MULTIDISCIPLINARY SCIENTIFIC RESEARCH
BJMSR VOL 9 NO 3 (2024) P-ISSN 2687-850X E-ISSN 2687-8518
Available online at https://www.cribfb.com
Journal homepage: https://www.cribfb.com/journal/index.php/BJMSR
Published by CRIBFB, USA

MUSIC EMOTION RECOGNITION AND CLASSIFICATION USING HYBRID CNN-LSTM DEEP NEURAL NETWORK

Jumpi Dutta [(a)1] *Dipankar Chanda* [(b)]

[(a)]*Research Scholar, Department of Electrical Engineering, Assam Engineering College, Guwahati, Assam, India; E-mail: jumpiofcit@gmail.com*
[(b)]*Professor, Department of Electrical Engineering, Assam Engineering College, Guwahati, Assam, India; E-mail: dchanda2007@rediffmail.com*

ARTICLE INFO

Article History:

Received: 14th May 2024
Reviewed & Revised: 14th May to 25th July 2024
Accepted: 27th July 2024
Published: 3rd August 2024

Keywords:

Emotion Recognition, Assamese, Feature Extraction, Mel Frequency Cepstrum Coefficient, Deep Learning, Long Short-Term Memory Network, Convolutional Neural Network.

JEL Classification Codes:

C880, Y100, Y800

Peer-Review Model:

External peer review was done through double-blind method

ABSTRACT

In music information retrieval (MIR), emotion-based classification is a complex and challenging task for researchers. In modern-day information technology, understanding music emotions through human-computer interaction plays a vital role in capturing the attention of both researchers and the music industry. This paper presents a learning algorithm by adopting the decision tree, random forest, k-nearest neighbors, multi-layer perceptron (MLP), long short-term memory (LSTM) neural network, convolutional neural networks (CNNs), and CNN-LSTM hybrid deep learning approaches with relevant feature extraction techniques. Many researchers have performed emotion recognition in music for different languages such as English, Chinese, Spanish, Turkish, Hindi, etc. However, languages like Assamese have drawn very little attention in the research of music emotion recognition (MER). This work aims to perform a novel approach to emotion recognition in Assamese songs. In this study, a newly created Assamese dataset of 200 song samples is used with four different emotions and another dataset used is the RAVDESS emotional song database which consists of 1012 song samples with six different music emotions. Relevant features such as mel-frequency cepstrum coefficients (MFCC), mel spectrogram, and chroma features are extracted from the song samples to investigate the performance of the proposed method. A comparative analysis using different classifiers is carried out and the findings of this study suggest that the CNN-LSTM model has shown better accuracy with both datasets. The accuracy is 85.00% with the Assamese dataset, and with the RAVDESS dataset, the accuracy is 89.66% compared to the other classifiers used in this work.

INTRODUCTION

Emotion is a state of feeling that reflects the conscious mental reaction based on an event, a situation, or a set of circumstances, accompanied by physiological and behavioral changes in the body (Domínguez-Jiménez et al., 2020). Music emotion recognition (MER) is one of the important subfields of music information retrieval (MIR) that has grown over the last few years to enhance human-machine interaction (HMI). The emotion recognition task in music has many applications, such as music recommender systems, automated playlist generation, music therapy, mental health monitoring, and so on (Hizlisoy et al., 2021; Modran et al., 2023). With enormous song generating every day on the internet and offline digital music, the task of organizing and retrieving information from such music becomes quite challenging. Thus, music emotion recognition is an emerging area of research with several issues that need to be dealt with, such as emotion labelling, selection of appropriate feature extraction techniques, and classification algorithm.

Traditional music retrieval methods of retrieving music information are widely used based on classification tags such as artists' names, song names, and album names. However, these traditional methods are considered insufficient to meet people's demand for advanced and personalized music recommendations. As a result, new approaches have been needed to enhance the music retrieval experience, and therefore, many music websites have introduced music recommendation services. In recent years, various listening platforms have provided song recommendation services for different moods by analysing users' preferences and listening history (Jitendra & Radhika, 2021).

