Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Attention-based CNN-BiGRU for Bengali Music Emotion Classification

Subhasish Ghosh, Md. Omar Faruk Riad
subhasish@bgctub.ac.bd, omarfaruk@bgctub.ac.bd
Department of Computer Science and Engineering, BGC Trust University Bangladesh

Article Information

Submitted : 07 Nov 2022
Reviewed: 17 Nov 2022
Accepted : 15 Dec 2022

Keywords

Conv1D; BiGRU; MFCCs; Bahdanau attention mechanism; Bengali music classification.

Abstract

For Bengali music emotion classification, deep learning models, particularly CNN and RNN are frequently used. To extract meaningful knowledge, however, past studies' shortcomings of low accuracy and overfitting have to be addressed. We have proposed a model combining Conv1D, Bi-GRU and the Bahdanau attention mechanism for music emotion classification of our Bengali music dataset. The model integrates distinct MFCCs wav preprocessing methods with deep learning methods and attention-based methods. The attention mechanism has increased the accuracy of the proposed classification model. The music is finally classified into one of the four emotion classes: Angry, Happy, Relax, Sad. The proposed Conv1D+BiGRU+Attention model is validated as more effective and efficient at classifying emotions in the Bengali music dataset than baseline methods, according to comparisons with baseline models. For our Bengali music dataset, the performance of our proposed model is 95%.

A. Introduction

Emotion is the formation of human feelings and sentiments that have an influence on how people behave over time and trigger physical and psychological changes [18]. According to statistics from the Harvard, Stanford, and Carnegie foundations, emotional intelligence is responsible for 85-87% of our success [16]. The maximum top performers have excellent emotional intelligence [17]. Emotion is the most important component of songs, beyond rhythm, tune, fusion, performer, and genre, as it directly affects listeners' mood and decision. Most of the Bengali music genres can be classified into four types depending on emotions like Angry, Happy, Relax and Sad. But the majority of the songs that are now accessible are not automatically classified into proper emotions. So, the ability to quickly classify musical emotions is crucial in today's music industry. Several researches have innovated varieties of Music Emotion Classification (MEC) Systems using different types of deep learning models. BiGRU model, convolutional long short-term memory deep neural network (CLDNN) model, CNN-LSTM model, etc. are recently proposed deep learning algorithms to classify musical emotions [26][38][39]. But previous researches had the flaws of low accuracy and overfitting problem. In this research, attention-based Conv1D and BiGRU model is designed for music emotion classification and comparative experimentation shows that the proposed model is classifying emotions more accurate. A Bangla emotional musical dataset has been created using different Bengali music genres and features vectors are extracted from wav files by using Mel Frequency Cepstral Coefficients (MFCCs). The MFCCs vectors are used to train the proposed model.

B. Related Works

A deep neural network made up of CNN+LSTM+DNN serves as the system's main component for classifying music emotions. The whole model's recognition accuracy is roughly 90%, which produces the desired result [19]. In a research [20] evaluated the precision of several classifiers for aesthetic classification by extracting appropriate acoustical data. According to the findings, logistic regression had the best accuracy at 65.37%. The gradient descent-based Spiking Neural Network (SNN) classifier identifies the ideal weight values for lowering the training error when learning. The proposed approach achieves 94.55% [21]. Another research categorizes and quantifies music emotion based on the mapping link between musical qualities and emotion through the quantification of musical aspects. According to simulation data, the model's accuracy in identifying emotions is up to 93.78% [22]. Multiple linear regression outperforms support vector regression in terms of accuracy, providing a 61.29 % accuracy rate with a precision of 65 %, recall of 61 %, and f-measure of 60 % [23].

CNN, gcForest, ResNet50 and ResNet50_trust achieved accuracy is 59.20%, 65.67%, 66.37% and 71.56% respectively in another comparative study of music emotion classification [24]. Using an upgraded Deep Belief Network to extract features for classification provides increased classification accuracy 79.4% while lowering the possibility of overfitting [25]. The BiGRU emotion recognition model is developed in another research, and it is compared to other models. Up to 79 % and 81.01 % of the time, respectively, BiGRU can properly detect music with joyful and sad emotions [26]. According to the findings, Naive Bayes had the highest

classification accuracy for musical mood, at 86.64% [27]. The upgraded DBN network combined with the SVM classification algorithm can classify music emotions with an accuracy of 88.31% [28]. The experimental findings demonstrate that the suggested model performs competitively on the Bi-modal (84.91%), 4Q emotion (92.04%), and Soundtrack (87.24%) datasets [29]. Jia, X. have presented a deep learning-based and better attention mechanism-based technique for categorizing musical emotions. The output results of the CNN-LSTM model and deep neural network are combined, to achieve average classification accuracy is 84.8% [30]. A convolutional neural network-based music emotion identification algorithm is proposed in this paper. The proposed method's recognition accuracy is 92.06 %, and its loss function value is about 0.98 [31]. But Shallow neural networks outperform a range of regression models, with the highest performing networks accounting for 64.4 % in arousal prediction and 65.4 % in valence prediction [32].

