there aren't many good datasets available online for Bengali music. To do so, we have scrapped out the data from 2 different websites 'www.music.com.bd' [10] and 'www.mp3haat.com' [11] containing Bengali songs. For each genre, we collected a good amount of data. To collect the data we used the python library '**BeautifulSoup**' [12]. It is a library that scrapes out data out of HTML and XML files. We have extracted the songs in a specific folder for each genre of music and saved the file link, name of the file, and the genre label in a separate CSV file. For 6 classes of genres, we have labeled the song with a numeric value starting from 0 to 5. We have annotated the data automatically while collecting the files for each genre. Next, after the raw data is collected we have extracted 100 seconds of each audio file. This is done as initially majority of the songs start with a melody or musical tone without lyrics in it which does not give out prominent audio features. Moreover the large duration of an audio file additionally increases the computational time. So, we segment all the music files starting from 20 seconds to 120 seconds. Most of the music files are about 240seconds(4 minutes) long. So, reducing the duration of the audio will not effect the features extracted from the audio.

4.2 Extracting Features from Audio Signals

Feature extraction is the process of highlighting the most discriminating and impactful features of a signal through numerical representation to characterize an audio. There are different kinds of feature extraction techniques. In our project, we use 9 types of feature extraction techniques to analyze the audio files. The features that have been extracted are sent to our model as input to train the data. We extracted both **time** and **frequency** domain features. The time domain expresses our signal as a sequence of samples, the frequency domain expresses our signal as a superposition of sinusoids of varying magnitudes, frequencies, and phase offsets. The *time-domain* features that we used are ZCR and Tempo. Different *frequency-domain* features that we used are Spectral Centroid, Spectral Rolloff, Spectral Bandwith, Spectral flux, MFCC and Chroma Frequency, etc. Pitch is a feature which is based upon a combination of time-domain and frequency-domain [13]. To implement these features a python package named **librosa** [14] is used. It provides the building blocks necessary to create music information retrieval systems and analyze music.

For every feature extraction method we calculate the mean and variance for every feature. To extract any feature, at first, we will read the audio MP3 as a time series. And then

we will extract the expected features and stack them up into CSV files. The techniques used to extract these features are discussed as follows:

4.2.1 Zero-crossing rate(ZCR):

Zero-Crossing Rate is the rate of change of a signal's sign. Or the rate of the changes to positive from negative or the vice-versa. ZCR is used to identify human speech on a music signal.The ZCR method of deciding whether a speech frame is voiced, unvoiced, or silent is quick and convenient. It is calculated by the following equation:

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{R<0}(S_t S_{t-1}) \tag{4.1}$$

Here, S is a signal of length T and $1_{R<0}$ [15]. The function 'librosa.zero_crossings()' is used to determine each of the frames zero crossing. Mean of different classes are shown in figure 4.1.

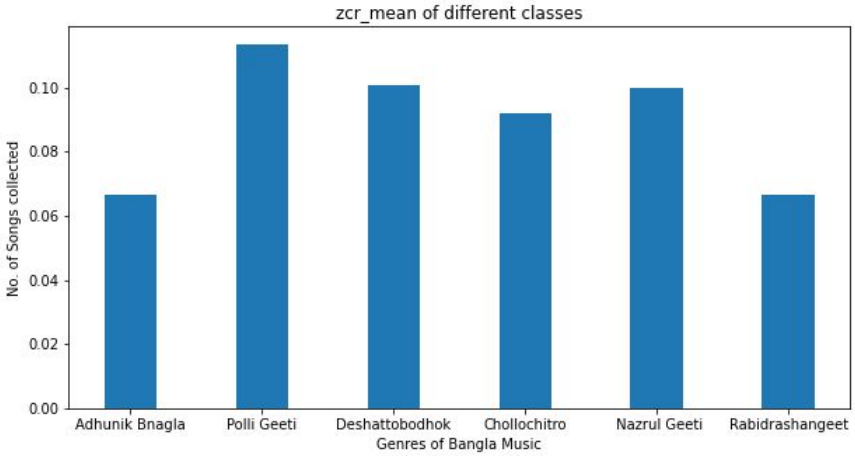

Figure 4.2: Mean ZCR of Different Classes

4.2.2 Spectral Centroid:

The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located. The following formula is used to calculate the centroid:

$$\text{Spectral Centroid} = \frac{\Sigma_{n=0}^{N-1} f(n)X(n)}{\Sigma_{n=0}^{N-1} X(n)} \tag{4.2}$$

where x(n) represents the weighted frequency value, or magnitude, of bin(the range of values) number n, and f(n) represents the center frequency of that bin [16]. The average value of spectral centroid for each value is calculated using 'librosa.spectral_centroid()' Figure visualizes the mean values of Spectral Centroids from different classes from the dataset. The max value goes to 'Cholochitro'.