In MER, one of the researchers' challenges is dealing with copyrighted audio data. The datasets available in various studies have restrictions in the modification and distribution of the audio material. Researchers may be unable to share the

[1]Corresponding author: *ORCID ID: 0009-0008-8265-9242*

https://doi.org/10.46281/bjmsr.v9i3.2230

To cite this article: Dutta, J., & Chanda, D. (2024). MUSIC EMOTION RECOGNITION AND CLASSIFICATION USING HYBRID CNN-LSTM DEEP NEURAL NETWORK. *Bangladesh Journal of Multidisciplinary Scientific Research*, *9*(3), 21-32. https://doi.org/10.46281/bjmsr.v9i3.2230

dataset due to legal constraints, which may restrict other researchers from accessing the same data for further research. Researchers may consider the annotations and metadata obtained from social music websites to evaluate MER algorithms. Moreover, the quality and reliability of annotation from such sources may vary (Aljanaki et al., 2017).

A self-built Assamese database consisting of 4-emotions is used in the proposed work. Assamese is a regional language from North-Eastern India, primarily spoken in the state of Assam. This work on emotion recognition in Assamese not only addresses a significant research gap but also contributes to both academic knowledge and cultural preservation. The proposed work also uses a RAVDESS emotional song database consisting of 6 emotions. Many researchers have performed their work using RAVDESS speech databases (Christy et al., 2020; Farooq et al., 2020), but to our best knowledge, the publications still need to address the RAVDESS song database using a deep-learning approach. This work focuses on selecting appropriate machine learning algorithms; therefore, a CNN-LSTM (Agga et al., 2022) deep learning approach is used for music emotion recognition.

The remaining part of the paper is structured as follows: Literature reviews have been discussed in section 2. The materials and methods including feature extraction and classification has been discussed in section 3. Section 4 presents the results and discussions. Section 5 concludes the paper with some possible future directions.

LITERATURE REVIEW

For the analysis of emotion recognition, an annotated emotional database is necessary. Among the various approaches for labeling emotions in music, mainly the categorical (Panda et al., 2015) and dimensional (Er & Esin, 2021; X. Liu et al., 2017) approaches can be found in research papers related to music emotion recognition. In the categorical approach, emotions are identifying by giving labels to music excerpts such as happy, sad, angry, fearful, relaxed etc. (Patra et al., 2016). Hevner's affective ring model (Er & Esin, 2021; Patra et al., 2018) is one of the popular models in emotion research with 8 groups of emotions. In the dimensional approach, emotion is represented with dimensional space. (Russell, 1980) proposed a 2D general model of affection that consists of valence and arousal as the primary dimensions (Delbouys et al., 2018; Weninger et al., 2014). Panda et al. (2020) adopted Russell's emotion quadrant to evaluate the performance of their work for music emotion recognition by creating a public dataset of 900 audio clips.

Distinguishing relevant features is essential for recognizing music's emotion. Zhang et al. (2016) used a random forest classifier with four emotions: happy, sad, fear, and relaxed. Music audio signals are extracted using mel-frequency cepstrum coefficients (MFCC), RMS energy, zero-crossing rate (ZCR), and fundamental frequency F0. The model tested with an emotional APM music database and the classification accuracy has shown to be 83.29%. Panda et al. (2015) have shown the classification of emotional state in audio music based on both standard (253) and melodic (98) audio features. It is evident from their experiment that melodic features perform better than the standard audio features. Patra et al. (2018) use acoustic features (Deng & Leung, 2012; Kaya et al., 2020), such as rhythm, timbre, and intensity, to classify various moods of Hindi and Western songs. Masood et al. (2016) use spectral features such as root mean square energy, brightness, roughness, spectral roll-off, skewness, and flatness for singer identification in Hindi songs. RMSE (root mean square error), MAE (mean absolute error), and R^2 (square coefficient) are used by Yang (2021) in music emotion recognition using neural network technology.