An RNN-based MIDI music emotion classification method achieved accuracy of up to 75.4% [33]. Liu, H., Fang, Y. and Huang, Q. have trained and tested in two label spaces: calm-excited and joy-sad. This can predict music excitement with an accuracy of 81% and joylessness with an accuracy of 78% [34]. In terms of valence and arousal, a simpler MLP binary classifier produces results that accurately identify the positive and negative regions of the V/A plane (valence is correctly classified at 73%, and arousal at 69%) [35]. A CNN-LSTM combined network to classify music emotions and offers a multifeature combined network classifier accuracy is 74% [36]. The results show that the MFCC feature combined with RNN increases instrument emotion recognition performance, with a recognition rate of 89.3 %. On other hand, MFCC with SVM produced an 85.7% recognition rate [37]. According to the experimental findings, the proposed approach has a recognition accuracy of 71% for joyful emotions and 68.8% for sad emotions [38]. For three experiments, MECS 1 obtains accuracy of 91 %, 88 %, and 86 %, MECS 2 obtains accuracy of 87 %, 82 %, and 79 %, MECS 3 obtains 82.3 %, 82 percent, and 81.6 % accuracy [39] [40].

C. Proposed Architecture

This research has proposed the architecture illustrated in Fig. 1.

3.1 Wav Pre-processing

Loudness, pitch, and timbre—three characteristics of musical signals—can be captured by these short-time feature parameters. The music signal is time-varying and unstable, but the region of the signal between 10 ms to 30 ms is often smooth, our all-Bangla music's wav is the duration of 30s, therefore the spectrum waveform may be seen as a short-time and smooth operations. Pre-Emphasis is the first step in a noise reduction method called pre-emphasis is used to boost the high-frequency component of a signal [1][2]. Framing & Windowing is the technique of framing means dividing the audio into small frame, the process of framing involves breaking the voice stream into brief frames, which are generally between 5 and 50 milliseconds length [3]. Windowing is the technique of adjusting each frame's window to minimize discontinuities and leakage at start and end of the each frame [4][5].

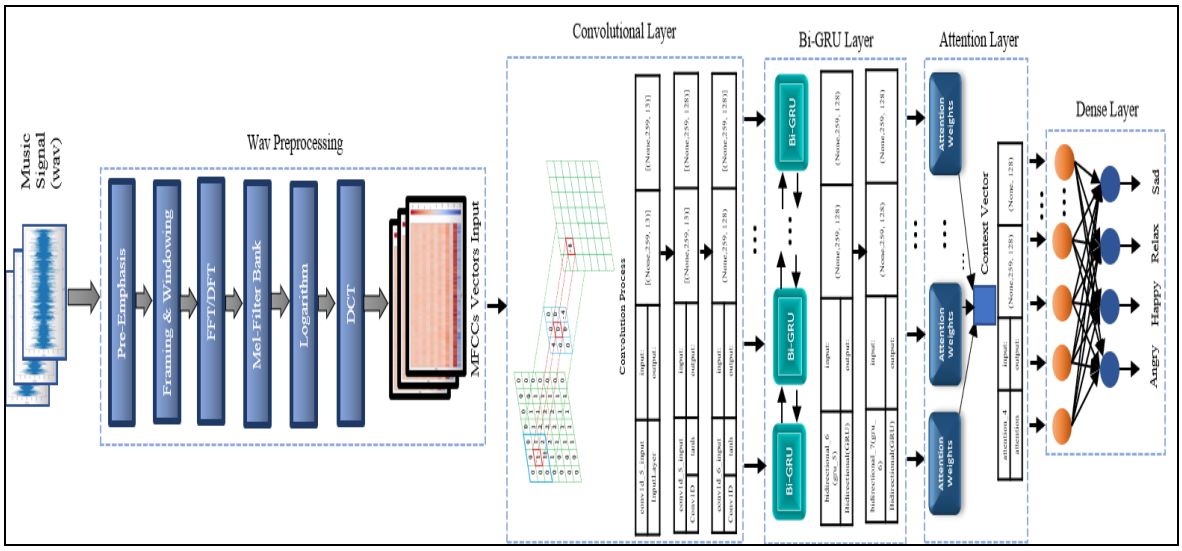

Fig 1: The overall architecture of the attention-based Conv1D-BiGRU model

FFT/DFT is an algorithm known as fast Fourier transform (FFT) computes a sequence's Discrete Fourier transform (DFT) or its inverse (IDFT)[6].In our research our all-music wav's sample rate is 44.1 kHz, window length is 512 which means 512 samples in each frame. After obtaining the signal's spectrum value, the energy spectrum is created by square rooting the result [7]. Mel filter banks are made up of m triangular filter banks that adhere to the Mel frequency scale. We use the logarithm of the output signal to derive the spectrum estimate error with significant robustness [10]. A Discrete Cosine Transform (DCT) is applied to the filter banks in order to create MFCCs, with number of resultant coefficients and rest being discarded [11]. Higher sounds have higher pitches, and lower sounds have lower pitches. The Chroma filter bank can be used to create Chroma filters. All the energy of the recorded sound is supposed to be projected into 12 bins by the filter bank. The all 12 bins notes are A, A#, B, C, C#, D, D#, E, F, F#, G, G#.

