Received 23 September 2022, accepted 8 October 2022, date of publication 12 October 2022, date of current version 17 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3213818

RESEARCH ARTICLE

BMNet-5: A Novel Approach of Neural Network to Classify the Genre of Bengali Music Based on Audio Features

KHAN MD. HASIB[1], (Member, IEEE), ANIKA TANZIM[1], JUNGPIL SHIN[2], (Senior Member, IEEE), KAZI OMAR FARUK[3], (Member, IEEE), JUBAYER AL MAHMUD[4], AND M. F. MRIDHA[5], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka 1208, Bangladesh
[2]Department of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu-shi, Fukushima 965-8580, Japan
[3]Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh
[4]Department of Computer Science and Engineering, University of Dhaka, Dhaka 1000, Bangladesh
[5]Department of Computer Science, American International University-Bangladesh, Dhaka 1229, Bangladesh

Corresponding authors: M. F. Mridha (firoz.mridha@aiub.edu) and Jungpil Shin (jpshin@u-aizu.ac.jp)

ABSTRACT Music genre classification (MGC) is the process of putting genre labels on music by analyzing the sounds or words. With the rapid growth of music data repositories, MGC can be used in a lot of ways to organize and manage music recommendation systems, advertising, and streaming services. But there have been a lot of works on classifying English music using different statistical and machine learning methods, but there hasn't been much progress in classifying Bengali music. Also, Deep Learning (DL) methods have been used in a few important ways to classify different types of music. The content and uniqueness of Bengali music make it much more interesting. Also, there is still a lot to learn about how to use the DL approach in Bengali music. So, Bengali music genre classification is a pretty new area of research in the field of Deep Learning. In this paper, we developed a unique technique called BMNet-5 to perform a multiclass classification of Bangla music genres such as "Bangla Adhunik," "Bangla Hip-Hop," "Bangla Band Music," "Nazrulgeeti," "Palligeeti," and "Rabindra Sangeet." We show the effectiveness of the suggested technique by extracting features from a dataset of 1742 Bangla music pieces and evaluating the automated classification judgments. The proposed BMNet-5 is based on a neural network designed to predict music genre from audio inputs. Our suggested model outperformed the corresponding previous research with an accuracy of 90.32%. The BMNet-5 model is then tested for performance consistency using K-fold cross validation with varying k values. Finally, we use the suggested model to train the interpretable SHAP model for all the genre of the Bangla music dataset, and the development of an explainable outcome may have a significant advantage.

INDEX TERMS Bangla music genre classification, deep learning, feature extraction, audio classification, explainable AI, SHAP.

I. INTRODUCTION

Music is a unifying, cross-cultural activity that exists in almost every culture around the world. According to a survey, Americans spend more money on music then they do on prescribed pharmaceuticals, owing to the present global music surge [1]. Background and purposeful music listening add up to more than 5 hours of music viewing every day for the average American [2]. This study reveals the widespread popularity of music and the influence it plays in everyday people's lives.

Although much research has been done on English music genre classification, there is still a lot of room for Bengali music genre classification using Machine learning

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

and Deep learning classification techniques. Bangladesh has a culture that encompasses music with a sense of its heritage and origin, as evidenced by its long history. Music genre labels are used to classify and describe the enormous world of music based on shared tradition or set of conventions. The start of music in Bangladesh starts long back with the Buddhist monks or saints between the 7th to 12th centuries composing a music genre named Charyacharyavinishcha, generally referred to as charyagiti [3]. Bengali music is split into a wide range of genres and for this research, we have focused on 6 genres namely: Nazrulgeeti, Rabindra sangeet, Polligeeti, Band music, Hip Hop, Adhunik Gaan (Modern music). For each genre, there are about 250 - 350 songs (.MP3 files) which provides a wide number of the dataset to learn about each genre. Classification of music itself can be a critical task and Bengali music has a comprehensive set of tones and diversities which makes the genre classification more challenging. However, if we extract and investigate several characteristics in order to conduct an analysis of the connection sequence across them, we can see a noteworthy structure in the same genre. As feature extraction and classifier learning can have a significant impact on classification system performance [4], we focused on extracting nine features in this study to analyze the audio files namely: Zero-Crossing Rate (ZCR), Spectral Centroid, Mel Frequency Cepstral Coefficients(MFCC), Spectral Roll-off, Chroma Frequency, Spectral Bandwidth, Spectral Flux, Pitch and Tempo. Before feature extraction, each MP3 file read as a time series has been sliced to 120 seconds to reduce the computational time and mainly collect only the prominent features of the entire audio. These feature extraction techniques resulted in a better performance leading us to determine the genres more accurately.

Automatic music genre classification is commonly used to produce and classify playlists of the same genre. When there is a large number of unlabeled songs, manually annotating the genre and recommending them to consumers becomes tough. As a result, well-known applications like *Napster*, *Spotify*, and *Soundcloud*, apply genre classification to recommend music to their users. Deep Neural Network, CNN has been popularly used for classifying music [5], [6], [7] over the years. In our research, we proposed a model BMNet-5 which is a deep neural network with 5 layer neuron to extract Bangla music genre. It has also experimented with Logistic Regression (LR), Convolutional neural networks (CNN), Deep NN model 1,2,3,4. After that, the proposed BMNet-5 model is put through a performance consistency test with the help of K-fold cross validation utilizing a range of different k values. In conclusion, we make use of the proposed model in order to train the interpretable SHAP model for a specific category of our dataset. The creation of an explainable result may have a substantial benefit, as it may allow for more accurate predictions.

The remainder of the study is arranged as follows: Section II describes our research aim and scope. Section III analyzes similar papers on various forms of news classification in Bangla and other languages. Section IV outlined the research methodology, including datasets and suggested approaches. Section V displays the analysis of the results. Finally, in Section VI, this task is completed, and future directions are provided.

II. RESEARCH AIM AND SCOPE

This research aims to extract the key features from the audio files and develop a robust classifier BMNet-5 that can classify the 6 genres of Bengali Music as accurately as possible. The following are the principal contributions of this study:

- Extracting nine various features from each bengali song of the dataset to classify the genres and find connections between the sound signals.
- Proposed a Deep NN Model 5 which we called BMNet-5, is a 5-layer modified neural network with 187 epochs and a batch size of 50 to predict the genre and to do a multiclass classification of Bangla music.
- The proposed BMNet-5 is evaluated using various performance metrices: Accuracy, Precision, Recall, and F1-Score.
- Evaluated the effectiveness of the proposed BMNet-5 model by generating confusion matrix, ROC curve and PR curve.
- Checked the performance consistency of the proposed BMNet-5 model rigorously, the model is trained multiple times using K-fold cross validation with different K values.
- Explainable AI technique SHAP is used to explain the prediction of some genres using proposed BMNet-5 in terms of interpretability.

III. RELATED WORKS

In regard to making predictions and classifications, artificial intelligence and machine learning algorithms have gained a lot of traction in recent years [8]. Development in music genre classification has been ongoing with the increasing amount of music every year. Below are some related work using deep learning algorithms and machine learning models.

It has been established that neural networks function efficiently with time series data [9], and CNN for music classification lately gained popularity due to its high accuracy and predictions. A neural network is a system primarily made up of intricately linked components that process data by responding actively to input datasets. [10] Considering the multi-scale time-frequency information and using CNN architecture on popular datasets such as GTZAN, Ballroom, and Extended Ballroom Liu et al. [11] achieved 93.9% 96.7% 97.2% classification accuracies. Although the number of tracks for each genre in the GTZAN and Ballroom dataset is around 100, using the broadcast module (BM) with dense convolution layers a good performance has been curated. CNN has also been used as a classifier by Vishnupriya et al. [12] where they extract only two features Mel Frequency spectrum and Mel Frequency Cepstral Coefficient (MFCC) from the audio files and train and classify the genres. Using a dataset of 1000 audio files and 100 files per genre, and accuracy of 76% for Mel Spec and 47% for MFCC is received. Using more features could have resulted in better performance. Furthermore, the research by Hareesh Bahuleyan et al. [13] presented a method for automatically classifying music by assigning tags to songs in the user's library. Using Convolutional Neural Network and machine It investigates both Neural networks and classic Machine Learning techniques at first, CNN is trained end to end by extracting features from audio signals. After extracting features such as Mel-frequency Cepstral Coefficients (MFCC), Chroma Features, and Spectral Centroid (XGB), the second method employs several Machine Learning techniques such as Logistic Regression, Random Forest, Support Vector Machines (SVM), and others. They concluded that the VGG-16 CNN model provided the highest accuracy after analyzing the two approaches separately. The optimal model with 0.894 accuracy was produced by developing an ensemble classifier of VGG-16 CNN and XGB. In another paper [14] evaluation was done with multi label feed-forward Deep neural network on a smaller dataset consisting of only 400 songs. Similar feature extraction methods were put into action and an accuracy of 97.8% was received. Increasing the number of dataset and audio files for each genre could have showed more realistic result and the actual scenario of the classifier.

Machine learning algorithms like K-Nearest Neighbour(KNN) and Support Vector Machine (SVM) showed prominent results in genre classification when experimented with the GTZAN dataset having 30-second audio files [15]. Using MARSYAS, a public software framework for computer audio applications MFCC feature is extracted and trained with KNN and SVM. SVM model gives an accuracy of 77% which was greater than the accuracy received from the KNN model. Jaime et al. [16] experimented with different machine Learning paradigms: Probabilistic Graphical Models (Naive Bayes), Feed-forward and Recurrent Neural Networks, and Support Vector Machines (SVMs) to classify genres of retrieved from web applications. In earlier research, Li and Ogihara [17] recommend Daubechies Wavelet Coefficient Histograms (DWCHs) as a set of skills for categorizing music types. They presented DWCHs as another feature extraction approach for music genre grouping in their research. They looked into the relative merits of various feature extraction and grouping procedures, as well as the sequence in which different characterization tactics should be applied to specific feature sets. Using the DWCH method they extracted 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. Three different reduction methods are used to extend SVM for multi-class: pairwise, one against-the-rest, and multi-class objective functions. For experiments involving SVMs [18], they tested them with linear, polynomial, and radius-based kernels. Moreover, they also implemented Gaussian Mixture Models for each music genre. In comparison to Beat or Pitch the performance with FFT and MFCC was much better. In the case of the models used they have shown a comparison with different features and modeling techniques. Overall they looked at the relative merits of various feature extraction and grouping procedures, as well as the sequence in which various characterization tactics should be applied to specific feature sets.

Low-level language such as Bangla Music Genre Classification (MGC) still has a lot of scopes based on the uniqueness of the songs. A Gated recurrent unit(GRU) consisting of 3 layers has shown the notable result of 80.4% accuracy and 80.6% F1-score when experimented with by Moumita et al. using a dataset of 6 genres and a total of 2944 Bengali songs. Using only one MFCC feature and using other Neural network algorithms like LSTM, Feed-Forward Neural Network, and GRU(1,2,3) layers this research has been executed. A similar approach of using neural networks was undertaken by Md. Afif et al. [19] where they used 5 layers of neural networks to experiment with 8 feature vectors and. 6 genres of Bengali music has been classified with a dataset of 1742 songs. Despite having a variety of feature vectors the highest accuracy they received was 74%. Other machine learning models were used to train the dataset, for instance, SVM, logistic regression, linear regression, and kNN. Genres of music can often depend upon the mood and in the research made by Deepti et al. [20] they have focused on classifying four genres of Hindi music - Classical, Folk, Ghazal, and Sufi based on positive arousal, negative arousal, positive valence, and negative valence by considering arousal and valence as parameters. The correlation coefficient is used to determine the linear dependence between the characteristics of two signals. [21]. Keeping the correlation in consideration they extracted timbral feature, rhythmic feature, and time-frequency features and used different parameters to evaluate their proposed models. A dataset of 1000 songs of 30s each is used to determine and demonstrate the emotions based on a song. On the whole, an annotation process is made for classification keeping the correlation in mind and trained with different models like hash tag tree generation, CNN, SVM, and kNN out of which hash tag tree generation outperforms others. The mood of songs can make a clear distinction between the genres as seen in the paper by Patra et al. [22] where Lib-SVM and Feed-forward neural networks were used to develop frameworks for categorizing emotions based on acoustic, lyric, and a combination of the two. F-measure values were noted by the authors to be 0.751 and 0.835. In another work, by S. Jothilakshmi et al. [23] machine learning model Gaussian mixture model(GMM) outperformed k- nearest neighbor (kNN) model when the classification of 5 genres of Indian music was evaluated based on features extraction. Each genre consisted of 100 audio files of 30sec. Features like skewness, flatness, Spectral irregularity, perceptual features, and others were extracted and experimented with. An accuracy of 91.25% was found which took place due to the feature combination although it was evaluated comparatively on a smaller dataset with only 500 songs. The right combination of features can make a positive impact on the accuracy of genre selection as seen in the research of Betsy et al. [24]. Compared to the conventional MFCC feature giving an accuracy of 84.21% combination of spectral roll-off, spectral flux, spectral skewness, and spectral kurtosis, combined with fractional MFCC features, outperforms all other feature combinations showing an accuracy of 96.05%.