Nowadays, machine learning and deep learning approaches are used by many researchers to recognize music emotions in a dataset. Deep learning architectures achieve excellent results along with some suitable feature extraction techniques. A deep learning method of one-dimensional residual convolutional neural network (1D CNN) with the inception gate recurrent unit (GRU) has been proposed by Han et al. (2023). They experimented on the Soundtrack dataset and observed that the proposed model performs effectively in emotion detection and classification tasks in music with an accuracy of 84%. Hizlisoy et al. (2021) adopted a new Turkish emotional music database with a duration of 30 seconds each, classified using the long short-term memory (LSTM) and deep neural network (DNN) that has shown 99.19% accuracy. The performance of LSTM+DNN is compared with KNN, SVM, and random forest classifiers. X. Liu et al. (2017) proposed a MER method using a deep convolutional neural network (Aziz, 2020). and performed the experiment on the standard CAL500 and CAL500exp datasets. Additional effort on extracting specific features is optional in their model as it is included in the training procedure of the CNN model. He and Ferguson (2022) adopted VA model to experiment using Bidirectional LSTM model to predict emotion for the music excerpt.

Aljanaki et al. (2017) developed a new music dataset, MediaEval Database for Emotional Analysis in Music (DEAM), one of the most extensive datasets of dynamic annotation consisting of 1,802 excerpts. De Benito-Gorron et al. (2019) used the Google AudioSet dataset for speech and music event detection, classified using convolutional and LSTM recurrent networks that have produced an accuracy of 85%. A database AMG1608 created by Chen et al. (2015) contains 1,608 songs annotated by 665 subjects and represents the emotion in valence and arousal. Soleymani et al. (2013) also collected a large emotional database that consisted of 1000 song samples**.** MoodSwing dataset, developed by Schmidt et al. (2010), contains 240 segments of US pop songs, each lasting 15 seconds. The dataset is annotated with valence-arousal (VA) annotations per second. SVM, KNN, and ANN classifiers are used by Er and Esin (2021) to test their model on a new 4-class Turkish music dataset for music emotion recognition.

MATERIALS AND METHODS

The architecture adopted for the proposed MER has been depicted in Figure 1.