In this research the number of MFCCs coefficient is 13 for each frame, number of intervals for FFT is 2048, window length is 512 and the number of segments is 5. Here total number of MFCC= 3098, divided into Training MFCC= 1858, Validation MFCC=620 and Testing MFCC=620. The MFCC vectors are stored in a JSON formatted dataset. The data dictionary of the JSON dataset [43] is

```
data = {
  "mapping": [],
  "labels": [],
  "mfcc": []
 }
```

In the above, "mapping" stores the class name indicates the "mfcc" vector belongs to one of angry, happy, relax and sad classes. The "labels" are numerical values representing each classes serially 0,1,2,3.

3.2 Convolutional neural network

Row-based MFCCs Vector representations are given to the two convolutional layers. Assume that the ith row input feature matrix for a single channel is $V_i \in R^d$, where V_i

is the ith music wav's d-dimensional feature vector. $V_{i:i+l} \in R^{i*d}$. The formulation of feature h_i following extraction is as follows:

$$h_i = f(W * V_{i:i+l} + b + z) \quad (1)$$

We have total two convolutional layers in our research. The following is a list of the feature that the CNN layers provided:

$$h_n = [h_1, h_2, \ldots, h_n], n = 1 \quad (2)$$
$$h_n = [h_1, h_2, \ldots, h_n], n = 2 \quad (3)$$

3.3 Gated recurrent unit (GRU)

GRU is a variation of LSTM that is similar but it only contains two gates: update gates (z_t) and reset gates (r_t) [13]. Together, these two gates regulate the state's information updating process. The contribution of the past state h_{t-1} to the candidate state $\tilde{h}_t$ is controlled by reset gate r_t, and the smaller the value, the greater the ignored rate. The reset gate operates as follows at time t.

$$r_t = \sigma (W_r h_{t-1} + U_r x_t + b_r) \quad (4)$$

To decide how much of the stream's data can be preserved or forgotten for at timestamps t and t-1, the update z_t is used. The expression of update gate as follows:

$$z_t = \sigma (W_z h_{t-1} + U_z x_t + b_z) \quad (5)$$

Here, σ is the logistic sigmoid function. x_t represents the input and h_{t-1} represents previous hidden states. Equation (6) is used to calculate the GRU state at the time interval t and the candidate state $\tilde{h}_t$ is calculated using Eq (7).

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot + h_{t-1} \quad (6)$$
$$\tilde{h}_t = \tanh (W_h (h_{t-1} \odot r_t) + U_h x_t + b_h) \quad (7)$$

W, U are learning weights, and b is the bias term, where $\odot$ is multiplication of the vector element.

3.4 Bi-GRU

In our proposed methodology Bi-GRU has been used to gain richer vector information, and it generalizes a bit more quickly or with less data. Bi-GRU produces feature vector matrix from the vectors acquired from the previous steps [12].

$$\vec{h}_{f_{gru}} = \overrightarrow{GRU} (p_n), n \in [1, m] \quad (8)$$
$$\overleftarrow{h}_{b_{gru}} = \overleftarrow{GRU} (p_n), n \in [m, 1] \quad (9)$$

Here, equation (8) and (9) represents forward and backward GRUs respectively. For each input vector, we now obtain music score annotation.

$$h_{t_{gru}} = GRU [\vec{h}_{f_{gru}}, \overleftarrow{h}_{b_{gru}}] \quad (10)$$

where, $h_{t_{gru}}$ is the concatenating output. This output of the long dependencies feature that is extracted from BiGRU by forwarding ($\vec{h}_{f_{gru}}$) and backward ($\overleftarrow{h}_{b_{gru}}$).

3.5 Attention Mechanism

In this research, the Bahdanau attention mechanism also known as additive attention is used [41, 42]. Because various MFCCs vectors have various meanings, in order to differentiate the emotion of each music, we apply attention on Bi-GRU generated features. The $h_{t_{gru}}$ music score annotation is passed through one layer perceptron to get $u_{t_{gru}}$ as the hidden representation of $h_{t_{gru}}$ th input that formulated as follows:

$$u_{t_{gru}} = \tanh (w * h_{t_{gru}} + b) \quad (11)$$

where w stands for weight and b for bias. The single vector created by combining the important weights A_t into V is indicated by the Eq (12) [15].

$$V = \sum(A_{t_{gru}} * h_t) \quad (12)$$

3.6 Dense Layer

This layer completes the music emotion classification operation based on the output from attention layer, as shown in Fig.1. The softmax function is taken into consideration for multi-class music emotion classification, while sparse categorical cross-entropy is responsible for measuring the loss and difference between the actual and predicted appropriate emotion - Angry or Happy or Relax or Sad.