Reviewing all the papers we could realize that there were few works related to Bangla music genre classification and if proper work is conducted using Deep neural networks and machine learning models it may perform well in the classification of music genres. However, feature extraction and its combination play a vital role in these classifiers. The correct combination of features and a clear understanding of the tunes, and acoustics of the songs can improve the accuracy and evaluate the models with precision.

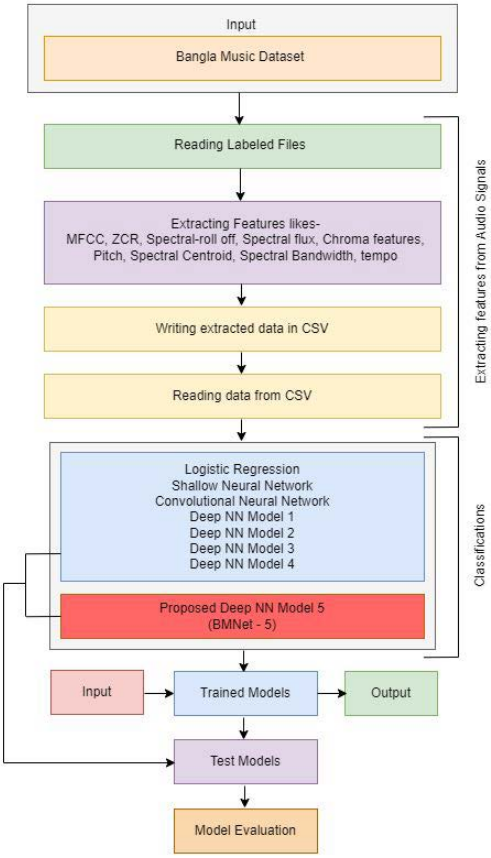

FIGURE 1. Workflow diagram of the work.

IV. PROPOSED METHODOLOGY

The main goal of this work is to classify and predict from the Bangla Music dataset with the help of our proposed model BMNet-5 by looking at the feature extraction of audio signals. Fig. 1 shows the process of how this work is done. First, some steps are taken to prepare the raw music files that have been collected from the existing dataset. Then, features are taken from the data that has been prepared, and the feature vector is sent to the training and validation step. The model is then judged by how well it predicts the test set.

A. DATASET

A dataset is the most important thing you need in order to build a model. We did all of our tests on raw data from 1742 Bangla music brought in by Mamun [19]. The data we have worked with that includes audio files from YouTube and other places for this research. The audio files by filtering artists, bands, playlists of uncategorized music, and other things. About 7.3 GB (Gigabytes) of raw MP3 files have been used. After getting all of these MP3 files, the data checked each one by hand to see what genre it was in. Here, the dataset was prepared by Mamun [19], each raw audio files were opened to check if they were in an accessible condition. In the paper [19], they manually checked every file and determined their genre on the basis of data collection process from Youtube and other sources. The data is kept different types of music in their own directories and put their files in those directories and also get rid of bad or unnecessary files. After the data was cleaned up, we used a Python library called librosa [25] to pull out different parts of the audio files.

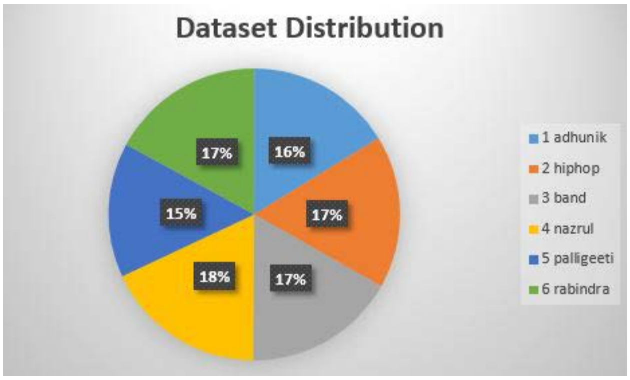

FIGURE 2. Distribution of the dataset.

To make the time it takes for audio files to load faster, we only used parts of the files that were a certain length, starting at 20 seconds and going up to 100 seconds. We figured that none of our songs are shorter than 120 seconds. Most of the songs are about 240 (4 minute) seconds long, on average. So, it won't change anything when reading features from audio files. We used a base sample rate of 22050Hz, 100 seconds from each MP3 file (starting at the 20th second and ending at the 120th), and the mono channel to cut down on redundancy. We used a base sample rate because the native sample rate could mess up the classification result since different audio files have different sample rates. The extracted features were then saved in a CSV file. Fig. 2 displays the distribution of the dataset that we will use to train our model. From the figure, we analyzed that nazrul has highest number of instances while palligeeti has lowest no of samples.

Table 1 shows that the CSV file has examples of different classes. This table shows how many songs we have for each

TABLE 1. Number of Bangla songs per genre.

No.	Genre Name	Collection
1	Bangla Adhunik	283
2	Bangla Hip-Hop	295
3	Bangla Band Music	295
4	Nazrulgeeti	312
5	Palligeeti	260
6	Rabindra Sangeet	297
	Total	1742

class. It is essential that we discuss the applications that were used throughout the gathering and preparation of this dataset. The following is a condensed explanation of the utilities that were used in the process of preparing this dataset.

1) IBROSA

This library was brought up previously while we were discussing other topics. A variety of time domain and frequency domain properties could be extracted from an audio signal with the assistance of this library, which then supplied the results as numeric values [26].

2) YouTube-dl

Another Linux package that we made use of to obtain music from YouTube was called youtube-dl. This library offers a Command Line Interface (CLI), which allows users to download audio and video files from YouTube in the format of their choice. We were able to significantly cut down on the amount of time spent collecting data thanks to the functionality on YouTube that allows users to download playlists [27].

3) FFMPEG

ffmpeg is a library that can transform several video codecs into the format of the user's choosing. A backend library called ffmpeg is used by the librosa library. Therefore, librosa to function, we need to have ffmpeg already installed [28].

B. FEATURE EXTRACTION

We are unable to utilize the raw audio segments as input for a neural network since this kind of network only understands numbers though some modified neural network works [29]. We tried to explain that Machine learning does not deal with audio data in raw form. We need to convert the audio data into number or text or images. It was the best approach to convert the audio into spectrograms. That's why we need to transform the audio into some numerical values so that we can use it as input into the model that we have presented. In order to do this, we will need to extract a variety of attributes.

In general, there are two types of audio features:

- The term "physical features" refers to the mathematical measurements that are computed directly from the sound wave, such as the energy function, the spectrum, the cepstral coefficients, the fundamental frequency, and so on. Other examples of physical features include the fundamental frequency and the cepstral coefficients [30].
- Loudness, brightness, pitch, timbre, rhythm, and other aspects of sound are examples of perceptual characteristics. Perceptual features are words that are subjective and are associated with how humans perceive sounds [31].

For the purpose of our study, we are going to extract nine significant variables from our dataset that will assist us in categorizing the different types of music. They are: Zero-crossing rate, MFCC (Mel-Frequency Cepstral Coefficients, Spectral-roll off, Spectral flux, Chroma features, Pitch, Spectral centroid, Spectral Bandwidth and tempo. The proposed features are selected on the importance of each features. For example, ZCR approach is used widely to extract music information because it has higher values for highly rhythmic sounds. Spectral centroid is used to determine the highest frequencies in a sound. This is how all the importance of the features are checked and used for analyzing music genre. We also compute the mean and variance of each feature for every feature, regardless of the technique used to extract the features. Here, numerical values are used as input and there are 29 inputs in total. Mean values are used to get all those 29 inputs. Mean and variance help to get better results and select the correct genre precisely. To begin the process of extracting any feature, we will first read the audio MP3 as a time series. After that, we will isolate the characteristics that are anticipated and compile them into CSV files. The following is a discussion of the strategies that were used in order to extract these features:

1) ZERO-CROSSING RATE(ZCR)

The Zero-Crossing Frequency is the rate at which the sign of a signal changes or the pace at which the move from negative to positive or vice versa. ZCR is a technique for detecting human voices in a music signal [25]. The ZCR approach is rapid and easy for determining whether a speech frame is voiced, unvoiced, or silent. The following equation is used to compute it.

$$ZCR = \frac{1}{H-1}\sum_{h=1}^{H-1} 1_{K<0}\left(C_h C_{h-1}\right) \tag{1}$$

Here, C is a signal of length H and $1_{\mathrm{K}<0}$ [25]. The *librosa.zero crossings()* function is used to find the zero crossing for each frame. The means of several classes are shown in Fig. 3. We may examine the mean ZCR values for several classes in our dataset. Here, *Bangla Band Music* has the highest value, whilst *Rabindra Sangeet* has the lowest.

2) SPECTRAL CENTROID

In digital signal processing, a spectrum is characterized by its spectral centroid. It reveals the position of the spectrum's gravitational core. The centroid is determined using the following formula:

$$\text{Spectral Centroid} = \frac{\sum_{m=0}^{M-1} f(m)Y(n)}{\sum_{m=0}^{M-1} Y(m)} \tag{2}$$

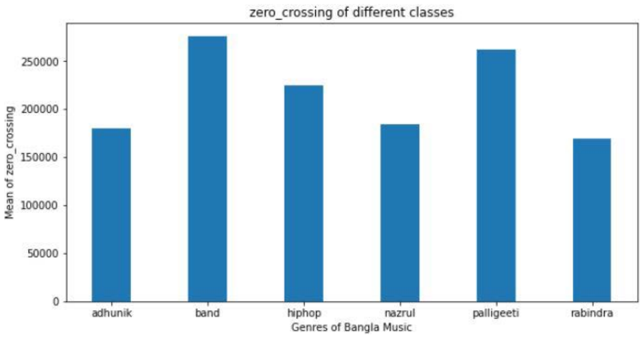

FIGURE 3. Mean ZCR of different classes.

where Y(n) is the weighted frequency value, or magnitude, of bin(the range of values) number m and f(m) is the bin's center frequency [32]. *librosa.spectral centroid()* is used to determine the average spectral centroid value for each value.

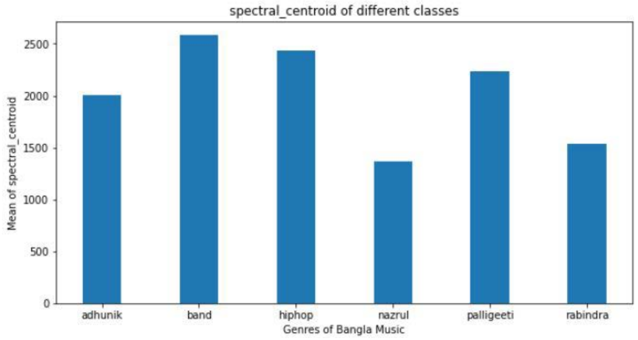

FIGURE 4. Mean spectral centroids of different classes.

Fig. 4 depicts the mean values of Spectral Centroids from various classes in the dataset. The maximum value is assigned to *Bangla Band Music*, while the minimum value is assigned to *Nazrulgeeti*.