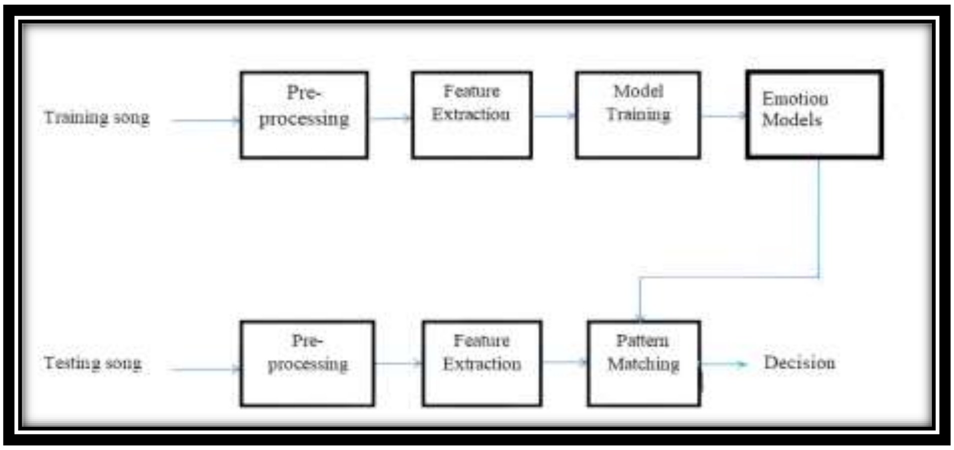


Figure 1. MER block diagram of the proposed model

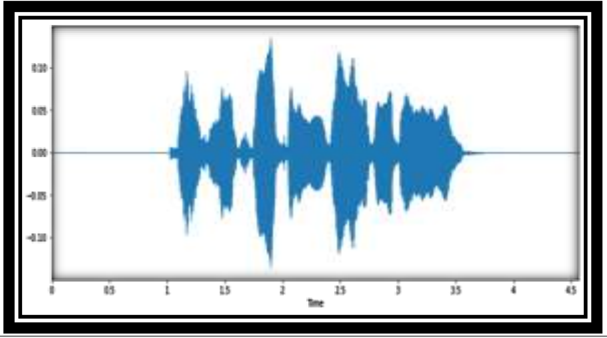
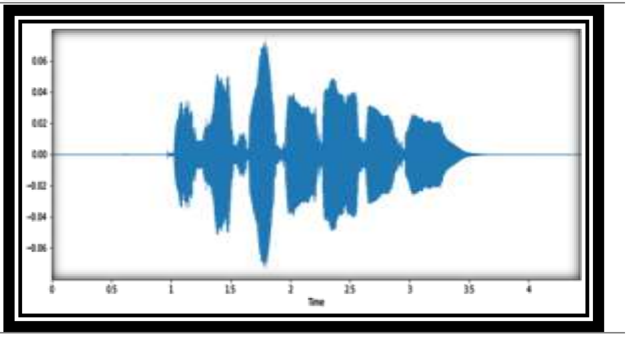
The song databases used in our work are as follows:

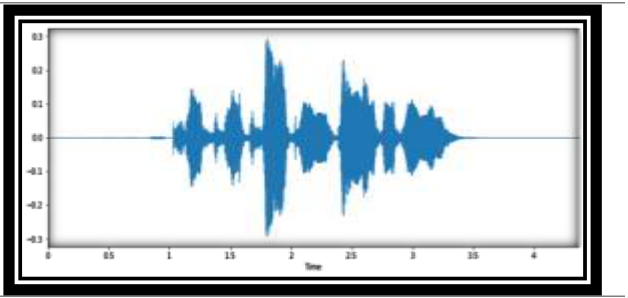
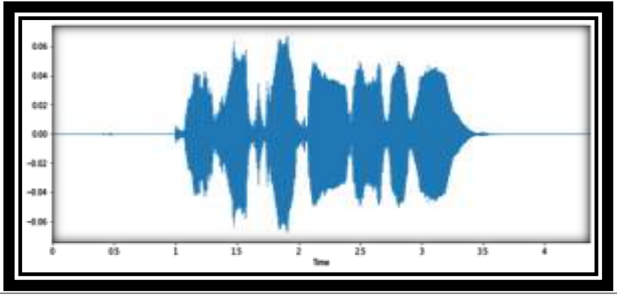
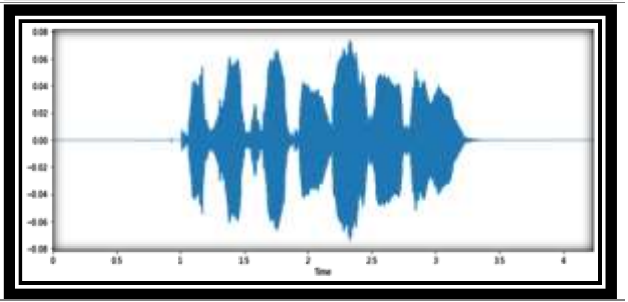
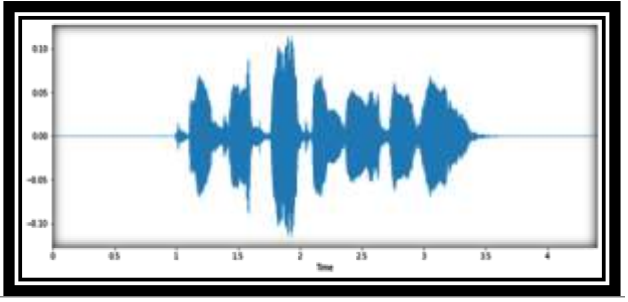
RAVDESS Song Database: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database is used to illustrate the study. The RAVDESS database consists of both emotional speech and song samples. It consists of 24 professional actors (12 male and 12 female) and is recorded in a North American accent. Song samples contain six emotions: happy, sad, angry, calm, fearful, and neutral (no song file of actor 18). Our work uses the RAVDESS song database comprising 1012 song samples (Livingstone & Russo, 2018).