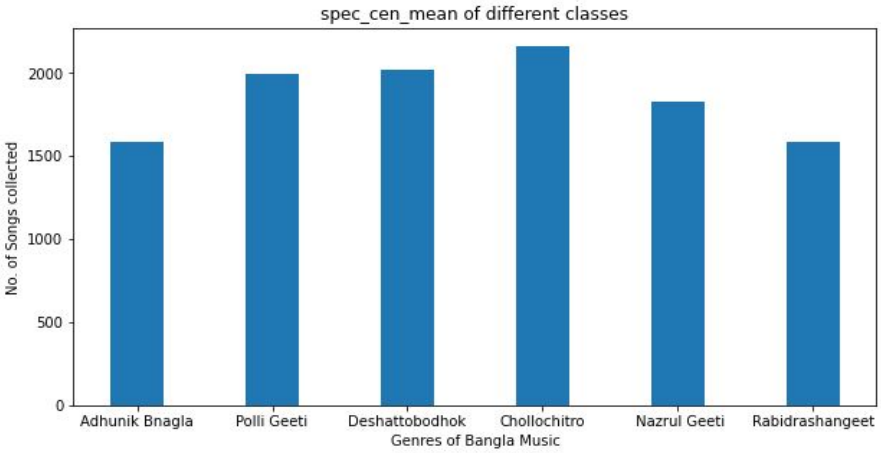

Figure 4.3: Mean Spectral Centroids of Different Classes

4.2.3 MFCC(Mel-Frequency Cepstral Coeicients):

The mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10-20) which concisely describe the overall shape of a spectral envelope. In MIR, it is often used to describe timbre. The MFCCs are calculated by applying the discrete cosine transform (DCT) to a mel-frequency spectrogram. MFCC can usually extract upto 20 features but taking 12-13 features are considered to be good for feature extraction this is why we take 13 features. We calculate the mean and variance for these 13 features using 'librosa.feature.mfcc'

4.2.4 Spectral Roll-off:

Spectral rolloff is the frequency below which a specified percentage of the total spectral energy. The equation used to calculate the spectral rolloff is as follows:

$$\text{Spectral rolloff} = \sum M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \tag{4.3}$$

The frequency below which 85 percent of the magnitude spectrum is concentrated is known as the spectral rolloff where $M_t[n]$ is the magnitude of the Fourier transform at frame t

and frequency bin. we use ‘ librosa.feature.spectral_rolloff’ to calculate the spectral roll off mean and variance value for each file. The mean calculated is shown in the followinf bar graph for all the genres.

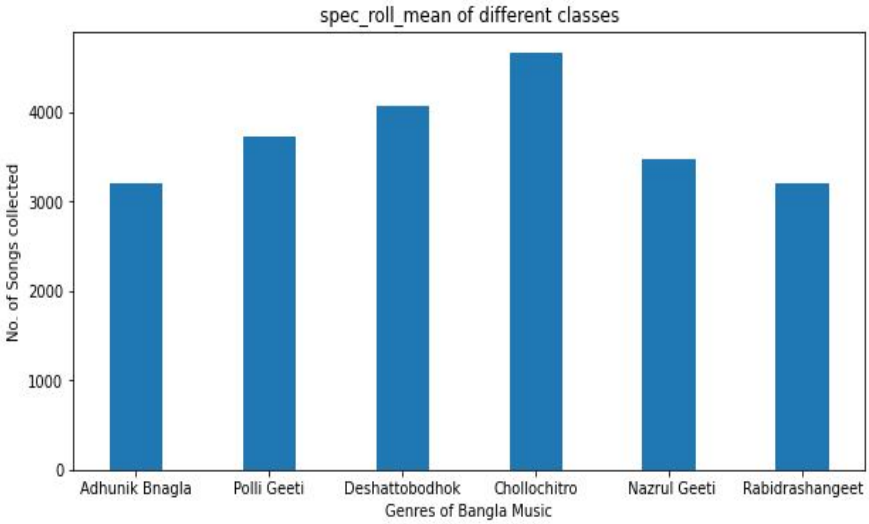

Figure 4.4: Mean Spectral Roll-off of Different Classes

4.2.5 Chroma Frequency:

The human perception of pitch is periodic in the sense that two pitches are perceived as similar in“color” (playing a similar harmonic role) if they differ by one or several octaves (where, in our scale,an octave is defined as the distance of 12 pitches).The main idea of chroma features is to aggregate all spectral information that relates to a given pitch class into a single coefficient. One main property of chroma features is that they capture harmonic and melodic characteristics of music. This is why taking this feature for our genre classification. The following bar graph shows that all the mean values are almost close to one another.

4.2.6 Spectral Bandwidth:

Spectral Bandwidth points out the frequency at which the energy of a spectrum is centered. The bandwidth determines the resolution of a signal. It is the wavelength interval in which a radiated spectral quantity is not less than half its maximum value.

$$\text{Spectral Bandwidth} = (\Sigma_k S(k)(f(k) - f_c)^p)^{\frac{1}{p}} \tag{4.4}$$

where S(k) is the spectral magnitude at frequency bin k, f(k) is the frequency at bin k, and f_c is the spectral centroid. we use ‘ *librosa.feature.spectral_bandwidth* to calculate the spec-

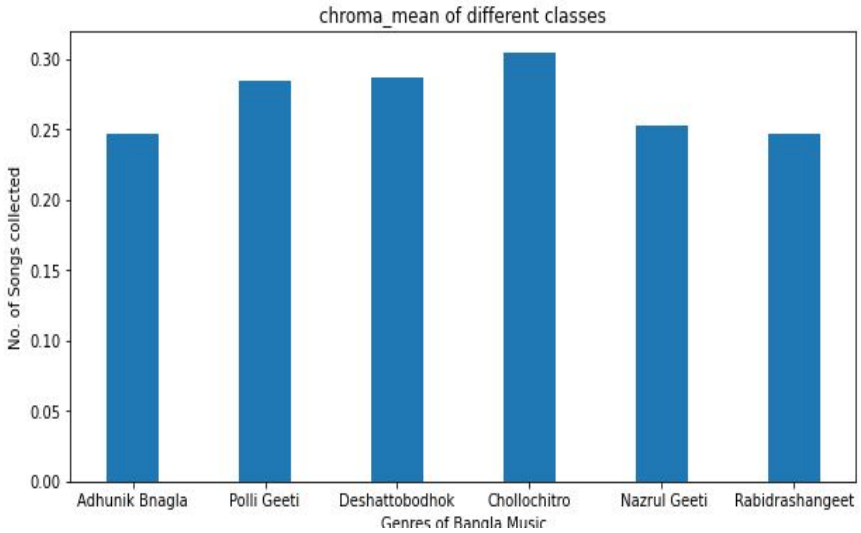

Figure 4.5: Mean Chroma Frequency of Different Classes

tral Bandwidtth mean and variance value for each file. 'librosa.feature.spectral_bandwidth' computes the order-p spectral bandwidth. Polligeeti has the lowest mean value in terms of Spectral Bandwidth.