3) MFCC(MEL-FREQUENCY CEPSTRAL COEFFICIENTS)

A signal's mel frequency cepstral coefficients (MFCCs) are a limited group of characteristics (typically 10-20) that simply define the overall shape of a spectral envelope [33]. It is often used to describe timbre in MIR. The MFCCs are computed using a mel-frequency spectrogram and the discrete cosine transform (DCT). MFCC can normally extract up to 20 features, but 12-13 features are thought to be excellent for feature extraction, thus we pick 13. *librosa.feature.mfcc* is used to compute the mean and variance for these 13 features.

We picked a sample of mfcc features where from Fig. 5, we analyzed that *Nazrulgeeti* has the maximum value while *Bangla Band Music* has the lowest.

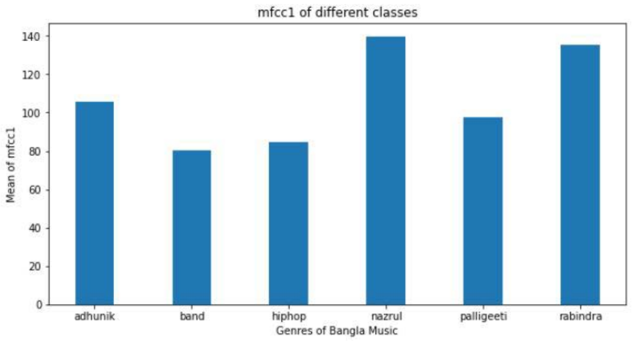

FIGURE 5. Mean MFCC1 of different classes.

4) SPECTRAL ROLL-OFF

When the entire spectral energy falls below a certain frequency, this phenomenon is known as spectral rolloff [32]. The spectral rolloff is calculated using the following equation:

$$\text{Spectral roll-off } = \sum N_g[m] = 0.85 * \sum_{m=1}^{M} N_g[m] \quad (3)$$

The bandwidth under which 85% of the magnitude spectrum is focused is referred to as the spectral rolloff, where $N_g[m]$ is the Fourier transform magnitude at frame t and frequency bin. We use *librosa.feature.spectral rolloff* to determine the mean and variance of the spectral roll off for each file.

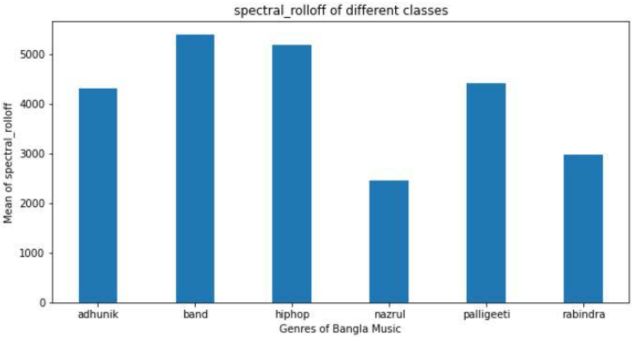

FIGURE 6. Mean spectral roll-off values.

When we look at Fig. 6 and look at the mean values for Spectral Roll-off, we can see that the values for the classes *Bangla HipHop* and *Bangla Band Music* are very close to each other. *Nazulgeeti* has the minimal value.

5) CHROMA FREQUENCY

The way people hear pitch is periodic, meaning that two pitches that are different by one or more octaves are heard as having the same color, or harmonic role (where, in our scale, an octave is defined as the distance of 12 pitches). The main idea behind chroma features is to combine all spectral information about a given pitch class into a single coefficient [34]. One of the most important things about chroma features is that they capture the harmony and melody of music. This is why we are using this aspect to classify the genre.

This is how we figure out the mean value of a chromagram from a music signal of a given signal. We figure this out with

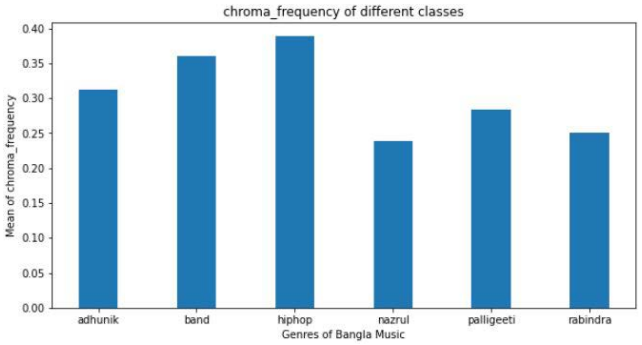

FIGURE 7. Mean chroma frequency values.

the *librosa.feature.chroma stft()* function in the frequency domain. Fig. 7 illustrates the average Chroma Frequencies for each class. Here, *Bangla HipHop* is the highest and *Nazrulgeeti* is the lowest.

6) SPECTRAL BANDWIDTH

Spectral Bandwidth identifies the frequency at which the energy of a spectrum is concentrated. A signal's resolution is determined by its bandwidth [35]. Radiated spectral quantities are not less than half their maximum value in this wavelength range.

$$\text{Spectral Bandwidth } = \left(\Sigma_p V(p)\left(z(p) - z_d\right)^q\right)^{\frac{1}{q}} \tag{4}$$

where z(k) is the frequency in bin p, z(p) is the spectral magnitude, and z(d) is the spectral magnitude at z(p). For each file, we compute the mean and standard deviation of spectral Bandwidth using *librosa.feature.spectral bandwidth*. The order-q spectral bandwidth is calculated using *librosa.feature.spectral bandwidth*.

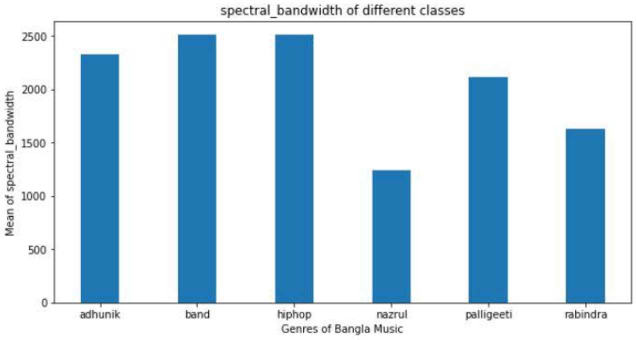

FIGURE 8. Mean spectral bandwidth of different classes.

Spectral Bandwidth values for several classes are shown in a bar graph in Fig. 8. *Bangla HipHop* and *Bangla Band Music* have almost identical peak values here, whereas *Nazrulgeeti* has the lowest mean value.

7) SPECTRAL FLUX

Spectral flux measures the change in the spectrum between two frames [36]. It is calculated as the squared difference between the two normalized magnitudes of the two frames spectral distributions.

$$Fl(i, i-1) = \sum_{k=1}^{Wf_L} (EN_i(k) - EN_{i-1}(k))^2 \tag{5}$$

where $\mathrm{EN_i(k)}$ is the k^{th} normalized DFT coeffcient at the i^{th} frame. From the librosa Python package, *librosa.onset.onset strength* was used to figure out the spectral flux in our research.

8) PITCH

Most of the time, the pitch of an audio signal is used to describe the vibration of a frequency. Many researcher can say that a high frequency audio wave has a high pitch, and a low frequency audio wave has a low pitch [37]. *librosa.piptrack* figures out how high or low the sound files are.

9) TEMPO

In music, tempo means how fast or slow a piece of music is played [38]. For example, a tempo of 60 BPM means that there is one beat every second. We used *librosa.beat.tempo* to find the numerical data for each file based on this feature.

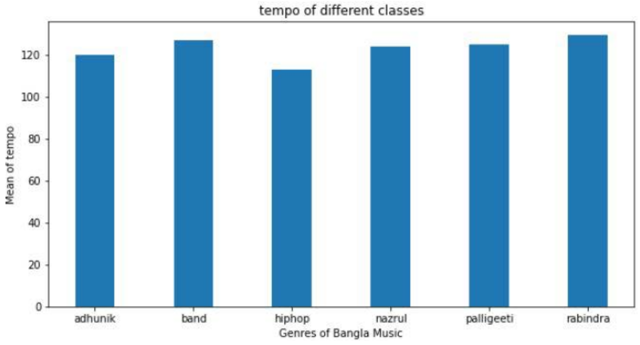

FIGURE 9. Mean tempo values.

Fig. 9 shows how the averages of Tempos from different classes look. We can see that the values for Tempo are almost the same for each class.

C. SPLIT THE DATASET INTO TRAINING AND TESTING SETS

We divided our dataset in a 70:30 ratio for training and testing. We have 1220 data in our training dataset and 522 data in our test dataset after dividing by this ratio. We used *SubsetRandomSampler* and *DataLoader* library to shuffle and partition our dataset. We can get a more generic model by shuffling.

D. BUILDING AND TRAINING MODELS

While splitting the dataset, we used SubsetRandomSampler and DataLoader to shuffle and partition our dataset. The input audio signal must be classified once the feature extraction procedure has been completed. Classification is a necessary step in assigning a label to a specific kind of music.

Different sample classes are separated from each other using a classifier in the feature space. We can get a more generic model by shuffling. We experimented with several machine learning algorithms, including Logistic Regression, Shallow Neural Network, and Deep Neural Network Model along with our proposed BMNet-5 model. We have also given our Deep Neural Network Model varied parameters and functionality. It shows that the NN model performs better than other techniques for the given dataset. It is also well known that as the dataset expands, a neural network performs better. The following is a brief explanation of our methods:

1) LOGISTIC REGRESSION (LR)

LR is a classification approach that assumes one or more independent factors influence the result. Although logistic regression (LR) is primarily a binary classifier, it may be adapted to address multi-class situations using one-vs-rest logistic regression (OVR) or multinomial logistic regression (MLR) [39]. A ridge value was advocated for the log probability calculation in the work of Le Cessie and Van Houwelingen [40]. There are adjustments for categorization purposes. If n cases with m features have k classes, the m*(k-1) matrix points towards component B being computed. The probability for class j with the exception of the class is as in:

$$P_j\left(X_i\right)=\frac{\exp\left(X_iB_j\right)}{\sum_{j=1}^{k-1}\exp\left(X_iB_j\right)+1} \tag{6}$$

The last class has probability as shown in the eqn. 2:

$$1-\sum_{j=1}^{k-1}P_j\left(X_i\right)=\frac{1}{\sum_{j=1}^{k-1}\exp\left(X_iB_j\right)+1} \tag{7}$$

A Quasi-Newton process is employed for discovering enhanced values of m*(k-1) elements to locate matrix B where L is reduced. The matrix B is compressed to a m*(k-1) vector prior to the optimization approach [41].

We utilized Stochastic Gradient Descent to optimize the model and Cross entropy Loss Function to compute the loss in our model, which uses Logistic Regression with our dataset. We trained at a 0.1 learning rate, with a batch size of 32, and iterations totaling 1500.

2) SHALLOW NEURAL NETWORK MODEL

There are just one or two hidden layers in shallow neural networks [42]. The neural network shown in the picture below has three layers: an input layer, a hidden layer, and an output layer.

Fig. 10 depicts the structure of Shallow Neural Network. In our model, we have implemented it with ReLU activation functions are used in our model. The Batch size was 32, the learning rate was 0.1, and iterations totaled 1500. We utilized Stochastic Gradient Descent and the Cross entropy Loss Function to determine the loss.

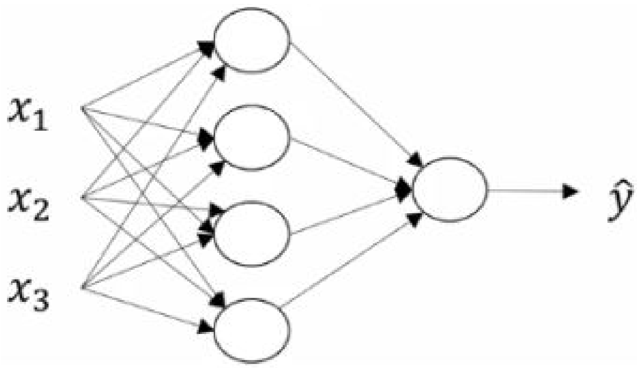

FIGURE 10. Structure of shallow neural network.