Calm, happy, sad, angry, and fearful emotions were recorded with two levels of expression: strong and normal, except for the neutral emotion. Neutral emotion was recorded in normal expression only. Each sentence sample had two repetitions (Livingstone & Russo, 2018).

Assamese Song Database: In the proposed work, a newly created Assamese song database is used, collected from five well-known singers of Assam. Based on selecting songs with different emotions, pre-processing is performed using Audacity software, where different emotions in the lyrics are considered and separate the singer's voice from the instrumental part. The dataset contains 200 song samples with both male and female singers, and each input audio sample is 5 seconds long. The samples contain four emotions namely calm, happy, neutral, and sad. Table 1 indicates the types of emotions with their properties.

Table 1. Types of emotions and their properties

Emotion	Description/ Properties	Signal
Happy	The pitch value is high.	
Sad	The pitch value is low.	

Angry	High-intensity value.	
Calm	Audio with calm remains uniform	
Neutral	Audio without any emotion is neutral.	
Fearful	Fear is considered as a negative emotion. Not easy to detect.	

Feature Extraction

MFCC: MFCCs are a popular and effective analytical tool used in the music analysis domain. The computation of MFCC mainly involves the following operations: pre-emphasis, framing, windowing, FFT, mel filter bank, and DCT. In the initial steps, the music signal is divided into short frames and then fast Fourier transform is applied to each frame. A Bank of filters is applied to the DFT/FFT spectrum and then converted to log mel spectrum. In the final step, discrete cosine transform (DCT) is used to calculate MFCCs from the log mel spectrum (Christy et al., 2020). Frequency f hertz can be converted to Mel scale by using equation (1).

$$M(f) = 1125 \ln(1 + f/700) \text{ ------------------------- } (1)$$

In this work, a total of forty MFCC coefficients are used. The process of extraction of MFCCs is depicted in Figure 2.

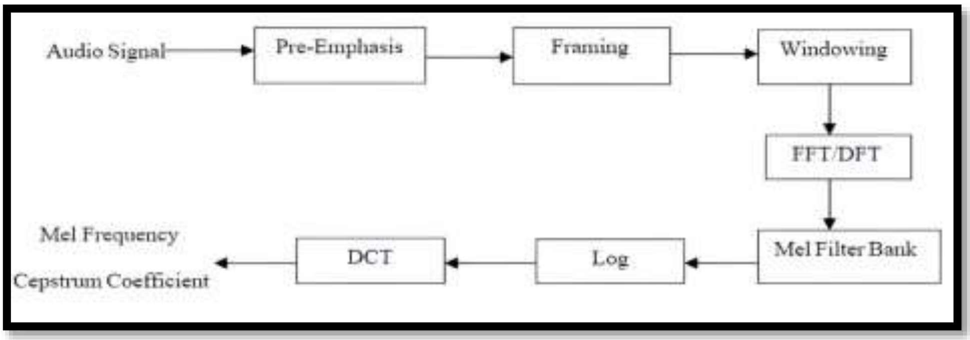


Figure 2. MFCC block diagram

Mel Spectrogram: The Mel spectrogram contains a short-time Fourier transform (STFT), which converts each frame from frequency scale to logarithmic Mel-scale. The audio signals are passed through a bank of filters to get the Mel spectrogram (Zhou et al., 2021). In this work, we considered 128 mel features.

Chroma Feature: Chroma feature or chromatogram is a powerful tool in music audio analysis. Also referred to as pitch class profiles, pitches can often be classified into twelve classes. Each octave incorporated with 12 semitones or chroma, a feature vector of length 12 is obtained (Ellis & Poliner, 2007; Murthy et al., 2018). In this work, twelve chroma features are considered.

Feature Selection

Feature selection is the process of choosing the relevant features from the extracted features according to certain evaluation criterion. This is necessary to improve the recognition accuracy of the learning algorithm.

Classification Methods

Decision Tree: The decision tree is one of the widely used machine learning techniques for supervised classification where predictions are made by considering each feature in the dataset. The entire dataset is visualized as a tree with nodes in the decision tree algorithm. From the node point, the tree splits according to the value of some specific dataset feature (Zuber & Vidhya, 2022). First, we need to select the best attribute or feature from the entire features list of the dataset for the root node and sub-nodes. After the selection of the best feature, the dataset is divided into smaller subsets and the leaf node of a class constitutes the decision (Christy et al., 2020).