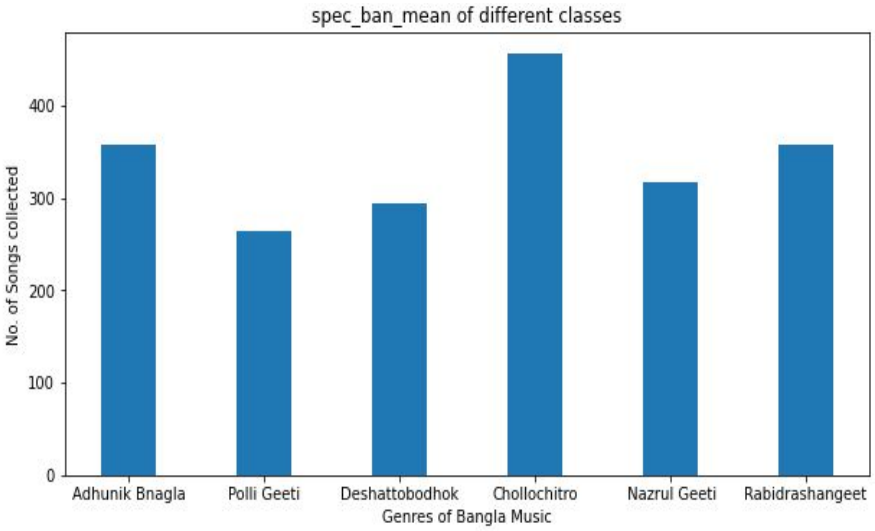

Figure 4.6: Mean Spectral Bandwidth of Different Classes

4.2.7 Spectral Flux:

Spectral flux calculates the spectral change between two successive frames and is computed as the squared difference between the 2 normalized magnitudes of the successive spectral distributions.

$$Fl(i,i-1) = \sum_{k=1}^{Wf_L} (EN_i(k) - EN_{i-1}(k))^2 \tag{4.5}$$

where $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_l} X_i(l)}$, i.e., $EN_i(k)$ is the k^th normalized DFT coefficient at the i^th frame. 'librosa.onset.onset_strength' have been used from librosa python package to calculate the spectral flux. All the means values are almost equal for the classes.

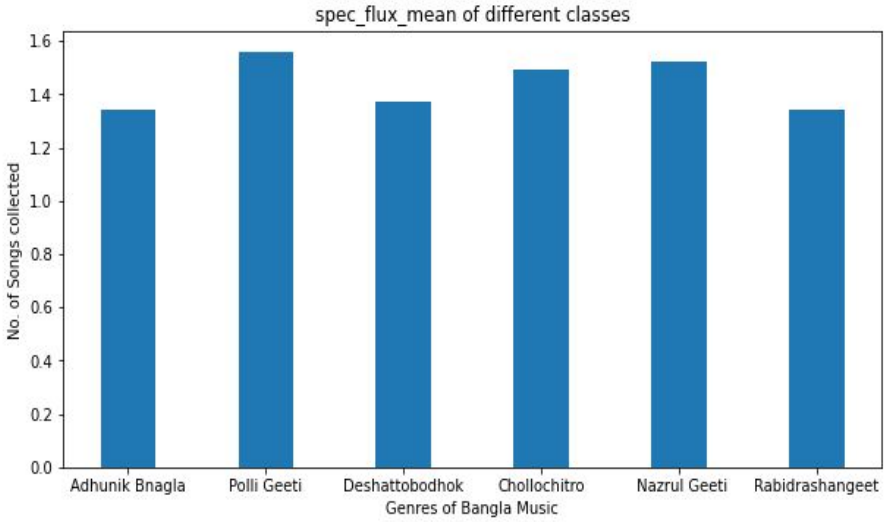

Figure 4.7: Mean Spectral Flux of Different Classes

4.2.8 Pitch:

The vibration of a frequency is generally referred to as the pitch of a audio signal. For a high frequency audio wave it can be said is is a high pitch audio and a low pitch audio implies to a low frequency audio wave. 'librosa.piptrack' calculates the pitch of the audio files. Adhunik Bangla and NazrulSangeet has the lowest mean whereas Polligeeti has the higest pitch value.

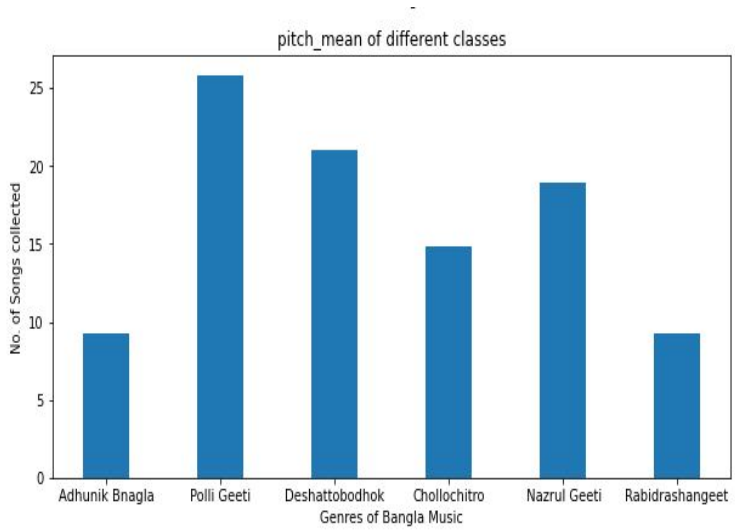

Figure 4.8: Mean Pitch of Different Classes

4.2.9 Tempo:

In terms of music, Tempo is the speed at which a segment of music is played. A tempo of 60 BPM, for example, means that a beat occurs precisely once per second. We have used 'librosa.beat.tempo' to find the numerical data of each file based on this feature. The tempo for all the genres seems top have similarity in mean.