3) CONVOLUTIONAL NEURAL NETWORK (CNN)

The initial use of the Convolutional neural network (CNN), a version of MLP influenced by biology, was in digit recognition. The local receptive fields, shared weights, and sub-sampling used by CNN provide some degree of shift, scale, and distortion invariance for the model [43]. Based on the convolution of spectrograms, these principles may be used to music categorization.

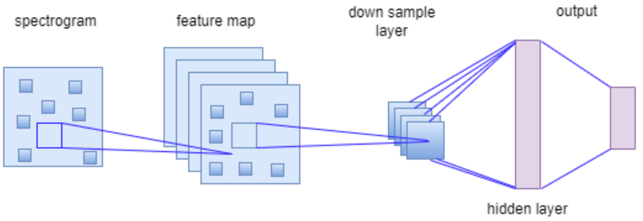

FIGURE 11. Structure of convolution neural network.

When we use these filters to perform spectrogram convolution, we receive different feature maps as illustrated in Fig. 11. These filters may be used to acquire high-level features since various genres will have distinct components. After obtaining the feature map, a sub-sample layer is applied to each feature. The sub-sample layer is important for two reasons. Using a sub-sample layer helps lower the overall sample size. The amount of weights will be enormous if we just utilize the original feature map. To build the model, we define its key features, such as input and output dimensions and the size of each hidden layer's graph of nodes. A feature extraction's input dimension is the total number of features that were initially gathered. We've already taken 9 characteristics over 40 dimensions for extraction. Having taken many MFCCs and accounting for each feature's mean and variance yields a total of 40 dimensions. In this case, 40 is the input dimension. Further, our "dimension" will be the 6 categories into which we divide the dataset. Additionally, the output of a subsample layer may be invariant to the pitch offset and tempo contraction and expansion during translation.

4) DEEP NEURAL NETWORK MODEL

The accuracy of neural networks may be improved over time by using training data. These algorithms are useful in computer science and artificial intelligence after they have been fine-tuned for accuracy, enabling us to categorize and cluster data at a fast rate of speed [44].

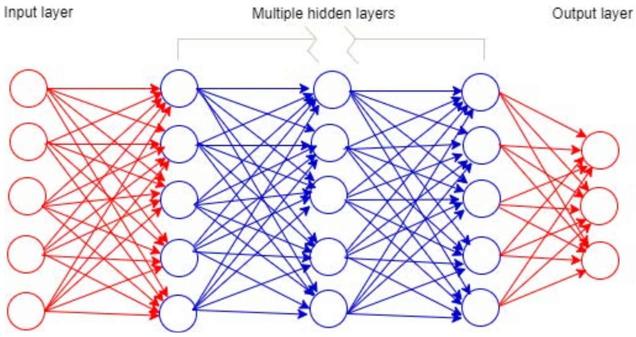

FIGURE 12. **Structure of deep neural network.**

An unstructured data set may be used to train a deep-learning network, which has more input options than a machine-learning network. Training a neural network on a large amount of data is likely to result in a more accurate model. Unlike most typical machine-learning algorithms, deep-learning networks automatically extract features. Fig. 12 depicts the basic structure of deep neural network. It is possible to extract significant characteristics for classification by using DNN, a multi-layer, non-linear model. MLPs, also known as Feed Forward Networks, are the most common kind of neural network. With DNN's deep layer architecture, it's possible to automatically extract and learn features from even the most complicated and high-dimensional of data.

In our experiment, we employed 1 layer, 2 layer, 3 layer and 4 layer of DNN to classify the genre of Bangla music. Fully-connected neural networks, such as the DNN utilized in this study, comprise of an input layer, an output layer, and other layers that are hidden from view. There is a direct correlation between the number of input neurons and the size of the input feature vectors, whereas the number of output neurons is proportional to how many musical styles are being examined. The model's units' inputs and outputs follow the single neuron's fundamental logic.

We have experimented with a number of layered models before settling on the one that provides the most accurate results. Hyperparameters are also specified differently in each layer of the model. The model type must be initialized and the forward pass must be declared before the model class can be set up [45]. Linear layers are used as a starting point for our model. *torch.nn.Linear* transformation is applied to the incoming data in this case:

$$y = W^T \times x + b \tag{8}$$

E. CONSTRUCT LOSS AND OPTIMIZER CLASS

To develop the model we used in our experiment, some terminology plays a vital role such as: Optimizer, Stochastic Gradient Descent (SGD), Loss Function, Cross Entropy Loss function.

1) OPTIMIZER

The weights and other parameters of our model must be changed during training in order to minimize the loss function. As a result, we want to make sure that our classifications are as accurate as possible. To get the most out of the system. We update the model with the help of an optimizer that links the loss function and hyperparameters [46]. Optimizers come in a variety of shapes and sizes, each with its own set of advantages. We have chosen Stochastic Gradient Descent in our suggested model as our optimizer instead of Adagrad, RMSprop, Adam, etc.

2) STOCHASTIC GRADIENT DESCENT (SGD)

Using SGD, an objective function may be optimized over time. This minimizes the computing cost in high-dimensional optimization problems, resulting in quicker iterations in exchange for a slower convergence rate [47]. After each cycle, it takes a step in the direction of the steepest downward slope. A global optimum, or at least a close approximation to it, is what it aims for.

For the linear regression problem, a sample is denoted as $\mathrm{y_i} = (\mathrm{x1; x2; \ldots; xn})$, assuming that the function and loss function are defined as follows:

$$h_\theta(x_1, x_2, \ldots, x_n) = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$$
$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=0}^{m} (h_\theta(x_0, x_1, \ldots, x_n) - y_i)^2 \tag{9}$$

where $(\theta)_\mathrm{i}$ is the parameter of the model and $\mathrm{x_i}$ is the n^{th} eigenvalue of each sample.

3) LOSS FUNCTION

The goal of supervised machine learning algorithms is to reduce the error for each training sample throughout the learning process. Gradient descent and other optimization techniques are used in this process. And the loss function is to blame for this issue. An optimizer's guide is the loss function, which tells it whether it's heading in the right or incorrect path [48]. For example, the squared error function and the cross entropy loss function are two examples of loss functions. In our suggested model, we've included a Cross Entropy loss function.

4) CROSS ENTROPY LOSS FUNCTION

It is the difference between two probability distributions for a given random variable or sequence of events that is measured using the Cross Entropy Loss function. Log loss is also closely linked to and often confused with logistic loss.

The feature discrimination network is often subjected to the cross-entropy loss function. Here are the equations you need to know:

$$L = \begin{cases} -y \ln p, \quad y = 1 \\ -(1-y)\ln(1-p), \quad y = 0 \\ -\alpha(1-p)^{\gamma} y \ln p \end{cases}$$

$$FL = \begin{cases} -y = 1 \\ -(1-\alpha)p^{\gamma}(1-y)\ln(1-p), y = 0 \end{cases}$$

$$L_{\text{total}} = \begin{cases} AL + BL + CL, & \text{use } L \\ AFL + BFL + CL, & \text{use } FL \end{cases} \tag{10}$$

Here, the cross-entropy loss function is given by L, y represents the actual category label and p refers to the anticipated category label probability. As shown in equation 10, FL is used to balance the influence of negative and positive classes on loss function values, whereas K is used to examine the impact of stiff and easy samples on loss function values.. For example, L_{total} is an indicator of the total amount of training model loss that can be attributed to the loss function weights of G, H (the two classification models) and C (the feature categorization network).

F. PROPOSED DEEP NN MODEL 5

Music may be broken down into a number of distinct parts. As a rule of thumb, melody and harmony are the two most important aspects of a song. Both vertical and horizontal dimensions include harmony, although melody dominates the horizontal one. In this way, it is possible to say that each musical style varies from one another in terms of time intervals. As a result, by examining the temporal connection represented by MFCC characteristics, the Deep Neural Network (DNN) may play an important role in identifying musical genres with the extension of number of layer.

To classify the genre of Bangla music, we proposed a modified deep neural network model with layer 5 which we called BMNet-5 in terms of 512, 256, 128, 64 & 7 nodes with 29 features in the input layer which is depicted in Fig. 13. As there are six classes to forecast, the output layer comprises six nodes with softmax activation. Other layers were activated using RELU (Rectified Linear Unit). As a first-order gradient-based optimization technique, Adam [49] has been employed in our model since it is already implemented in Keras, which saves memory and processing power. "Sparse Categorical Crossentropy," since we encode our labels and our intended output is an integer, was employed for loss computation. A class is represented by an integer in the model's output. It has a numeric range of 0 to 5. At first we used 250-300 epochs, but after some basic adjustment we found that utilizing only 187 epochs and a batch size of 50 training data for each epoch yielded superior results.

There are tensor nodes in the input layer (IL) that represent the 2048-bit tensor length, using the formula IL = IW1,IW2,IW3,…, IWa. The feature vector from the Fully linked layer is imported into this layer. The embedding dimension of the input layer tensor is Tensor (2048, ED). There are no words in this part; instead, we use the number of words and the length of the features to create a tensor called Tensor(height, width, filter). The number of filters is used to replace the original filter.

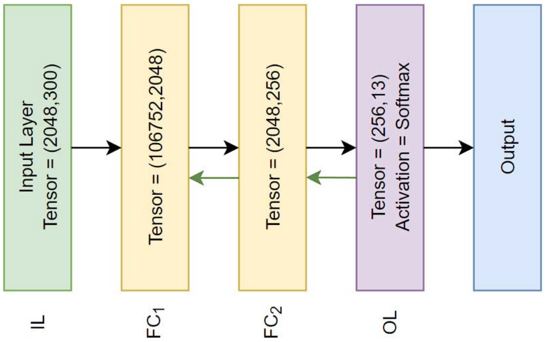

FIGURE 13. Proposed BMNet-5 model architecture. Left green arrow indicates the backpropagation and right black arrow denotes the forward propagation.

The suggested design uses two fully linked levels (FC1 and FC2). In the internal tensor nodes, the subscripts (n m) specify node (FC1 or FC2) lengths, which are represented by Z1,Z2,Z3,…, Z(nm). High-level characteristics are extracted by the convolution layer and flattened by the fully connected layer. There are 106,752 neurons in the first layer, and 2,048 neurons in the next layer are linked to them. In the Eqn 11, the fully connected layers use a linear model to transmit feature values.

$$W^i = W^{i-1} \times \omega^i \tag{11}$$

where, Flattened feature values are denoted by W^i (the i^{th}layer weight) and ω^i (the flattened value). Controlling model overfitting is accomplished by the use of a dropout technique. High-level characteristics may be generalized by the dropout technique.

Here, the feature map (Tensor(256, 1)) is fed into an input tensor, which then generates the desired genre name. OL's internal nodes correspond to the flattened feature map's length in the subscript (x), which is represented as O1, O2, O3,…, Ox. Eqn 12 is used to compute the predicted score:

$$E_i = \frac{\sum_{j=0}^{g} W(j,i) \times X(j)^i}{\sum_{k=1}^{CL} \sum_{j=0}^{g} W(k,j) \times X(j)^i} \tag{12}$$

Here, X(j) symbolizes the feature vector for the j^{th} class, while E_i represents the i^{th} class expectation g is the length of the feature, which is 256 characters. According to the values of g the classifier's accuracy may be improved. To improve classifier performance, we set the value of g to 256 in this implementation. The Soft-Max method is used to normalize the class value.

1) ACTIVATION FUNCTIONS

There are a variety of activation functions that we have tested in various layers and conditions. Without activation, a neural network acts like a linear regression. Activation

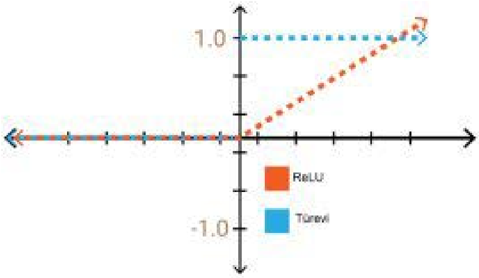

FIGURE 14. **ReLU activation function of BMNet-5 model.**

functions may be used in multilayered deep neural networks to learn meaningful features from input. As a result of their derivative function, which is tied to the inputs, they support backpropagation.