K-nearest Neighbors: K-nearest neighbor is a relatively simple classification algorithm. In this approach, all the data points are plotted in space, and to classify a new song, the system will observe its k-nearest points in space and decide by comparing the new song with the other songs in the training dataset (Jamdar et al., 2015). In KNN, K is known as the nearest neighbor. The algorithm is known as the nearest neighbors algorithm when the value of K is equal to 1. In this approach, selecting the optimal value of K is quite challenging. The process is repeated for different values of K to find its optimal value.

Random Forest: Random forest is a supervised machine learning algorithm with many decision trees. It acts as ensemble learning by using which complex problems can be solved by combining many classifiers. The algorithm establishes output based on the outcomes of the decision trees. The results are determined by considering the mean or average of the outcomes from various trees. Since the random forest is a collection of decision trees, increasing the number of trees gives better results and can reduce the issue of overfitting in the decision tree. Thus the algorithm is more accurate than the decision tree algorithm. This algorithm can also handle missing data (Christy et al., 2020; Murthy et al., 2018).

Multi-Layer Perceptron (MLP): MLP is a feedforward ANN consisting of three main layers of nodes: input, hidden, and output. Usually, MLPs are applied to supervised learning problems (Bhatkar & Kharat, 2015). The backpropagation algorithm is used to train the network to build a higher value at the output, representing the appropriate class as output, and all other values are low (Iversen et al., 2006).

Long Short-Term Memory (LSTM): LSTM network is a modified version of recurrent neural networks (RNNs). It can handle the problem of vanishing gradient problem faced by traditional RNNs (H. Liu et al., 2018). The LSTM network architecture consists of three gates; the first gate is called forget gate, the second is called the input gate, the third is the output gate, and the memory unit is known as cell. The memory cell is the core of the LSTM network. LSTM, just like a RNN, also has a hidden state known as short-term memory and the cell state is known as long-term memory. The first gate can keep the useful information coming from the previous time and forget the useless information. In the input gate, the cell tries to quantify the importance of new information. In the output gate, the cell passes the latest updated information to the next timestamp (H. Liu et al., 2018; Qing & Niu, 2018). In Bidirectional LSTM, the output is generated by a forward and backward layer. This work creates an LSTM model with two Bidirectional LSTM layers. The first layer has a parameter return sequences which is set to zero to get all the hidden state. 'Relu' activation function is used with this layer. It is followed by another Bidirectional LSTM layer with 100 neurons.

Convolutional Neural Networks: Convolutional neural networks (CNN) are one of the most popular algorithms in the field of deep learning. CNN can be applied in speech, audio, image, and video processing. CNN combines three layers: convolutional layers, pooling layers, and fully connected layers (Mustaqeem & Kwon, 2019). The first convolutional layer has some filters to apply to input and extracts features from the data matrix. The pooling layer is used to reduce or down-sample the data matrix. Different pooling operations used in CNNs are max pooling, min pooling, mean pooling and average pooling. The last component of CNN architecture is the fully connected (FC) layer. The final pooling layer output is flattened and fed to the fully connected layer. This layer is used for extracting features and then fed to a suitable classifier like 'softmax'.

In our model, we have been using a 1D convolution layer to deal with audio signals. Initially, we chose 64 numbers of filters and 'relu' as the activation function. The pooling layer with a pool size of 8 follows it. We have created a fully connected layer using the 256 'dense' layers. The final output ‘dense’ layer used the ‘softmax’ activation function with six nodes.

Proposed CNN-LSTM Model: CNN-LSTM is a hybrid deep learning (HDL) model that combines CNN and LSTM layers (Aksan et al., 2023; Kim et al., 2013; Tasdelen & Sen, 2021). In this hybrid model, the CNN layer extracts features from the input data, and then CNN output is fed to the LSTM layer as input data to provide sequence prediction (Tasdelen & Sen, 2021). This helps the LSTM to get features from the input data that CNN has recognized. The CNN-LSTM can be applied for activity recognition, image, and video labeling.