D. Experiments

To find the proper audio data for this system, the emotion of a track is mainly focused. There are four Emotion types in the dataset: Angry, Happy, Relax and Sad.

4.1 Song Selection Method

The Bengali Robindra Sangeet, Nazrul Geeti, Lalon Geeti and many other Bengali Folk Music etc. are selected. They are using Dotara, Tabla-bayan, Harmonium, Bansuri, Sitar etc. instruments. In this research of Bengali modern music, it is discovered that emotions are much more extensive in this updated Bangla music. They used Violin, Keyboard, Guitar, Bass, Saxophone, Harp, Flute, Cello, Drums and so many varied instruments, also used so many Virtual Studio Technology (VST) and different types of voice textures to express the various types of emotions in music. As a result, the range of emotions in Bengali songs is expanded. The five primary techniques in the Bengali song collection are taken into consideration as follows. The scale is a basic technique for musical tonic construction. Average Energy (AE) refers to the entire wave sequence's average energy (AE) is commonly used to determine the loudness of the music. Rhythm is the tempo or beat of a song. The beats per minute are commonly used to determine tempo (BPM). Generally soft, relax, sad or upset songs have slow Beats per minute (tempo). On another side, normal or more than normal amount of beat can be used for happy or joyful songs. Aside from that, fast tempo is also used for excited or cheerful songs. On other hand,

the speediest tempo is generally used for angry or grumpy songs. The Voice texture of a song is very important for deciding the emotion of a song. We also concentrated on how musical instruments have been played in the music to express the emotion of a song.

4.2 Rendering Process

Firstly, we have listened to a whole song from our collected Bangla song. Then we find out a best part of a song which can express a particular emotion better. We have used "Reaper" as Digital Audio Workstation (DAW) for rendering each Bangla song.

Table 1. Rendering Process settings parameters.

Command	Parameters
1. Time Selection	30 seconds
2. Render Bounds	Time Selection
3. Directory	Particular Emotion folder
4. File Name	Particular Emotion and set a serial number
5. Sample Rate	44100 Hz
6. Channels	Stereo
7. Resample Mode	512pt HQ Sinc
8. Output Format	WAV
9. WAV Depth	16-bit PCM

Our all-rendering wav sample rate is 44100 Hz which is the standard sample rate of the modern music production.

SAMPLES_PER_TRACK=SAMPLE_RATE*TRACK_DURATION

In pre-processing all wav signals contains the values of multiplication results of sample rate and duration of a track. So, our each signal or sample per track is 1,323,000 (44100*30).

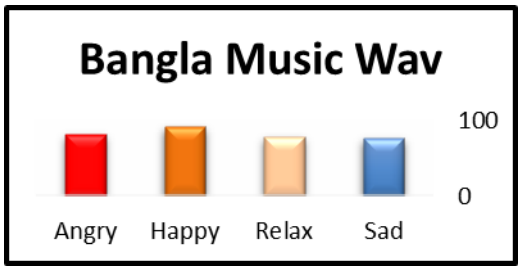

Fig 2. Number of wav files for each emotion
[Angry-81, Happy-91, Relax-78, Sad-76]

***4.3** Development Enviroment*

We have used google colaboratory environment in our all experimentation. The best feature of google colaboratory gives acces to powerful deep learning utilities like GPUs and TPUs.

***4.4** Training procedure and hyper-parameters setting*

The hyper-parameters settings for the proposed model are given in Table II. The proposed model and baseline models are implemented using keras API developed in python programming language. The grid search optimization methods are used to fine tune the hyperparameters. In keras, the input layer is a tensor. The input layer is the starting tensor sent to the first hidden layer. This tensor should have the equal shape as the training data. For training our attention-based convolutional BiGRU model 60% of dataset used for training, 20% used for validation and 20% used for testing the model. All extracted MFCCs vectors are passed as input to the model's convolution layer.

Conv1D and BiGRU require the input shape with a certain number of dimensions like (batch size, sequence length, features). The input shape of the training data used for training the proposed Conv1D+BiGRU+Attention model is (1858, 259, 13), for validation (620,259,13) and for testing (620,259,13). As the batch size is 64 configured for both training and testing henceforth for each of the iterations during training the input shape will be (64, 259, 13) and this is same for validation as well as testing. For training the proposed model we have used 64 batch size and 100 numbers of epochs. The number of training data in one forward or one backward pass is referred to as the batch size since one epoch is defined as one forward pass and one backward pass of all the training data, and each of the iterations is referred to as the number of passes where each pass is using batch size number of training data. The amount of samples that are propagated through the network is likewise determined by the batch size. The network's parameter is updated after one epoch when the total amount of training data is passed. Due to the significantly increased quantity of updates, the network converges more quickly.