We have employed a variety of activation functions in various settings and layers, including ReLU, Leaky ReLU, Hyperbolic Tangent, Sigmoid, Step, and Linear Function, among others. It is possible to use ReLU as an all-purpose activation function. Only the hidden layers will make use of the ReLU function. We used a leaking ReLU function if we come across any dead neurons in our networks.

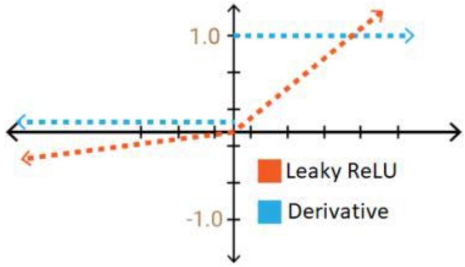

FIGURE 15. **Leaky ReLU activation function of BMNet-5 model.**

- A non-linear activation function, ReLU, is efficient and non-linear. It speeds up the convergence of the network. ReLU(x) = $(x)^+$ = the maximum possible value (0, x). Fig. 14 depicts the result.
- Leaky ReLU activation function is a variant of LeakyReLU(x), which has the equation max(0, x)* negativeslope min (0, x). Fig. 15 depicts the result.

G. INTERPRET THE MODEL PREDICTIONS

Machine and deep learning algorithms are very beneficial in our daily lives, yet they remain mostly unexplored.

Building confidence in the machine and deep learning model requires that predictions be explained or interpreted. SHapley Additive Explanations (SHAP) is one of the most extensively used techniques for understanding the predictions of any machine or deep learning model. Fig. 16 depicts the process of an explainable artificial intelligence in action. As a result of the method, a list of explanations may be found at the bottom of the output document.

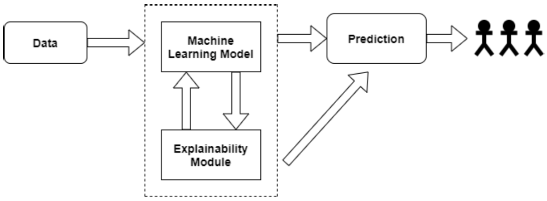

FIGURE 16. **Working procedure of explainable AI [50].**

1) SHAP

It is SHAP's job to calculate each feature's Shapley values, which represent the weight each feature has on the prediction as a whole. Aside from adding shapley explanations to our research, we have employed both locally interpretable and model-independent shapley explanations in our study. The output value (prediction) is the sum of these forces and the base value (average prediction across the validation set) [51]. Color-coded violin plots from all predictions allow us to summarize the relevance of a feature and (ii) link low/high feature values to an increase/decrease in output values.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we have presented the results of our experiments and the results of our performance evaluations. To begin, we constructed a heat map to demonstrate the relationship between the features and the model we were proposing. Next, we looked into how well the classifiers predicted Bangla music outcomes and genre using evaluation metrics. Additionally, we evaluate the suggested Model BMNet-5's performance. A confusion matrix, ROC curve, PR curve, and additional K-fold cross validations are used to determine the model's effectiveness. Finally, the results of our research are compared to those of previous studies.

A. HEAT MAP OF FEATURES REPRESENTATION

We think that the inability of certain algorithms to detect the most significant and strongly associated attributes is one explanation for their poor performance. First, we would want to come up with an approach that can determine the best set of characteristics, and then the best algorithms to use with them. Fig. 17 represents the correlated features of the work. As far as we could tell, the algorithms who performed well had access to the derived feature-set that was well-correlated, whereas those who didn't performed as well were unable to correctly analyze the correlation structure of their feature-sets.

Based on the important feature selection technique's selection of 29 strongly associated features with the expected attribute (num), the following figure was created. The attribute values are shown on the right side (ranging from 0.2 to -0.8). From the figure, we analyzed a substantial association between spectral_centroid and the spectral_bandwidth, mfcc0, mfcc2 characteristics, with a correlation value of around 0.2. The lowest correlation was found for the mfcc1 feature, with a correlation value of about -0.8. Similarly, there

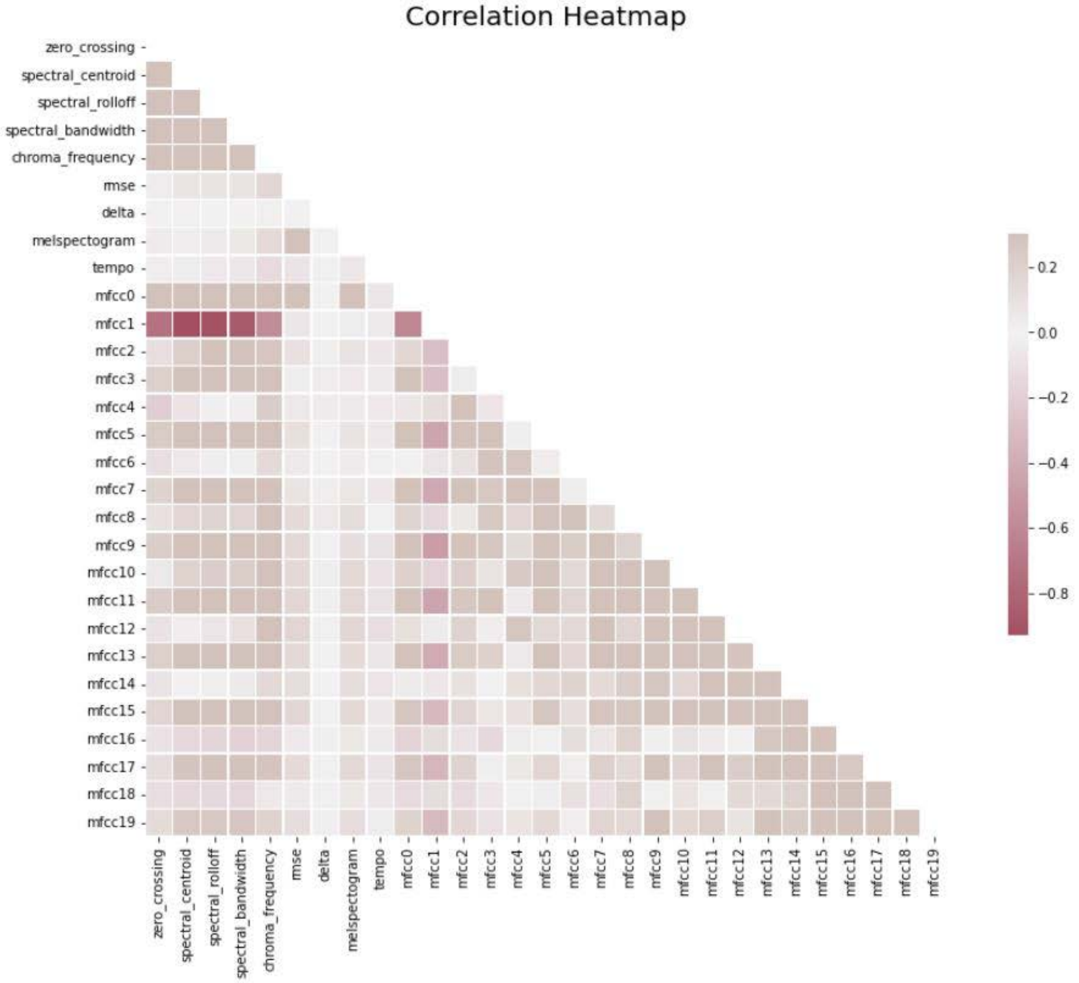

FIGURE 17. Correlated features of the work.

is a strong link between mfcc9 and chroma_frequency. But other characteristics' linked values ranged from 0.0 to −0.6, which isn't very high.

B. ENVIRONMENTAL SETUP OF THE STUDY

The whole procedure was carried out using a multi-core computer that had an Intel Core i7 CPU running at 3.4 GHz, an NVIDIA Geforce GTX 980M graphics processing unit (GPU) with 8 GB of memory, and other components. Python version 3.7 was used in conjunction with the deep learning framework Keras, which used Tensorflow as its backend.

C. EVALUATION MEASURES

In order to evaluate the quality of our custom models, we will use the evaluation criteria provided by the scikit-learn package [52]. Our first objective is to choose the model that works best for our particular circumstances. In order to do our calculations, we will make use of the score meter, the score calculating f1, and the uncertainty matrix. Metrics like as accuracy, precision, recall, and f1-score need to be measured in order for us to be able to compare different algorithms.

Accuracy, precision, recall, and f-score are all represented in the form of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in equations 13 through 16. These equations may be found below.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

$$F1 * \text{ score} = \frac{2 \text{ precision recall}}{\text{precision } + \text{ recall}} \tag{16}$$

An explanation of each of the aforementioned measures is provided in the following paragraphs:

- True Positive (TP) is the total number of occurrences that have been properly labeled as having a positive value or yes (1) by the created model M* after the labeled instances have been updated.
- True Negative (TN) refers to the total number of instances that were properly identified by the created model as having a negative value or the value zero.

TABLE 2. Hyperparameters of the classifiers along with proposed BMNet-5 model.

Model	No. of Hidden Layer	Optimizer	Activation Function	Batch Size	No. of Iteration	No. of Epochs	Learning Rate	No. of Nodes per Hidden Layer
Logistic Regression	0	SGD	Sigmoid	32	1500	39	0.1	0
Shallow Neural Network	1	SGD	ReLU	32	1500	62	0.1	50
Convolutional Neural Network	2	Adam	ReLU	100	610	50	0.001	50
Deep NN Model 1	1	SGD	ReLU	32	1200	39	0.1	32
Deep NN Model 2	2	SGD	ReLU	32	1500	39	0.1	150
Deep NN Model 3	3	SGD	ReLU	32	1500	39	0.1	100
Deep NN Model 4	4	SGD	LeakyReLU	32	1500	39	0.1	50, 30
Proposed BMNet-5 Model	**5**	**Adam**	**LeakyReLU**	**50**	**4560**	**187**	**0.0005**	**100**

- False Positive (FP) is the total amount of occurrences classified incorrectly, which means that the machine predicts the value as positive/yes (1) but its actual value is negative/no (0) by the generated model M* after updating the labeled instances. In other words, FP is the total amount of occurrences that were incorrectly classified.
- A computer forecasts the value as negative/no (0), but its real value is positive/yes (1) as determined by the created model. This is what is meant by the term "false negative," which refers to the total quantity of occurrences categorized wrongly.
- Accuracy: The accuracy of the test is determined by the percentage of the total data that is correctly classified.
- Precision is a statistic that is used to evaluate how precise a class is in comparison to the actual world.
- Recall: A recall is a kind of measure that is used to evaluate how well prepared a class is.
- F-Score: The F-score was designed in order to take into consideration the possibility of both false positive and false negative results.

D. HYPERPARAMETER OPTIMIZATION

It has been noted that better models may be discovered in a shorter amount of time by randomly choosing values for hyperparameters as opposed to doing an exhaustive

grid search [53]. On the basis of the dataset that we created, we tweaked the hyperparameters of the model that was used in addition to proposing BMNet-5. The values of the hyper parameters were picked using a method called random searching, which included going through a list of all of the available values and picking the ones that had the best overall performance. Table 2 lists the hyper parameters of the models that were used as well as the BMNet-5 model that was recommended. From the Table, we analyzed that in terms of Logistic Regression, there is 0 hidden layer and also the no. of nodes per hidden layer is 0 whereas Shallow Neural Network has 1 no. of hidden layer and 50 no. of nodes per hidden layer. Again, in case of convolutional neural, we used Adam as an optimizer and ReLU for activation function and the batch size is 100. Both Deep NN Model 1 and 2, we used SGD optimizer and ReLU activation function whereas the no. iteration is different with the value of 1200 and 1500. On the other hand, there is a correlation between the changes in the hyperparameters and the changes in the performance of Deep NN Model 3 and 4. The amount of hidden layers in Model 4 is different from the previous models. It employs 30 hidden nodes for the fourth layer, whereas Model 3 gives 100 hidden nodes for each of the first three layers it travels through. For the first three layers, it uses 50 hidden nodes. Adjusting the values of the hyperparameters may have a significant impact on the findings. There is no other way to fine-tune the settings outside experimenting with the many options available. Nevertheless, improving the activation function and optimizer function via analysis may lead to an improvement in overall performance. Our proposed BMNet-5 model produced 5 no. of hidden layer, Adam as an optimizer, 50 as a batch size with 187 epochs. The activation functions that were employed were called Leaky ReLU and ReLU with 100 hidden layers. The activation functions' ability to alter expected outcomes is shown by this model.