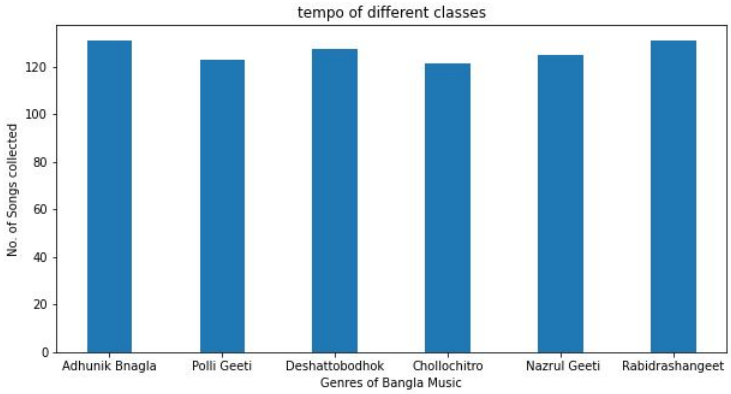

Figure 4.9: Mean Tempo of Different Classes

4.3 Training Our Model

4.3.1 Compose Transforms

We need to drop the unwanted columns and standardize data. Standardization of a dataset is a common requirement. They might behave badly if the individual features do not more or less look like standard normally distributed data. So, we have dropped 'Document Link', 'Name', 'Label',and 'Sampling rate column as we don not need them as input. Then, we have normalized the dataset for better result.

4.3.2 Make data iterable

After normalizing data, we need to make data iterable. We split our dataset into 90:10 ratio for training and testing respectively. By splitting into this ratio, we have 621 data in our training dataset and 69 data in our test dataset. We have shuffled our dataset while splitting them with **SubsetRandomSampler** and **DataLoader**. By shuffling we can have a more generalized model.

We have defined the parameters in this step. There are two types of parameters passed into NN model. They are - Modelparameter and Hyperparameter. We have tuned the hyperparameters in our model.

Modelparameter: Model parameters are learned during training when we optimize a loss function using something like gradient descent.

Hyperparameter: Parameters which define the model architecture are referred to as hyperparameters. Hyperparameters are not model parameters and they cannot be directly trained from the data. These values are fixed before the training of the data begins. They deal with parameters such as learning rate,num iterations etc. We give different settings of hyperparameters in this part for different training models. The hyperparameters we have used are -

1. **Learning rate:** How quickly the model should be able to learn, how complicated the model is, and so on.
2. **Number of iterations:** It is used to specify the number of combinations that are randomly tried. If it is too less, finding the best combination is difficult, and if it is too large, the processing time increases. It is important to find a balanced value for number of iterations.1 iteration means one mini-batch forward and backward pass. That means a parameter (wights and biases) update.
3. **Minibatch:** Number of examples in 1 iteration
4. **Epochs:** How many times we are running the dataset. 1 epoch means running through the whole dataset once
5. **Number of nodes per hiddden layer:** number of nodes in a hidden layer. It can vary in every hidden layer.

4.3.3 Create a customized Class

At this point, we create our proposed model to classify genres of Bangla Music. So far, different researches have been done to classify genres for different language's music using different machine learning and deep learning approaches. We have tried different machine learning approaches namely Logistic Regression, shallow Neural Network and Deep Neural Network Model. We have also given different settings and different functions into our Deep Neural network Model. It turns out that the NN model tends to work better than other approaches for the dataset generated. It is also known that as the dataset grows a neural network tends to work better.

A brief description of our approcahes is geven below:

Logistic Regression:

Logistic regression is a very popular machine learning technique. Logistic regression is used when the dependent variable is categorical. The goal of logistic regression is to minimize the error between its predictions and training data. if given x feature vector,

$$Y = P(y = 1|x),\ where\ 0 <= Y <= 1$$

Logistic regression uses a sigmoid function to predict the output. The sigmoid function returns a value from 0 to 1.

$$s = \sigma(w^T \times x + b) = \sigma(z) = \frac{1}{1 + e^z}$$

The implementation of Multiclass classification follows the same ideas as the binary classification.In multi-class classification, we have more than two classes. In our work, we have 6 generes i.e 6 classes.

For multiple classes, logistic regression uses a softmax function to predict the output. Applying the softmax function, it gives an array of result. These represent the probability for the data point belonging to each class. The sum of all the values in the result array is 1.

$$a_i = \frac{e^{z_i}}{\Sigma_{k=1}^{c} \times e^{z_k}}\ where\ \Sigma_{i=1}^{c} a_i = 1$$

In our model, we have implemented Logistic Regression with our dataset where to optimze the model we have used Stochastic Gradient Descent and to calculate the loss, we have used Cros entropy Loss Function. Learning rate was 0.1 , batch size was 32, number of iterations were 1500.

Shallow Neural Network Model:

Shallow neural networks consist of only 1 or 2 hidden layers. The figure below shows a shallow neural network with 1 hidden layer, 1 input layer and 1 output layer.