Table 2. Hyperparameters settings for the proposed model

Hyperparameter	**Value**
Training input shape	1858×259×13
Number of epochs	100
Batch Size	64
Number of Conv1D Layers	2
Number of Filters	128
Convolutional kernel size	3
Number of BiGRU layer	2
BiGRU output shape per epoch	64×259×128
Dropout Rate	0.3
Number of Attention Mechanism	1
Number of Dense Layer	1
Learning Rate	0.001
Optimizer	Adam (AMSGrad=True)
Activation function types	tanh, softmax
Loss Function	keras.losses.SparseCategoricalCrossentropy (from_logits=False)

The kernel sizes of 3 for each of the 128 filters (kernels) are assigned to Conv1D layers consequently 128 different convolutions will take place. Different values in filters will create different output feature maps. In keras, the convolution and activation layers are added serially and activation is applied after each convolution. In each convolution the filter slides across the input and the filter is multiplied elementwise with the input at every position and the resulting values are summed up to get a single element of the output feature map. The tanh activation function is set for convolution layers in the proposed model so the output feature maps will be passed to the tanh activation function after convolution. The proposed model has a stack of 2 Conv1D layers which is done by settings of padding="same" parameter. If the padding is configured as such the output shape has the same dimension as the input.

The two BiGRU layers are configured to output the tensors with the shape of (64,259,128) in each epoch where batch size=64; timesteps=259; hidden state=128. The number of memory units is specified as 64 which are resulting in (64+64=128) for being bidirectional. The BiGRU sequence processing model consists of two GRUs taking inputs in forward and backward directions both. To stack two BiGRUs in keras return_sequences=true attribute is added so that all BiGRU layers except the last one output tensor has ndim=3. The tanh activation function is set for the BiGRU layers to get output of the network nodes. A weight constraint is employed as well as dropout with a dropout rate of 30% on these layers. After receiving features vectors from the bidirectional GRU layer the attention mechanism is applied on the received features vectors. When given a list of MFCCs, the generalized attention mechanism scores each key in the database using the query vector assigned to each individual MFCC features in the list. This depicts the relationship between the MFCC feature under examination and the other MFCC features in the sequence. The values are then scaled in accordance with the attention weights (calculated from the scores) in order to maintain the focus on the query-relevant features. As a result, the feature under consideration receives an output of attention. The attention layer will help training the model on the "context" of the MFCC feature vectors by returning the weighted sum as the output to the feedforward fully connected dense layer.

The dense layer output is transferred through the softmax activation function is used for multi-class classification and loss function Sparse Categorical Crossentropy is used to measure the performance of the multi-classifier model.

The AMSGrad optimizer is used to eliminate the need to manually tune the weights by reducing the overall loss and improve the accuracy. We have used Adam (AMSGrad=True) optimizer in our model where the learning rate is 0.001.

E. Analysis of Experiment Results

The experiment is executed on our Bengali music dataset to evaluate the model accuracy, recall, precision, and F1 score. The effect of the attention mechanism on the weight matrix improves the accuracy of the Attention based model. The experiment result validates that the proposed model (Conv1D+BiGRU+Attention) performed better and achieved better accuracy than the other methods in the Bengali music dataset, as can be seen from Table IV, in which the bolder entries

indicate the highest evaluation metrics like F1 score of each emotion classes and accuracy of the proposed model.

In Fig. 3, the confusion matrix shows that the proposed Conv1D+BiGRU+Attention model is doing the best in classifying the Relax emotion class of music. Table III, an overall glance of the performance of our model shows that the model accuracy is 95% having macro average of precision, recall, F1 score for all four classes (Angry, Happy, Relax, Sad) are 94.5%, 94% and 94.25% respectively. The high precision of each class shows that correct prediction of each class out of all predictions of that specific class, the high recall of each class shows that the number of correctly predicted specific music emotion class out of the number of that actual specific music emotion class in the dataset. The higher F1 score shows that the proposed model performs effective in classification of four emotion classes in the Bengali music dataset.

In Fig. 4, the ROC-AUC curve visualizes the high effectiveness of our model to distinguish among four emotion classes. The average ROC AUC score of the proposed model is 0.93873. The AUC is high which implies that the performance of the model is 93% capable of distinguishing among Angry, Happy, Relax and sad emotion classes.