When we compare it to the logistic regression, we can state that practically all of the parameters are the same. However, since the model has hidden layers, it is able to learn the data more and make more accurate predictions. It is reasonable to infer that the deep neural network performs well for the categorization of audio genres.

E. RESULT & DISCUSSION

In order to get the findings, we take into account the computation of accuracy, precision, and recall, as well as the measurement of f1-score. It is common knowledge that these four assessment criteria are used to evaluate the overall quality of the models along with proposed BMNet-5.

We have compared with different models i.e. LR, SNN, CNNs of different layers because we tried to experiment a number of layered models before settling on the one that provides the most accurate results. It helped to show that our proposed model is better and consistent than the traditional models. Table 3 shows a performance comparison of the classifiers we utilized along with proposed BMNet-5. In summary, our proposed model outperformed the other applied model with a high accuracy of 90.32%. The closest applied model is Deep NN Model 4 with an accuracy of 79.15%. The lowest accuracy comes from the logistic regression with an accuracy of 70.25%. Our proposed model also perfromed better than others in terms of precision, recall and f1-score with a value of 85.13%, 81.74% and 85.53% respectively.

TABLE 3. Experimental results of bangla music genre classification.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression (LR)	70.25%	68.28%	61.34%	64.44%
ShallowNueral Network	71.72%	68.12%	67.34%	69.44%
Convolutional Neural Network (CNN)	78.3%	70.24%	73.95%	74.43%
Deep NN Model 1	76.85%	76.86%	73.85%	75.50%
Deep NN Model 2	77.85%	76.11%	77.25%	74.30%
Deep NN Model 3	78.85%	76.26%	69.15%	75.50%
Deep NN Model 4	79.15%	76.16%	75.11%	72.52%
Proposed BMNet-5 Model	**90.32%**	**85.13%**	**81.74%**	**85.53%**

TABLE 4. Performance comparison with traditional ml classifiers.

Model	Accuracy	Precision	Recall	F1 Score
K-Nearest Neighbor (KNN)	73.38%	69.44%	64.65%	67.76%
Support Vector Machine (SVM)	78.23%	70.34%	68.47%	69.56%
Decision Tree (DT)	77.32%	74.31%	75.83%	76.83%
Random Forest (RF)	79.34%	77.21%	76.84%	73.55%
Proposed BMNet-5 Model	**90.32%**	**85.13%**	**81.74%**	**85.53%**

We also compared our results with some traditional ml classifiers such as: KNN, SVM, DT and RF. Table 4 represents the comparison results with our proposed BMNET-5. KNN produced accuracy of 73.38%, SVM got 78.23% while our proposed model came out the best results with an accuracy of 90.32%. In terms of ml classifiers, Random Forest (RF) came out the best results with an accuracy of 79.34% compare to KNN, SVM and DT.

Again, we also considered the per class of the dataset to get the evaluation score with the help of our proposed BMNet-5 model. Table 5 depicts the evaluation score of per class in Bangla Music Genre on the proposed BMNet-5 model.

TABLE 5. Evaluation score of per class in Bangla music genre on the proposed BMNet-5 model.

Class	Accuracy	Precision	Recall	F1 Score
Adhunik	80.45%	78.33%	66.38%	63.29%
Band Songs	75.32%	68.12%	67.34%	71.68%
Hip-Hop	83.34%	70.24%	73.95%	78.41%
Nazrulgeeti	79.54%	80.32%	72.34%	73.44%
Palligeeti	90.93%	86.22%	83.24%	76.23%
Rabindra Sangeet	83.32%	81.12%	78.64%	84.25%

We have shown from the table, that Palligeeti has highest accuracy with 90.93% whereas Band Songs got lowest of 75.32%. Rabindra Sangeet has the highest f1-score with a value of 84.25% while Adhunik has lowest with the value of 63.29% f1-score. In terms of precision and recall, Palligeeti performed better with the value of 86.22% and 83.24% while Rabindra Sangeet has the closest one with the result of 81.12% and 78.64%.

F. OPTIMAL PROPOSED MODEL EVALUATION

In terms of the Bangla Music dataset, our suggested BMNet-5 model outperformed all others as shown in Table 3. In order to do an in-depth examination of the model that performed the best, BMNet-5, a confusion matrix, ROC Curve, and PR curve are developed. The model's resilience is further tested by tests using a broad range of K-folds configurations. Finally, an interpretation model SHAP is utilized to calculate the likelihood of properly categorizing the instance and to provide an explanation for the enhanced classification transparency. Fig. 18 depicts the flow of the optimal evaluation of our proposed BMNet-5 model.

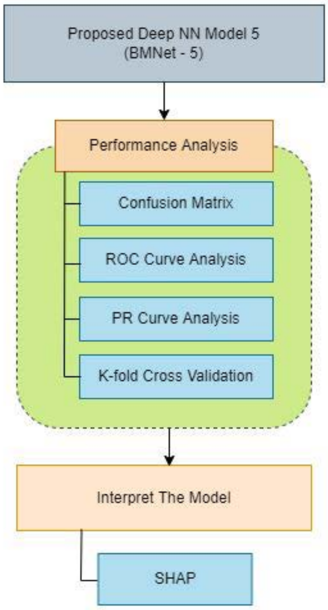

FIGURE 18. Flowchart of our proposed BMNet-5 model optimal evaluation.

1) CONFUSION MATRIX

The performance of a classification algorithm may be summarized using a confusion matrix. If the dataset comprises more than two classes or an unequal number of observations in each class, classification accuracy alone may be misleading [54]. So, in order to make the results easier to understand, the method mentioned above was used. In a confusion matrix, you may see how many positive and negative outcomes a model predicted accurately.

Fig. 19 illustrates the confusion matrix of the proposed BMNet-5 model. Predicted labels are shown on the x axis, while the actual labels are shown on the y axis. In case of class 0, we analyzed that 22 data are missclassified and 137 data are classified properly which denotes first row. Again, in terms of class 5, 151 data are classified properly while 22 data are missclassified which denotes the fifth row.

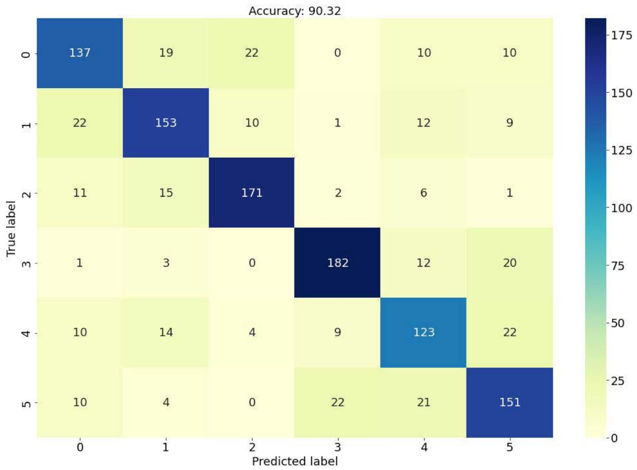

FIGURE 19. Confusion matrix of the proposed BMNet-5 model.

2) ROC CURVE ANALYSIS

The ROC probability curve is displayed for each of the six different classes of Bangle music using the proposed model BMNet-5 as well as the other classifiers that are shown in Fig. 20. The performance of the proposed model in distinguishing classes is shown by the ROC score of six different classes. Fig. 20 allows us to draw the conclusion that the ROC curve of Hip-Hop and Band Songs nearly reaches the top of the y-axis, and it has a false positive rate that is very close to zero and a true positive rate that is very close to one. The overall ROC value of the Hip-Hop and Band Songs genre is 97%.

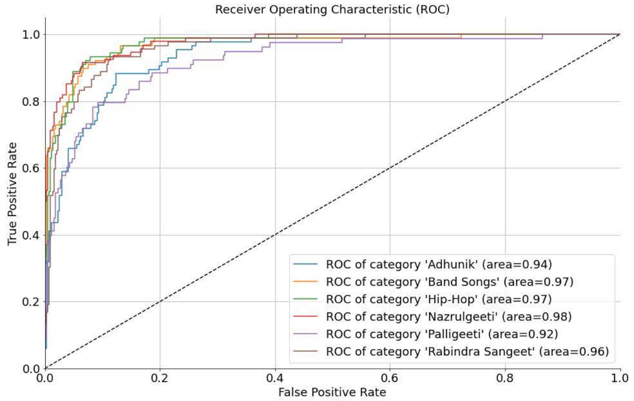

FIGURE 20. ROC curve analysis of proposed BMNet-5 model.

3) PR CURVE ANALYSIS

The second curve that we employ for the purpose of evaluating the results of the test is referred to as the Precision-Recall (PR) curve. The link between a recommender's memory and the accuracy of their forecasts is shown by this curve. The majority of the time, improving the degree of accuracy will result in a drop in the amount of recall that is achieved. This is due to the fact that a result that is more accurate and has fewer false positives (more precision) will often have a lower recall and lead to more false negatives than a result that is less accurate but has more false positives. This is true both in the forward and backward directions at the same time.

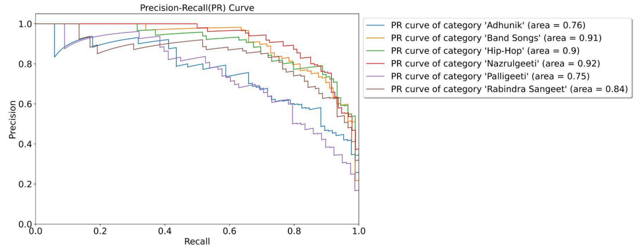

FIGURE 21. PR curve analysis of proposed BMNet-5 model.

Fig. 21 presents the PR curves in terms of our suggested BMNet-5 for each of the six distinct kinds of Bangla music genre. These curves may be found in terms of the popularity ratio. According to the information shown in this graph, the performance of the *Nazrulgeeti* class is noticeably superior to that of the other classes, with a value of 92%.

4) DIFFERENT K-FOLD CROSS VALIDATION

In order to give more evidence for the consistency of the proposed BMNet-5 model's performance relative to that of other classifiers, we conducted a series of experiments using a total of ten K-fold cross validation setups, each with a different K value ranging from 1 to 10.

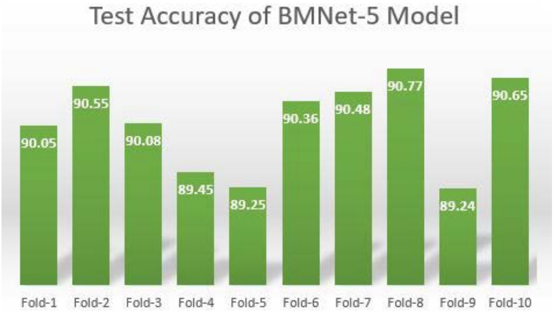

FIGURE 22. Performance evaluation using K-fold cross validations with various K values ranging from 1 to 10.

The findings of each K-fold cross validation are shown in Fig. 22. Clearly, the proposed BMNet-5 is capable of providing outstanding results (performance more than 89 percent) for all K-folds. The fact that performance did not drop considerably for any fold helps to the robustness and consistency of the proposed BMNet-5 model.

5) SHAP INTERPRETATION OF THE PROPOSED MODEL

SHAP values illustrate the impact of a single feature's value relative to the prediction we would make if this feature had a baseline value. The acronym SHAP refers to Shapley Additional Explanations. In addition to its supporting package, we used the SHAP library and applied the SHAP tree explanation object. The baseline for the Shapley values is the average of all predictions produced from the training set. In order to explain a specific prediction, the Shapley value of each feature is either added to or subtracted from this baseline.