The proposed CNN-LSTM model integrated several layers. Initially, we use two convolutional layers with the 'relu' activation function and with 64 and 128 numbers of neurons, respectively. This is followed by the max-pooling layer with a pool size of 4 and the dropout layer. After this, another convolutional layer with the activation function relu was built, followed by the max-pooling and dropout layers. LSTM layer is used with 100 neurons followed by a dropout layer. The final output 'dense' layer uses 6 neurons with a 'softmax' activation function. As the optimizer, 'Adam' with a learning rate of 0.001 and the 'categorical_crossentropy' loss function has been used. The structure of the CNN-LSTM model is shown in Figure 3.

The proposed CNN-LSTM structure of the layer can be described as follows: Con1D layer (filters:64, filter size:10, relu activation) + conv1D layer (filters:128, filter size:10, relu activation) + maxpooling1D (polling size:4) + dropout (0.4) + conv1D layer (filters:128, filter size:10, relu activation) + maxpooling1D (polling size:4) + dropout (0.4) + LSTM layer (neurons: 100) + dropout (0.4) + dense layer (neurons: 6, softmax activation).

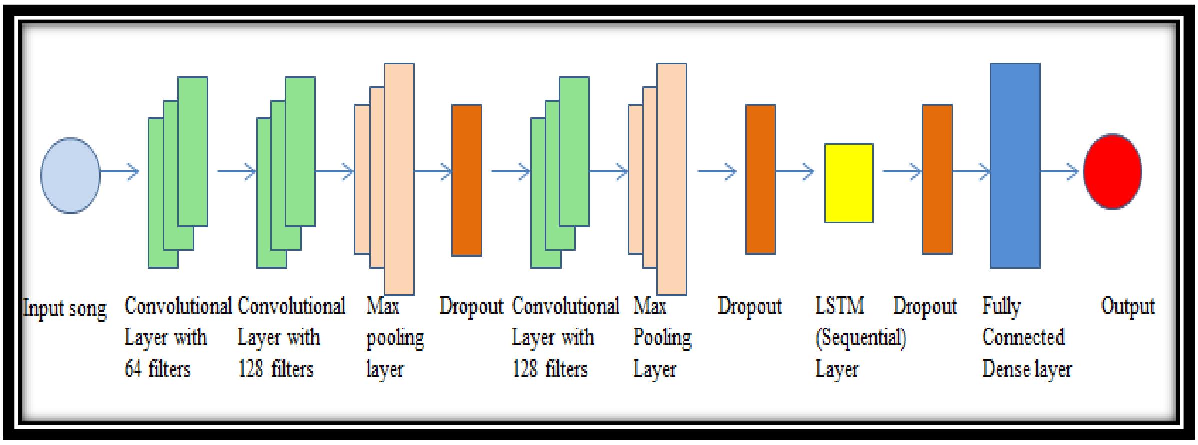

Input song
Convolutional Layer with 64 filters
Convolutional Layer with 128 filters
Max pooling layer
Dropout
Convolutional Layer with 128 filters
Max Pooling Layer
Dropout
LSTM (Sequential) Layer
Dropout
Fully Connected Dense layer
Output

Figure 3. The structure of the proposed CNN-LSTM model

RESULTS AND DISCUSSIONS

Both the Assamese and RAVDESS databases are divided into training and testing samples. Out of these samples, 80% of data are used for training, and 20% of data are used for testing purposes.

The individual performances of each feature category have been constructed, and the accuracy values obtained with different classifiers are shown in Table 2. The overall comparison of accuracies is depicted in Figure 4 for the entire classification model, in which the classification accuracy of the CNN-LSTM model is better compared to the other classification models. The results are presented in terms of accuracy (Eq.(2)), recall (Eq. (3)), precision (Eq.(4)), and F-measure (Eq. (5)) (Hizlisoy et al., 2021). The value of accuracy is calculated using the equation as defined in Eq. (2).