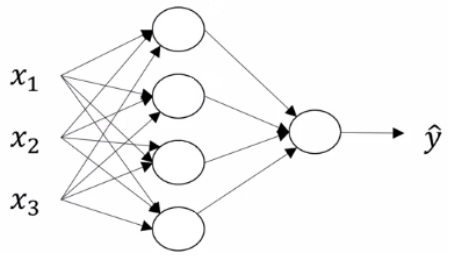

Figure 4.10: Structure of Shallow Neural Network

In our model, we have implemented our model with Shallow Neural Network with ReLU

activation function. Learning rate was 0.1 , batch size was 32, number of iterations were 1500. We have used Stochastic Gradient Descent and to calculate the loss, we have used Cros entropy Loss Function. We hav tuned our model with

Deep Neural Network Model:

According to wikipedia, "Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input." When we hear the name Neural Network, we feel that it consist of many and many hidden layers.

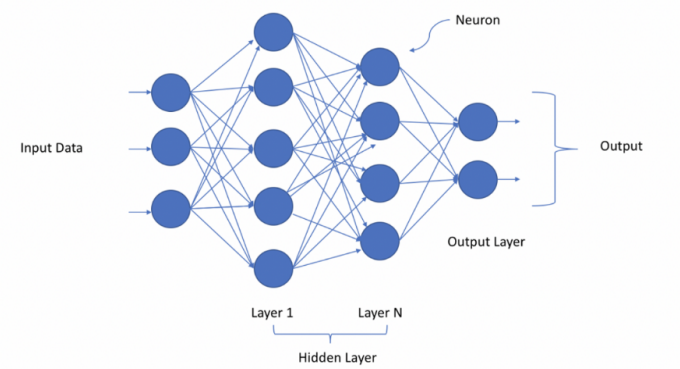

Figure 4.11: Structure of Deep Neural Network

Neuron:

The neuron is the atomic unit of a neural network. Given an input, it provides the output and passes that output as an input to the subsequent layer.A neuron can be thought of as a combination of 2 parts: The first part computes the output , using the inputs and the weights. The second part performs the activation on the output to give out the final output of the neuron.

Hidden Layer:

The hidden layer comprises of various neurons, each of which performs the above 2 calculations.

We have tried different layered models and choose one to propose that gives the best accuracy.We have also given different hyperparameter settings in each different layered model. In order to set up the model class, we need to initialize the model type and declare the forward pass. We initialize our model with this linear layer: **torch.nn.Linear** which applies a linear transformation to the incoming data:

$$y = W^T \times x + b$$

Activation functions:

We have tried different activation functions in diferent layers/ settings. We need these activation functions because without them, the output signal becomes a simple linear function. A non-activated neural network will act as a linear regression with limited learning power. Multilayered deep neural networks can learn meaningful featıures from data by using activation functions. They allow backpropagation because they have a derivative function which is related to the inputs.

Different activation functions that we have used in different settings and layers are - ReLU , Leaky ReLU, Hyperbolic Tangent, Sigmoid, Step, Linear function etc. ReLU function is a general activation function. ReLU function should only be used in the hidden layers. Leaky ReLU function is the best choice if we encounter a case of dead neurons in our networks. After all the settings, the model that we have proposed in our work, has been used ReLU activation function in 3 layers and Leaky ReLU activation function in one layer. ReLU that has a small positive slope in the negative area, so it does enable backpropagation, even for negative input values. This leaky value is given as a value of 0.01 if not +ve.

1. **ReLU activation function:** It is non-linear and computationally effecient. It allows the network to converge ver qickly. $\text{ReLU}(x) = (x)^+ = \max(0, x)$

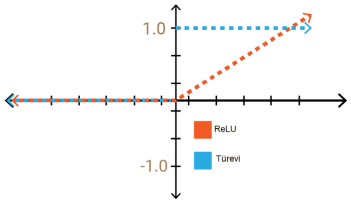

Figure 4.12: ReLU Activation Function

2. **Leaky ReLU activation function:** It is the variation of $\text{LeakyReLU}(x) = \max(0, x) + negativeslope * \min(0, x)$

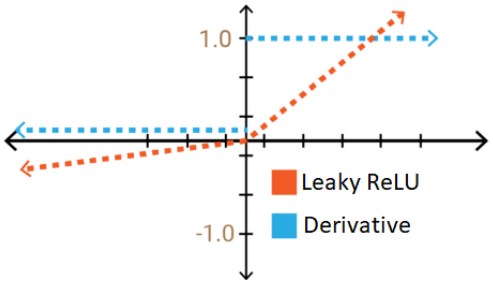

Figure 4.13: Leaky ReLU Activation Function

4.3.4 Instantiate Model Class

The model class that defined earlier needs to be instantiated to construct the model. We construct the model by setting out the parameters like- input dimensions, output dimensions, and a number of nodes per hidden layer. Input dimension is the number of features we extract in the earlier step. Earlier, we have extracted 9 features with 40 dimensions. As we have considered mean and variance of each feature and taken multiple MFCCs, we have 40 dimensions to consider. So, our input dimension is 40. And out dimension is the number of genres we classify from the dataset which will be 6.

4.3.5 Construct Loss and Optimizer Class

Optimizer:

While we are training our model, we need to change the weights and other model parameters to minimize the loss function. So that, our classifications become as correct and optimized as possible. To optimize We use an optimizer that connects the loss function and the hyperparameters and update the model.

There are various types of optimizers with their own qualities. For example - Stochastic Gradient Descent, Adagrad, RMSprop, Adam etc. we have used Stochastic Gradient Descent as our optimizer in our proposed model.

Stochastic Gradient Descent (SGD) is an iterative method for optimizing an objective function. n high-dimensional optimization problems this reduces the computational burden, achieving faster iterations in trade for a lower convergence rate. It starts at the initial point and take a step in the steepest downhill direction after each iteration. It will try to reach to the global optimum or somewhere near to the global optimum.