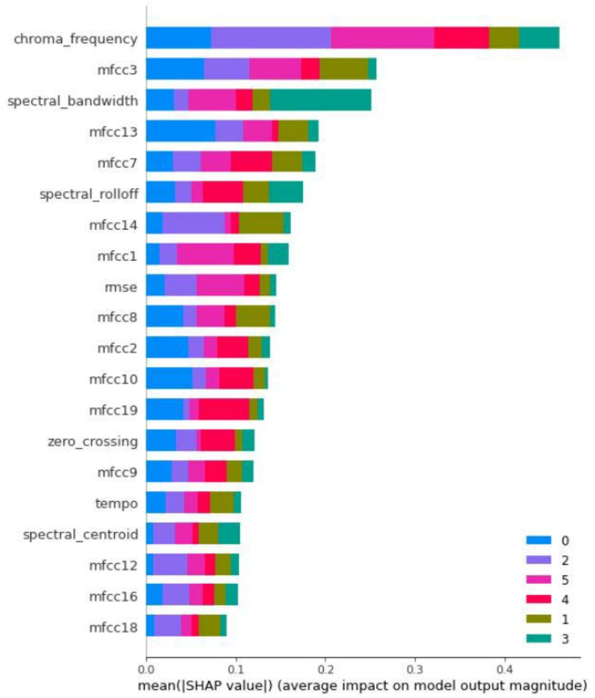

FIGURE 23. SHAP importance feature for Bangla music genre based on proposed BMNet-5 model.

Fig. 23 depicts the important consequences of the SHAP's primary characteristics, beginning with the most significant one. Here, the characteristics are ranked according to their impact on the model's prediction. The y-axis represents the features, while the x-axis displays the average absolute SHAP value of each feature.

Here, class 0 means Adhunik, class 1 means Band Songs, class 2 means Hip-Hop, class 3 means Nazrulgeeti, class 4 means Palligeeti and class 5 means Rabindra Sangeet where chroma_frequency feature extract all the classes simultaneously. On other hand, mfcc18 feature has the lowest range of classes. To make the model efficient we worked on the most relevant characteristics of the feature such for chroma_frequency, we analyzed from Fig. 23 that class 0 (blue) means adhunik has medium impact, class 2 (violet) denotes most impact, class 5 (pink) has less than most impact, class 4 (red) denotes less than medium impact, class 1 (paste) has less impact and lastly, class 3 (green) impact better than class 1 (paste). That's how we worked on the most relevant characteristics of specific feature to make the model more efficient.

The following Fig. 24 illustrates the most positive impact features which colored in red. Here, the only bad impact features are mfcc2 and chroma_frequency. While all the other features are impacted positive in terms of proposed BMNet-5 model for the Bangla music genre dataset. Fig. 25 depicts the most lowest effect aspects, which are indicated in blue. Only chroma frequency have a positive effect on this image which is denoted by red color. While all other characteristics are negatively influenced by the proposed BMNet-5 model for the Bangla music genre dataset in this scenerio.

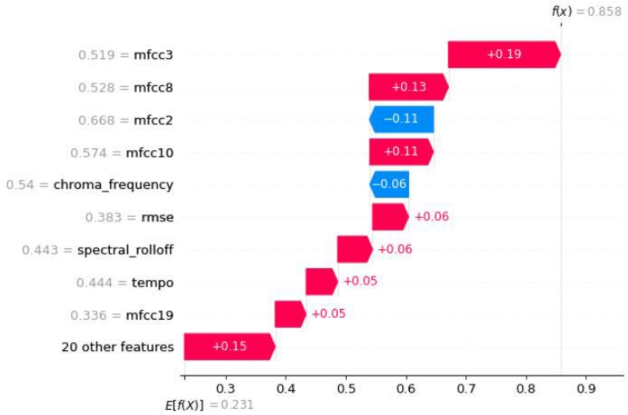

FIGURE 24. SHAP prediction for proposed BMNet-5 model in terms of positive impact.

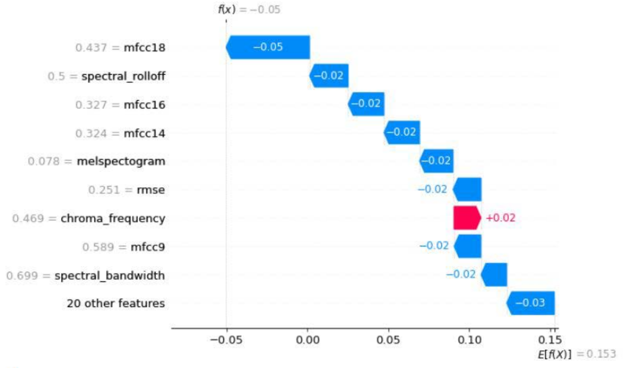

FIGURE 25. SHAP prediction for proposed BMNet-5 model in terms of negative impact.

G. COMPARISON OF ACCURACY WITH SOME EXISTING LITERATURES

In terms of its accuracy, we compared the performance of our model to that of previous works that are considered to be state-of-the-art, as shown in Table 5. In order to carry out the comparison, we decided to use the BMNet-5 suggested model since the performance of our model was superior to that of the other classifiers that were used.

From the Table, we analyzed that our proposed model BMNET-5 outperformed all the existing work regarding Bangla music genre classification. Most results in the Table 6 are in the range of 74%-85% which is lower than ours. Bhowmik et al. [55] used total no. of 1200 Bengali songs with 4 genres achieving an accuracy of 85% with SVM

TABLE 6. Comparison of accuracy with some existing literatures.

Paper	Dataset for classification	Genres	Best Method	Accuracy
Bhowmik et al., 2019 [55]	Total 1200 Bengali songs	4	SVM	85%
Mamun et al., 2019 [19]	Total 1742 Bengali songs	6	Feed-forward Neural Network	74%
Sarma et al., 2021 [56]	Total 2944 Bengali songs	6	GRU based model	80.4%
Proposed Approach	**Total 1742 Bengali songs**	**6**	**BMNet-5**	**90.32%**

classifier. Mamun et al. [19] proposed feed-forward neural network to classify 1742 bengali songs with 6 classifiers but ended with an accuracy of 74%. Finally, Sarma et al. [56] suggested GRU based model to classify Bengali music with an accuracy of 80.4%. But our proposed model BMNet-5 classified 1742 Bengali songs with 6 genres with an accuracy of 90.32% which denotes that our model is best among other existing works.

VI. CONCLUSION AND FUTURE WORK

Since musical instruments and technology advance quickly, the categorization of musical genres is emerging as an intriguing study area. Because there have been so few studies in this area, scholars are now paying more attention to Bengali music. Huge numbers of Bangla songs are generated daily as a result of the music industry's fast expansion in Bangladesh. The creation of songs from various genres involves a huge number of producers, lyricists, vocalists, and artists. Classification methods for music genres are essential for managing and using music databases. This research work presented a novel approach known as BMNet-5 to classify Bangla music genres as "Bangla Adhunik," "Bangla Hip-Hop," "Bangla Band Music," "Nazrulgeeti," "Palligeeti," and "Rabindra Sangeet". By extracting characteristics from a dataset of 1742 Bangla music compositions and testing their automatic categorization, we demonstrate the efficiency of the method. Based on an audio input, the proposed BMNet-5 uses a neural network to predict the genre of music. In comparison to the prior studies, our model was 90.32 percent accurate. Following this, the BMNet-5 model's performance consistency is verified using K-fold cross validation with different k values. To sum it up, we apply the recommended model to train the interpretable SHAP model for a specific category of our dataset, and the creation of an explainable result may be a substantial benefit.

In our proposed model, we couldn't observe the distinct feature rankings based on how well they operate with various deformations which, without the inclusion of large amounts of white noise, has poor generalization skills against all deformation kinds. We can handle this situation using modified harmonic CNN to specify the issues. If additional audio features can be added to our training set, the model may perform better. Additionally, if there were more audio files available, the model may perform better. By expanding the scope of our feature gathering domain, we can enhance feature extraction methods in future. Additionally, we want to use the vast Bengali music genre dataset to execute transfer learning with fine-tuning.

REFERENCES

[1] D. J. Levitin, *This is Your Brain on Music: The Science of a Human Obsession*. New York, NY, USA: Penguin, 2006.

[2] D. Huron, "Is music an evolutionary adaptation?" *Ann. New York Acad. Sci.*, vol. 930, no. 1, pp. 43–61, Jun. 2001.

[3] The Encyclopedia of Banglapedia. (2015). *Music*. [Online]. Available: http://en.banglapedia.org/index.php?title=Music

[4] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[5] A. Schindler, T. Lidy, and A. Rauber, "Comparing shallow versus deep neural network architectures for automatic music genre classification," in *Proc. FMT*, 2016, pp. 17–21.

[6] A. R. Rajanna, K. Aryafar, A. Shokoufandeh, and R. Ptucha, "Deep neural networks: A case study for music genre classification," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 655–660.

[7] N. Pelchat and C. M. Gelowitz, "Neural network music genre classification," *Can. J. Electr. Comput. Eng.*, vol. 43, no. 3, pp. 170–173, 2020.

[8] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[9] M. F. Mushtaq, U. Akram, M. Aamir, H. Ali, and M. Zulqarnain, "Neural network techniques for time series prediction: A review," *Int. J. Informat. Visualizat.*, vol. 3, no. 3, pp. 314–320, Aug. 2019.

[10] C.-Y. Chang, C.-K. Wu, C.-Y. Lo, C.-J. Wang, and P.-C. Chung, "Music emotion recognition with consideration of personal preference," in *Proc. Int. Workshop Multidimensional (nD) Syst.*, Sep. 2011, pp. 1–4.

[11] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," *Multimedia Tools Appl.*, vol. 80, no. 5, pp. 7313–7331, Feb. 2021.

[12] S. Vishnupriya and K. Meenakshi, "Automatic music genre classification using convolution neural network," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2018, pp. 1–4.

[13] H. Bahuleyan, "Music genre classification using machine learning techniques," 2018, *arXiv:1804.01149*.

[14] G. Jawaherlalnehru, S. Jothilakshmi, T. Nadu, and T. Nadu, "Music genre classification using deep neural networks," *Int. J. Sci. Res. Sci., Eng. Technol.*, vol. 4, no. 4, p. 935, 2018.

[15] M. Asim and Z. Ahmed, "Automatic music genres classification using machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 8, pp. 1–8, 2017.

[16] J. R. Castillo and M. J. Flores, "Web-based music genre classification for timeline song visualization and analysis," *IEEE Access*, vol. 9, pp. 18801–18816, 2021.

[17] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 282–289.

[18] K. M. Hasib, M. S. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. I. H. Showrov, S. Ahmed, and O. Rahman, "A survey of methods for managing the classification and solution of data imbalance problem," 2020, *arXiv:2012.11870*.

[19] M. A. A. Mamun, I. Kadir, A. S. A. Rabby, and A. A. Azmi, "Bangla music genre classification using neural network," in *Proc. 8th Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Nov. 2019, pp. 397–403.

[20] D. Chaudhary, N. P. Singh, and S. Singh, "Genre based classification of Hindi music," in *Proc. Int. Conf. Innov. Bio-Inspired Comput. Appl.* Cham, Switzerland: Springer, 2018, pp. 73–82.

[21] P.-K. Jao and Y.-H. Yang, "Music annotation and retrieval using unlabeled exemplars: Correlation and sparse codes," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1771–1775, Oct. 2015.

[22] B. G. Patra, D. Das, and S. Bandyopadhyay, "Multimodal mood classification of Hindi and western songs," *J. Intell. Inf. Syst.*, vol. 51, no. 3, pp. 579–596, Dec. 2018.

[23] S. Jothilakshmi and N. Kathiresan, ''Automatic music genre classification for Indian music,'' in *Proc. Int. Conf. Softw. Comput. App.*, 2012, pp. 1–5.