Table 3. F1-Score of classification
[0-Angry, 1-Happy, 2-Relax, 3-Sad]

	Precision	**Recall**	**F1-Score**	**Support**
0	0.92	0.88	0.90	136
1	0.92	0.92	0.92	118
2	0.99	0.99	0.99	180
3	0.95	0.97	0.96	186
Accuracy			0.95	620
Macro avg	0.94	0.94	0.94	620
Weighted avg	0.95	0.95	0.95	620

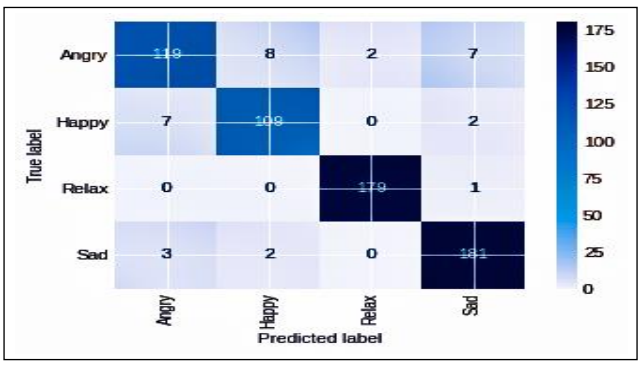

Fig 3. Confusion Matrix

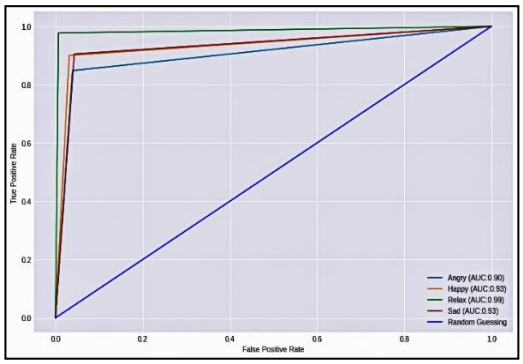

Fig 4. ROC-AUC Curve

The Fig. 5 is showing the convergence for our proposed model that attained reliable accuracy. The proposed model converged toward the optimal solution

after 100 epochs with consistent accuracy. This is evident that the Conv1D+BiGRU+Attention network reduces the overfitting problem and attains adequate accuracy.

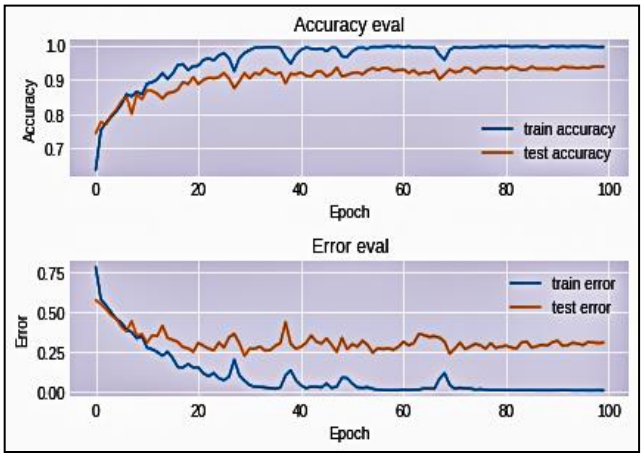

Fig 5. Train and Test Accuracy per epoch

5.1 *Comparison*

Table IV, shows that our proposed Conv1D+ BiGRU+ Attention model has achieved the highest accuracy of 95%. Our Conv1D+BiGRU+Attention model has outperformed with the highest accuracy and F1 score according to comparisons with baseline methods. Observations shown by the comparison implies that combining attention with the BiGRU and CNN has allowed the proposed model to focus on the context of the MFCC vector input samples enabling to recognize the pattern by accurately classifiable emotion of the music. The Bahdanau attention mechanism produces a context vector by performing weighted sum of the MFCC vectors where weights are calculated by the softmax activation function given the inputs as scores calculated by computing the dot product between the query vectors and key vectors. Thus each sample in the context vector is scaled according to the calculated attention weights to have focus on those samples.

Table 4. Comparative analysis of accuracy of the baseline models and proposed model [Conv1D+BiGRU+ Attention] on Bengali Music Dataset

Models	**F1-Score**				**Accuracy**
	Angry	***Happy***	***Relax***	***Sad***	
BiGRU+Attention+ Conv1D	0.86	0.87	0.98	0.90	0.90
BiLSTM+Attention [30]	0.86	0.87	0.99	0.90	0.91
BiGRU+Attention	0.82	0.86	0.99	0.87	0.89
Conv1D+Attention	0.86	0.91	0.99	0.91	0.92
CNN [39]	0.66	0.75	0.97	0.64	0.75
BiGRU [26]	0.73	0.85	0.97	0.83	0.85
CNN+LSTM [36]	0.68	0.81	0.98	0.79	0.81
CNN+GRU	0.70	0.82	0.97	0.83	0.84
CNN+BiGRU	0.72	0.83	0.99	0.84	0.84
Conv1D+BiGRU+ Attention	**0.90**	**0.92**	**0.99**	**0.96**	**0.95**