Loss Function:

In supervised machine learning algorithms, we want to minimize the error for each training example during the learning process. This is done using some optimization strategies like gradient descent. And this error comes from the loss function.The loss function is the guide to the an optimizer, telling it when it's moving in the right or wrong direction.

There are various loss functions like- Squared error Function, Cross entropy loss Function etc. We have used Cross Entropy loss function in our proposed model.

Cross Entropy Loss function is a measure of the difference between two probability distributions for a given random variable or set of events. It is also related to and often confused with logistic loss, called log loss.

4.3.6 Train the Model: Forward, Loss, Backward, Step

After constructing loss and optimizer class, we have trained our model. We calculate the total number of epoch from the length of the training set, number of iterations and batch size. We train our model in range of the number of epoch. We also calculate the loss and show it into graphs. We have calculated the accuracy, precision, recall, f1 score and confusion matrix here and show it into graph. We have saved every model as pickle file.

Chapter 5

Experiments

5.1 Dataset

Our work is based on a dataset consisting of **690** Bangla songs, divided into six genres: পল্লীগীতি (Polligeeti),দেশাত্মবোধক (Deshattobodhok), চলচ্চিত্র (Cholochitro), নজরুলগীতি (Nazrulgeeti), রবীন্দ্র সংগীত (RabindraSangeet), আধুনিক বাংলা (Adhunik Bangla) encoded in the mp3 format We have extracted the songs in a specific folder for each genre of music and saved the file link, name of the file, and the genre label in a separate CSV file. For 6 classes of genres, we have labeled the song with a numeric value starting from 0 to 5. We have annotated the data automatically while collecting the files for each genre.

Label	Genre	No of songs
0	আধুনিক বাংলা	150
1	পল্লীগীতি	184
2	দেশাত্মবোধক	37
3	চলচ্চিত্র	116
4	নজরুলগীতি	53
5	রবীন্দ্র সংগীত	150
	Total:	**690**

Table 5.1: Number of songs Per Genre

This tabulated data shows the number of songs we have per class. The comparison between the numbers of instances in each class is shown by the bar graph given in the next figure.

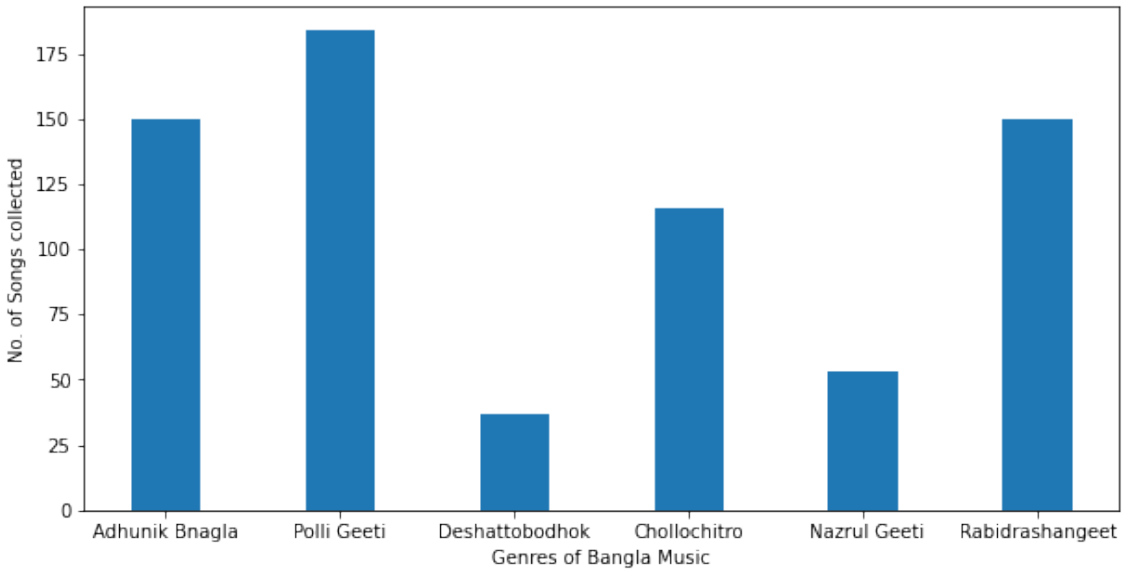

Figure 5.1: Bar Graph of Number of Instances Per Genre

We have split our dataset into 90:10 ratio. Every time, we have trained our model, we have shuffled the dataset for better accuracy. For example, when we have segmented the dataset, there are 621 data in train set and 69 data in test set.

	Label 0	**Label 1**	**Label 2**	**Label 3**	**label 4**	**Label5**	**total**
Train data	136	169	33	103	46	134	621
Test data	14	15	4	13	7	16	69

Table 5.2: Number of train and test data in each class

In the CSV file named 'audioscrape.csv' we have saved the file link, name of the file, and the genre label of the raw data. After that when we have extracted the features turn by turn , we have saved all the numeric values finally in the CSV named 'AllDetails6.csv'.