[24] B. Rajesh and D. G. Bhalke, ''Automatic genre classification of Indian Tamil and western music using fractional MFCC,'' *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 551–563, Sep. 2016.

[25] F. Gouyon, F. Pachet, and O. Delerue, ''On the use of zero-crossing rate for an application of classification of percussive sounds,'' in *Proc. COST G-6 Conf. Digit. Audio Effects*, Verona, Italy, vol. 5, 2000, p. 16.

[26] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, ''Librosa: Audio and music signal analysis in Python,'' in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[27] S. Choi, S. Jeong, J. Yoon, M. Yang, M. Ko, E. Park, J. Han, M. Lee, and S. Lee, ''VCTUBE: A library for automatic speech data annotation,'' in *Proc. Interspeech*, 2020, pp. 1013–1014.

[28] H.-T. Kim, ''Real-time flame detection using colour and dynamic features of flame based on FFmpeg,'' *J. Korea Inst. Electron. Commun. Sci.*, vol. 9, no. 9, pp. 977–982, Sep. 2014.

[29] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, ''Evaluation of CNN-based automatic music tagging models,'' 2020, *arXiv:2006.00751*.

[30] D. G. Rainham, C. J. Bates, C. M. Blanchard, T. J. Dummer, S. F. Kirk, and C. L. Shearer, ''Spatial classification of youth physical activity patterns,'' *Amer. J. Preventive Med.*, vol. 42, no. 5, pp. e87–e96, May 2012.

[31] T. Pohle, E. Pampalk, and G. Widmer, ''Evaluation of frequently used audio features for classification of music into perceptual categories,'' in *Proc. 4th Int. Workshop Content-Based Multimedia Indexing*, vol. 162, 2005, pp. 1–8.

[32] E. Priya, P. S. Reshma, and S. Sashaank, ''Temporal and spectral features based gender recognition from audio signals,'' in *Proc. Int. Conf. Commun., Comput. Internet Things (IC3IoT)*, Mar. 2022, pp. 1–5.

[33] U. Ayvaz, H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, and A. A. Bobomirzaevich, ''Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning,'' *Comput., Mater. Continua*, vol. 71, no. 3, pp. 5511–5521, 2022.

[34] J. V. T. Abraham, A. N. Khan, and A. Shahina, ''A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients,'' *Int. J. Speech Technol.*, vol. 2021, pp. 1–9, Aug. 2021.

[35] J. M. Ha, H. B. Shin, J. F. Joung, W. J. Chung, J.-E. Jeong, S. Kim, S. H. Hur, S.-Y. Bae, J.-Y. Kim, J. Y. Lee, S. Park, and H. Y. Woo, ''Rational molecular design of azaacene-based narrowband green-emitting fluorophores: Modulation of spectral bandwidth and vibronic transitions,'' *ACS Appl. Mater. Interfaces*, vol. 13, no. 22, pp. 26227–26236, Jun. 2021.

[36] C. Cox, W. Trojak, T. Dzanic, F. D. Witherden, and A. Jameson, ''Accuracy, stability, and performance comparison between the spectral difference and flux reconstruction schemes,'' *Comput. Fluids*, vol. 221, May 2021, Art. no. 104922.

[37] M. F. Asmussen, J. Liniger, and H. C. Pedersen, ''Fault detection and diagnosis methods for fluid power pitch system components—A review,'' *Energies*, vol. 14, no. 5, p. 1305, Feb. 2021.

[38] M. Ortiz, P. Vicente, E. Ianez, E. Montiel, and J. M. Azorin, ''Assessing footwear comfort by electroencephalography analysis,'' *IEEE Access*, vol. 9, pp. 134259–134269, 2021.

[39] S. Sazzed and S. Jayarathna, ''A sentiment classification in Bengali and machine translated English corpus,'' in *Proc. IEEE 20th Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Jul. 2019, pp. 107–114.

[40] J. C. Van Houwelingen, ''Ridge estimators in logistic regression,'' *J. Roy. Stat. Soc., C, Appl. Statist.*, vol. 43, no. 1, pp. 95–108, 1992.

[41] D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, ''Fault diagnosis of rolling element bearing by using multinomial logistic regression and wavelet packet transform,'' *Soft Comput.*, vol. 18, no. 2, pp. 255–266, Feb. 2014.

[42] M. Fan, ''Application of music industry based on the deep neural network,'' *Sci. Program.*, vol. 2022, pp. 1–6, Jan. 2022.

[43] J. Dias, V. Pillai, H. Deshmukh, and A. Shah, ''Music genre classification & recommendation system using CNN,'' *SSRN Electron. J.*, vol. 2022, pp. 1–7, Aug. 2022.

[44] M. Shah, N. Pujara, K. Mangaroliya, L. Gohil, T. Vyas, and S. Degadwala, ''Music genre classification using deep learning,'' in *Proc. 6th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Mar. 2022, pp. 974–978.

[45] J. Pakela and I. E. Naqa, ''Overview of deep machine learning methods,'' in *Machine and Deep Learning in Oncology, Medical Physics and Radiology*. Cham, Switzerland: Springer, 2022, pp. 51–77.

[46] Y. Mao, G. Zhong, H. Wang, and K. Huang, ''Music-CRN: An efficient content-based music classification and recommendation network,'' *Cognit. Comput.*, vol. 2022, pp. 1–11, Jul. 2022.

[47] M. T. Quasim, E. H. Alkhammash, M. A. Khan, and M. Hadjouni, ''Emotion-based music recommendation and classification using machine learning with IoT framework,'' *Soft Comput.*, vol. 25, no. 18, pp. 12249–12260, Sep. 2021.

[48] A. K. Sharma, G. Aggarwal, S. Bhardwaj, P. Chakrabarti, T. Chakrabarti, J. H. Abawajy, S. Bhattacharyya, R. Mishra, A. Das, and H. Mahdin, ''Classification of Indian classical music with time-series matching deep learning approach,'' *IEEE Access*, vol. 9, pp. 102041–102052, 2021.

[49] A. V. Ogal'tsov and A. I. Tyurin, ''A heuristic adaptive fast gradient method in stochastic optimization problems,'' *Comput. Math. Math. Phys.*, vol. 60, no. 7, pp. 1108–1115, Jul. 2020.

[50] K. M. Hasib, F. Rahman, R. Hasnat, and M. G. R. Alam, ''A machine learning and explainable AI approach for predicting secondary school Student performance,'' in *Proc. IEEE 12th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2022, pp. 0399–0405.

[51] H. Nori, S. Jenkins, P. Koch, and R. Caruana, ''InterpretML: A unified framework for machine learning interpretability,'' 2019, *arXiv:1909.09223*.

[52] K. M. Hasib, N. A. Towhid, and M. R. Islam, ''HSDLM: A hybrid sampling with deep learning method for imbalanced data classification,'' *Int. J. Cloud Appl. Comput.*, vol. 11, no. 4, pp. 1–13, Oct. 2021.

[53] J. Bergstra and Y. Bengio, ''Random search for hyper-parameter optimization,'' *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 1–25, 2012.

[54] K. M. Hasib, M. A. Habib, N. A. Towhid, and M. I. H. Showrov, ''A novel deep learning based sentiment analysis of Twitter data for U.S. Airline service,'' in *Proc. Int. Conf. Inf. Commun. Technol. Sustain. Develop. (ICICT4SD)*, Feb. 2021, pp. 450–455.

[55] A. Bhowmik and A. E. Chowdhury, ''Genre of Bangla music: A machine classification learning approach,'' *AIUB J. Sci. Eng.*, vol. 18, no. 2, pp. 66–72, Aug. 2019.

[56] M. S. Sarma and A. Das, ''BMGC: A deep learning approach to classify Bengali music genres,'' in *Proc. 4th Int. Conf. Netw., Inf. Syst. Acad. Manage. Perspect. Security.*, Apr. 2021, pp. 1–6.

KHAN MD. HASIB (Member, IEEE) received the B.Sc. degree from the Computer Science and Engineering Department, Ahsanullah University of Science and Technology (AUST), and the M.Sc. degree from the Computer Science and Engineering Department, BRAC University. He has more than three years of research and teaching experiences in computer science. He has been heavily involved in collaborative research activities especially in the fields of machine learning, deep learning, data mining, health informatics, computer vision, and natural language processing. He has published over 20 research papers in highly recognized journals and conference proceedings. He is working on several projects, such as airline industry customer review prediction, early Parkinson's disease prediction, and COVID-19 detection from medical image data and diabetes prediction using deep learning and machine learning algorithms.

ANIKA TANZIM received the B.Sc. degree in computer science and engineering from the Ahsanullah University of Science and Technology (AUST). She is currently pursuing the M.Sc. degree in big data and data science technology with advanced practice with Northumbria University, London Campus. Her research interests include machine learning, deep learning, question answering systems, and natural language processing. Her undergraduate thesis was natural language question answering system with DBpedia.

JUNGPIL SHIN (Senior Member, IEEE) received the B.Sc. degree in computer science and statistics and the M.Sc. degree in computer science from Pusan National University, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science and communication engineering from Kyushu University, Japan, in 1999, under a scholarship from the Japanese Government (MEXT). He was an Associate Professor, a Senior Associate Professor, and a Full Professor with the School of Computer Science and Engineering, The University of Aizu, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 250 published papers for widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, machine learning, human–computer interaction, non-touch interfaces, human gesture recognition, automatic control, Parkinson's disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, as well as handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He served as a program chair and as a program committee member for numerous international conferences. He serves as an Editor for IEEE journals and for *Sensors* (MDPI). He serves as a reviewer for several major IEEE and SCI journals.

KAZI OMAR FARUK (Member, IEEE) received the B.Sc. degree in computer science and engineering from the Ahsanullah University of Science and Technology (AUST). He is currently pursuing the M.Sc. degree in computer science and engineering with BRAC University. He has more than two years of research experience in advanced artificial intelligence, machine learning, and deep learning. He is doing extensive research in the field of genetic optimization, federated learning and its application, autoencoders, collaborative filtering, multicriteria decision making, recommendation system, application of natural language processing and deep learning in recommendation systems, and published six research papers in highly reputed conferences within a very short time.

JUBAYER AL MAHMUD received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Dhaka. He has been active in teaching and research for more than four years in the field of computer science and engineering. He worked at several universities as a Lecturer and later an Assistant Professor of computer science and engineering. His research interests include multiple fields of computer science, specially machine learning, computer vision, data mining, natural language processing, human–robot interaction, and human–computer interaction. He has published five research papers in well reputed journals and conference proceedings and many more under review. He is currently working on several research projects, such as COVID-19 vaccine efficacy and aftereffects in Bangladesh, automated gesture adaptation for human–robot interaction using machine learning algorithms.

M. F. MRIDHA (Senior Member, IEEE) received the Ph.D. degree in AI/ML from Jahangirnagar University, in 2017. He is currently working as an Associate Professor with the Department of Computer Science, American International University-Bangladesh (AIUB). Before that, he worked as an Associate Professor and the Chairman of the Department of CSE, Bangladesh University of Business and Technology. He also worked as a CSE Department Faculty Member at the University of Asia Pacific and as a Graduate Head, from 2012 to 2019. His research experience, within both academia and industry, results in over 120 journals and conference publications. For more than ten years, he has been with the master's and undergraduate students as a supervisor of their thesis work. His research work contributed to the reputed journal of *Scientific Reports* (Nature), *Knowledge-Based Systems*, *Artificial Intelligence Review*, IEEE Access, *Sensors*, *Cancers*, and *Applied Sciences*. His research interests include artificial intelligence (AI), machine learning, deep learning, big data analysis, and natural language processing (NLP). He has served as a program committee member in several international conferences/workshops. He served as an Associate Editor for several journals, including *PLOS One* journal. He has served as a Reviewer for reputed journals and IEEE conferences, such as HONET, ICIEV, ICCIT, IJCCI, ICAEE, ICCAIE, ICSIPA, SCORED, ISIEA, APACE, ICOS, ISCAIE, BEIAC, ISWTA, IC3e, CoAST, icIVPR, ICSCT, 3ICT, and DATA21.