F. Future Work and Recommendation

Future works mainly includes the following: (1) building transformer model for Bengali speech emotion recognition; (2) comparing with more wav preprocessing methods like Short Term Fourier Transformation (STFT), Constant-Q transform spectral envelope coefficients (CQT-SEC); (3) integration of deep generative model of raw audio waveforms in music emotion classification transformer models. The features included in the MFCC vectors are loudness, pitch and timbre. These are also the characteristics of music signals collectively used to classify the emotion of music. In confusion matrix Fig 3. we can see that the emotions angry and happy are sometimes incorrectly classified as those features are not differentiated properly by the model. Because these feature sets could be overlapped in the MFCC feature vectors. Experimenting with different types of wav pre-processing might help developing an efficient wav file pre-processing technique to increase the ability to discriminate among different emotion classes in the music.

G. Conclusion

In conclusion, the experimental results have shown that the attention based CNN-BiGU model can be used in Bengali music classification with higher accuracy and F1 score. Our Bengali emotion classification dataset is analyzed by the various peer reviewed deep learning models including our proposed Conv1D+BiGRU+Attention model. By comparison it can be inferred that by combining attention with the BiGRU and CNN has enabled the proposed model to classify emotion of Bengali music more accurately. The Conv1D layers have done dimensionality reduction of MFCC vectors which are input to BiGRU layers to learn the sample sequence in the MFCC vectors and context vectors are produced by applying attention mechanism. Finally, using the context vectors the fully connected dense layer has classified the emotion of the MFCC vector.

H. References

[1] S. R. Gulhane, S. D. Shirbahadurkar, and S. Badhe Sanjay. Self organizing feature map network for musical instrument sounds. *International journal of innovative technology and exploring Engineering*, vol. 8, no. 9S3, pp. 143–146, 2019.

[2] P. Y. Raj, B. Bhuwan, and L. Joonwhoan. Deep-learning-based multimodal emotion classification for music videos. *Sensors (Basel, Switzerland),* vol. 21, no. 14, pp. 4927–4931, 2021.

[3] Rana, D. and Jain, A. Effect of windowing on the calculation of MFCC statistical parameter for different gender in Hindi speech. *International Journal of Computer Applications, 98*(8), 2014.

[4] Jain, A., Prakash, N. and Agrawal, S.S. May. Evaluation of MFCC for emotion identification in Hindi speech. In *2011 IEEE 3rd International Conference on Communication Software and Networks* (pp. 189-193), 2011.

[5] Lee, D. Hornbostel-Sachs classification of musical instruments. *Knowledge Organization, 47*(1), pp.72-91, 2019.

[6] Heideman, Michael T.; Johnson, Don H.; Burrus, Charles Sidney ."Gauss and the history of the fast Fourier transform",1984.

[7] Ying, M., Kaiyong, L., Jiayu, H. and Zangjia, G. Analysis of Tibetan folk music style based on audio signal processing. *Journal of Electrical and Electronic Engineering, 7*(6), pp.151-154,, 2019.
[8] Prabavathy, S., Rathikarani, V. and Dhanalakshmi, P. Classification of Musical Instruments using SVM and KNN. *International Journal of Innovative Technology and Exploring Engineering, 9*(7), pp.1186-1190, 2020.
[9] Li, J., Luo, J., Ding, J., Zhao, X. and Yang, X.. Regional classification of Chinese folk songs based on CRF model. *Multimedia tools and applications, 78*(9), pp.11563-11584, 2019.
[10] Cheah, K.H., Nisar, H., Yap, V.V. and Lee, C.Y.. Convolutional neural networks for classification of music-listening EEG: comparing 1D convolutional kernels with 2D kernels and cerebral laterality of musical influence. *Neural Computing and Applications, 32(*13), pp.8867-8891,2020.
[11] Tamboli, A.I. and Kokate, R.D. An effective optimization-based neural network for musical note recognition. *Journal of Intelligent Systems, 28*(1), pp.173-183, 2019.
[12] Kamyab, M., Liu, G., Rasool, A. and Adjeisah, M. ACR-SA: attention-based deep model through two-channel CNN and Bi-RNN for sentiment analysis. *PeerJ Computer Science, 8*, p.e877, 2022.
[13] Dey, R. and Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* (pp. 1597-1600), 2017.
[15] Liu, J., Yang, Y., Lv, S., Wang, J. and Chen, H. Attention-based BiGRU-CNN for Chinese question classification. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-12, 2019.
[16] Hunckler, M., [Updated 2017 Feb 20]. Emotional Intelligence: Your Secret Weapon For Success In Business And Life. Available from: https://www.forbes.com/sites/matthunckler/2017/02/20/emotional-intelligence-in-business-and-life/?sh=3516c1687f6c
[17] DSilva, A., [Updated 2019 Nov 08]. Did you know that 90% of top performers have a high EQ? Available from: https://www.capacityhr.co.uk/did-you-know that-90-of-top-performers-have-a-high-eq#:~:
[18] Ackerman, E.C [Updated 2018 March 12]. Positive Emotions: A List of 26 Examples & Definition in Psychology. Available from: https://positivepsychology.com/positive-emotions-list-examples-definition -psychology/
[19] Yang, S., He, D. and Zhang, M. A Speaker System Based On CLDNN Music Emotion Recognition Algorithm. In *ICETIS 2022; 7th International Conference on Electronic Technology and Information Science* (pp. 1-7). VDE, 2022.
[20] Xie, L. and Gao, Y. A database for aesthetic classification of Chinese traditional music. *Cognitive Computation and Systems,* 2022.
[21] Tiple, B. and Patwardhan, M. Multi-label emotion recognition from Indian classical music using gradient descent SNN model*. Multimedia Tools and Applications, 81*(6), pp.8853-8870, 2022.
[22] He, J., 2022. Algorithm Composition and Emotion Recognition Based on Machine Learning. *Computational Intelligence and Neuroscience, 2022*.
[23] Satayarak, N. and Benjangkaprasert, C. On the Study of Thai Music Emotion