Document Link	Name	Label
http://www.mp3haat.net /mp3/bengali/ amar-mon-mane -na-asha.mp3	amar-mon-mane-na-asha.mp3	0
//www.music.com.bd/download/ Music/P/Polli Geeti /Polli Geeti - Aaha Lal Pagri (music.com.bd).mp3	Aaha Lal Pagri (music.com.bd).mp3	1
//www.music.com.bd/download/ Music/D/Deshattobodhok Gaan/Ek Nodi Rokta Paria (music.com.bd).mp3	Ek Nodi Rokta Paria (music.com.bd).mp3	2
//www.music.com.bd/download/ Music/O/OST/Hridoy Vanga Dheu/S. I. Tutul And Samina Chowdhury- Ekbar Bhalobashi Bolte (music.com.bd).mp3	S. I. Tutul And Samina Chowdhury - Ekbar Bhalobashi Bolte (music.com.bd).mp3	3
http://www.mp3haat.net/ mp3/bengali/ Dur-Dipo-Basini_ Asha-Bhosle.mp3	Dur-Dipo-Basini_Asha-Bhosle.mp3	4
http://www.mp3haat.net/ mp3/bengali/Lukale-Bolei-Khuje -Bahir-Kora_Rezwana- Choudhury-Bannya.mp3	Lukale-Bolei-Khuje-Bahir-Kora _Rezwana-Choudhury-Bannya.mp3	5

Table 5.3: Samples of each class in CSV

5.2 Evaluation Metric

In order to measure the quality of the mathematical or Neural Network process, evaluation metrics are used. For any project, evaluating NN models or algorithms is essential. There are several types of evaluation metrics available to evaluate a model. These include classification accuracy, logarithmic loss, confusion matrix etc. We compute accuracy, precision, recall, and F1 score to compare the performance of our model in different settings.

- **Accuracy**: Accuracy is the quintessential classification metric that is straightforward and true to its meaning. It is conveniently suitable for binary as well as a multiclass classification problem. The equation for accuracy:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

Accuracy is the ratio of accurate results among the total number of cases tested. Accuracy is the right choice for evaluating classification problems that are well balanced and not skewed or with no class imbalance. Here, accuracy is simply defined as the percentage of correct word suggestions, contextual error detected and Murad Takla sentence found by our model concerning the test dataset.

- **Precision**: Precision is the proportion of positive cases that are correctly identified. The equation for precision:

$$Precision = \frac{(TP)}{(TP + FP)}$$

Precision is the right choice of evaluation metric when we want to be very sure of our prediction. Here, precision is calculated by dividing the total number of accurate word suggestion/contextual spelling error detected by the total number of positives cases identified by our model for word suggestion and spelling error detection.

- **Recall**: A recall is the proportion of actual positive cases that are correctly identified. The equation for recall:

$$Recall = \frac{(TP)}{(TP + FN)}$$

A recall is the right evaluation metric when we want to capture as many positives as possible. Total number of correct words suggested and errors detected divided by the total number of true positive cases and false negative cases produce the recall.

- **F1 score**: F1 Score is the harmonic mean of precision and recall values for a classification problem. The F1 score is between 0 and 1 and is the harmonic mean of precision and recall. The equation for the F1 score:

$$F1Score = \frac{2 * ((precision * recall)}{(precision + recall))}$$

F1 score sort of maintains a balance between the precision and recall for our classifier. If our precision is low, the F1 is low, and if the recall is low again, our F1 score is low. It can be very helpful if the data is imbalanced.

5.3 Result

A loss graph for every 300 iteration shows that the loss is gradually decreasing with almost every iteration which is a good sign shown in figure 5.2. And the confusion matrix is shown in figure 5.3 for our proposed model which helps us to visualize the correct predicted values which are present in the diagonal position of the matrix.

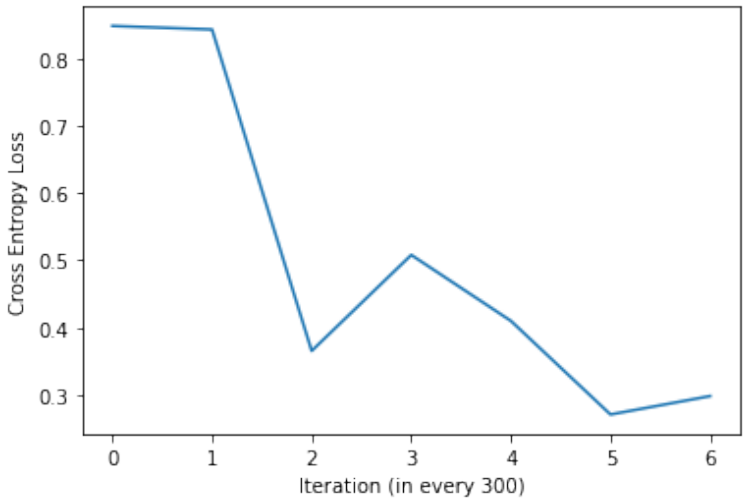

Figure 5.2: Loss in every 300 iteration of Our Proposed Deep Neural Network Model