Recognition Based on Western Music Model. In *Journal of Physics: Conference Series* (Vol. 2261, No. 1, p. 012018). IOP Publishing, 2022.
[24] Li, J., Han, L., Li, X., Zhu, J., Yuan, B. and Gou, Z. An evaluation of deep neural network models for music classification using spectrograms. *Multimedia Tools and Applications*, *81*(4), pp.4621-4647,2022.
[25] Wu, Z., 2022. Research on automatic classification method of ethnic music emotion based on machine learning. *Journal of Mathematics, 2022.*
[26] Niu, N., 2022. Music Emotion Recognition Model Using Gated Recurrent Unit Networks and Multi-Feature Extraction. *Mobile Information Systems, 2022.*
[27] Wang, C. and Ko, Y.C. Emotional representation of music in multi-source data by the Internet of Things and deep learning. *The Journal of Supercomputing*, pp.1 18, 2022.
[28] Tong, G. Music Emotion Classification Method Using Improved Deep Belief Network. *Mobile Information Systems, 2022.*
[29] Liao, Y.J., Wang, W.C., Ruan, S.J., Lee, Y.H. and Chen, S.C. A Music Playback Algorithm Based on Residual-Inception Blocks for Music Emotion Classification and Physiological Information. *Sensors, 22*(3), p.777, 2022.
[30] Jia, X. Music Emotion Classification Method Based on Deep Learning and Improved Attention Mechanism. *Computational Intelligence and Neuroscience, 2022.*
[31] Abdullah, S.M.S.A., Ameen, S.Y.A., Sadeeq, M.A. and Zeebaree, S. Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends, 2(02), pp.52-58, 2021.
[32] Zhao, W., Zhou, Y., Tie, Y. and Zhao, Y. Recurrent neural network for MIDI music emotion classification. In 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 2596-2600), 2018.
[33] Cunningham, S., Ridley, H., Weinel, J. and Picking, R. Supervised machine learning for audio emotion recognition. Personal and Ubiquitous Computing, 25(4), pp.637-650, 2021.
[34] Liu, H., Fang, Y. and Huang, Q. Music emotion recognition using a variant of recurrent neural network. In 2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018). Atlantis Press, 2019.
[35] Medina, Y.O., Beltrán, J.R. and Baldassarri, S. Emotional classification of music using neural networks with the MediaEval dataset. Personal and Ubiquitous Computing, pp.1-13, 2020.
[36] Chen, C. and Li, Q. A multimodal music emotion classification method based on multifeature combined network classifier. Mathematical Problems in Engineering, 2020.
[37] Rajesh, S. and Nalini, N.J. Musical instrument emotion recognition using deep recurrent neural network. Procedia Computer Science, 167, pp.16-25, 2020.
[38] Jia, X., 2022. Music Emotion Classification Method Based on Deep Learning and Explicit Sparse Attention Network. Computational Intelligence and Neuroscience, 2022.
[39] Chaudhary, D., Singh, N.P. and Singh, S. Development of music emotion classification system using convolution neural network. International Journal of Speech Technology, 24(3), pp.571-580, 2021.

[40] Na, W. and Yong, F., 2022. Music Recognition and Classification Algorithm considering Audio Emotion. Scientific Programming, 2022.
[41] Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y. Attention based models for speech recognition. Advances in neural information processing systems, 28, 2015.
[42] Bahdanau, D., Cho, K., Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
[43] Ghosh, S., and Riad, F. O. Bengali Emotion Classification Dataset, 2022. Available from:
https://drive.google.com/file/d/1aC2RM8s4uNR5UjvyzzkNaWks5lU1hcli/view?usp=sharing