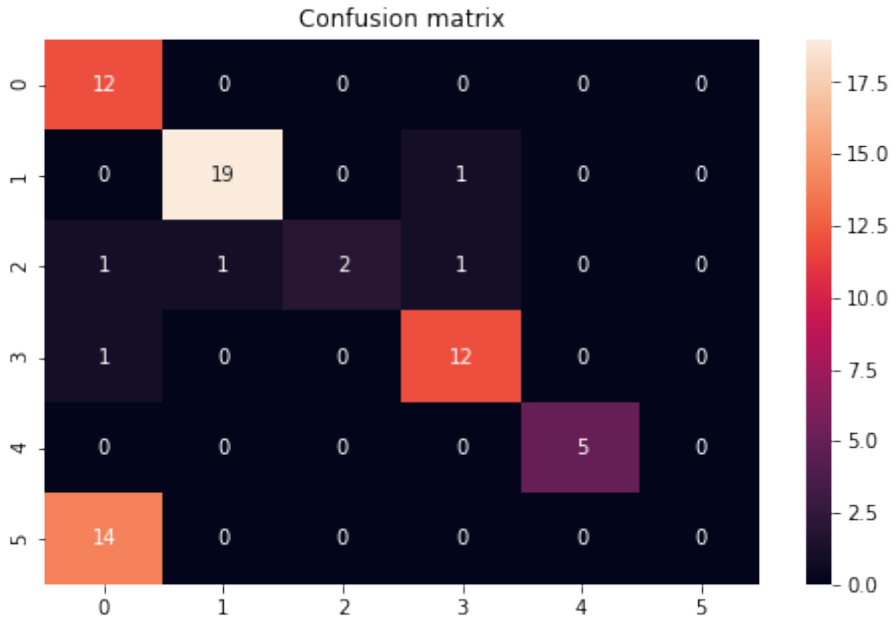

Figure 5.3: Condusion Matrix of Our Proposed Deep Neural Network Model

Model	Accuracy	Precision	Recall	F1 Score
Logisctic Regression	47.826	54.749	47.193	50.691
Shallow Neural Network	53.623	63.619	53.914	58.366
Deep NN Model 1	65.217	58.889	65.427	61.986
Deep NN Model 2	65.217	67.418	64.011	65.670
Deep NN Model 3	68.116	62.843	65.741	64.259
Deep NN Model 4	69.565	67.964	67.976	67.970
Proposed Deep NN Model 5	**72.464**	**70.595**	**71.218**	**70.905**

Table 5.4: Result Analysis Based On Different Models

We have presented our results in 5.4 as we can see the accuracy of our proposed Model-5 increases performance to a great extent comparing it to the machine learning logistic regression technique. It gives an accuracy of 72.46% and the highest F1-score for the classification of the 6 genres of music.

Model	No of Hidden Layers	Activation Function	Batch Size	No of Iterations	No of Epoch	Learning Rate	No of Nodes Per Hidden Layer	Accu -racy
Logistic Regression	0	Softmax	32	1500	69	0.1	0	47%
Shallow Neural Network	1	ReLU	32	1500	69	0.1	50	53%
Deep NN Model 1	4	3 Tanh, 1 ReLU	64	1500	139	0.1	32	65%
Deep NN Model 2	3	ReLU	64	1500	139	0.1	150	65%
Deep NN Model 3	3	Leaky ReLU	45	1500	97	0.1	100	68%
Deep NN Model 4	3	3 ReLU	100	1500	217	0.07	50,30	69%
Proposed Deep NN Model 5	4	3 leaky ReLU, 1 ReLU	32	2500	115	0.1	28	72%

Table 5.5: Performance of Ablation Experiments

Looking at the comparisons of our implemented models in table 5.5 we can see that Model-1 provides an accuracy of 65.217% when we use Tanh activation function with ReLU activation in the 4-layers of the model, which is equivalent to Model-2 accuracy, which is a 3-layer neural network using ReLU activation function, presented in the table 5.5.

Model 3 and 4 both are 3-layer neural network models but it is seen that the difference in the hyperparameters makes a difference in the performance. Model 4 provides a different number of hidden layers. For the first 3-layer, it passes 50 hidden nodes and for the 4^{th} layer it uses 30 hidden nodes. Moreover taking a learning rate of 0.07 helps the model to converge and learn the features better. Tuning the hyperparameters can make a major difference in the results. There is no fixed way to tune the settings rather than trying out different ones. But analyzing the better activation function and optimizer function can increase the overall performance.

Model-5, the proposed model that we used, is a 4-layer neural network with 115 epochs and a batch size of 32. Leaky ReLU and ReLU with 28 hidden layers were used as activation functions. This model shows how activation functions can influence predicted outcomes. Comparing it to the logistic regression we can say that almost all the parameters are identical however due to the addition of hidden layers, the model learns the data more and predicts better results. It is fair to assume that deep neural network works well for audio genre classification.

Chapter 6

Conclusion

In this project, we have presented an approach to classify different music genres based on deep neural networks and decomposition of the music signals in the time and frequency domain. 9 feature vector sets were created for each mp3 file segmented to 120 seconds. The choice of feature extraction was quite important as they determine how accurately a sequence of the music is presented in numerical dimensions. Our proposed 4 layer NN model outperforms the traditional logistic regression method. Adding layers has increased the performance measures as we have shown the comparative results in table 5.5. The hyperparameters played an important role in getting better results and improving the accuracy of our model. We used accuracy, precision, recall, and F1-score to evaluate our project.

The model might work better if more audio features can be added to our training set and on top of that if the number of audio files was more in that case our model might have worked better. Feature extraction techniques can be improved by widening the range of our domain for feature collection.

References

[1] D. J. Levitin, This is your brain on music: The science of a human obsession. Penguin, 2006.

[2] D. Huron, "Is music an evolutionary adaptation?," Annals of the New York Academy of sciences, vol. 930, no. 1, pp. 43–61, 2001.

[3] T. E. of Banglapedia, "Music," 2015.

[4] D. Gerhard, Audio signal classification: History and current techniques. Citeseer, 2003.

[5] M. C. Darji, "Audio signal processing: A review of audio signal classification features," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 2, pp. 227–230, 2017.

[6] M. A. Ali and Z. A. Siddiqui, "Automatic music genres classification using machine learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 8, pp. 337–344, 2017.

[7] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 282–289, 2003.

[8] A. Bhowmik and A. E. Chowdhury, "Genre of bangla music: A machine classification learning approach," AIUB Journal of Science and Engineering (AJSE), vol. 18, no. 2, pp. 66–72, 2019.

[9] M. A. Al Mamun, I. Kadir, A. S. A. Rabby, and A. Al Azmi, "Bangla music genre classification using neural network," in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 397–403, IEEE, 2019.

[10] music.com.bd, "www.music.com.bd/," 2014-2017.

[11] mp3haat, "Music," 2012-2021.

[12] L. Richardson, "Beautiful soup documentation," Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018], 2007.

[13] W. contributors, "Pitch," 2021.

[14] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291–1303, 2017.

[15] W. contributors, "Zero-crossing rate," 2021.

[16] W. contributors, "Spectral centroid," 2020.