Accuracy= (True Positive+True Negative)/(True Positive+True Negative+False Positive+False Negative) ----------- (2)

Table 2. Emotion recognition accuracy of different classifiers

Sl No.	Classifiers	Accuracy in % for Assamese dataset	Accuracy in % for RAVDESS Song dataset
1	Decision Tree	72.50	60.59
2	KNeighbors	65.00	78.33
3	Random Forest	82.50	81.28
4	LSTM	77.50	81.77
5	MLP	80.00	82.76
6	CNN	82.50	87.68
7	CNN-LSTM	85.00	89.66

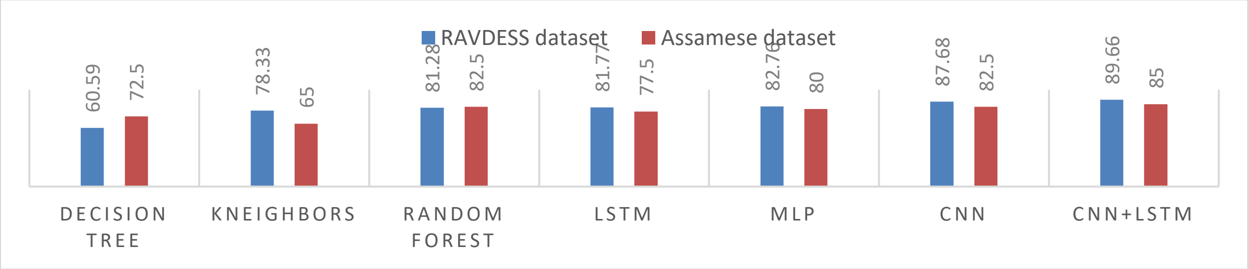


Figure 4. Overall classification accuracy

The confusion matrix of emotion recognition with CNN-LSTM classifiers are depicted in Table 3. The confusion matrix shows how the results of the classification model are distributed over the whole sets of emotion classes. The matrix evaluates the performance of a classification model. For example, in Table 3**,** in the third row, 42 samples belong to the target class happy. Out of which, 38 samples are correctly classified by the model and 4 samples are misclassified as calm. Therefore the classification accuracy for that particular emotion is 90.48%. This percentage is known as recall value.

Recall= $\frac{\text{True Positive}}{\text{True Positive+False Negative}}$ -------------------------- (3)

Similarly, in Table 4, in the fourth column, out of 9 samples, 8 samples correctly belong to the target class sad, and 1 sample is misclassified as calm. Thus the percentage of the correctly classified sample is 88.89%. This value is called the precision value.

Precision= $\frac{\text{True Positive}}{\text{True Positive+False Positive}}$ ------------------------ (4)

F-measure is the harmonic mean of precision and recall value and provides a test model's accuracy. It combines precision and recall into a single value representing both properties. If the value of precision and recall is 1, then this gives a best F-measure score of 1.

F-measure= $\frac{2 \times\text{precision} \times\text{recall}}{precision+recall}$ --------------------------------- (5)

Based on the confusion matrix in Table 3 and Table 4, a set of evaluation measures have been calculated, as shown in Table 5 and Table 6.

Table 3. Confusion matrix for RAVDESS song dataset using CNN-LSTM classifier

Emotions	**Neutral**	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**
Neutral	25	0	0	0	0	0
Calm	0	28	0	0	0	0
Happy	0	4	38	0	0	0
Sad	0	0	1	23	0	4
Angry	0	0	0	0	41	3
Fearful	0	1	0	5	3	27

Table 4. Confusion matrix for Assamese song dataset using CNN-LSTM classifier

Emotions	**Calm**	**Happy**	**Neutral**	**Sad**
Calm	10	4	0	1
Happy	1	6	0	0
Neutral	0	0	10	0
Sad	0	0	0	8

For the RAVDESS song dataset, the CNN-LSTM model trained with 281,286 parameters and with 200 epochs has shown an accuracy of 89.66%. The model accuracy with CNN-LSTM classifiers and the corresponding model loss is shown in Figure 5

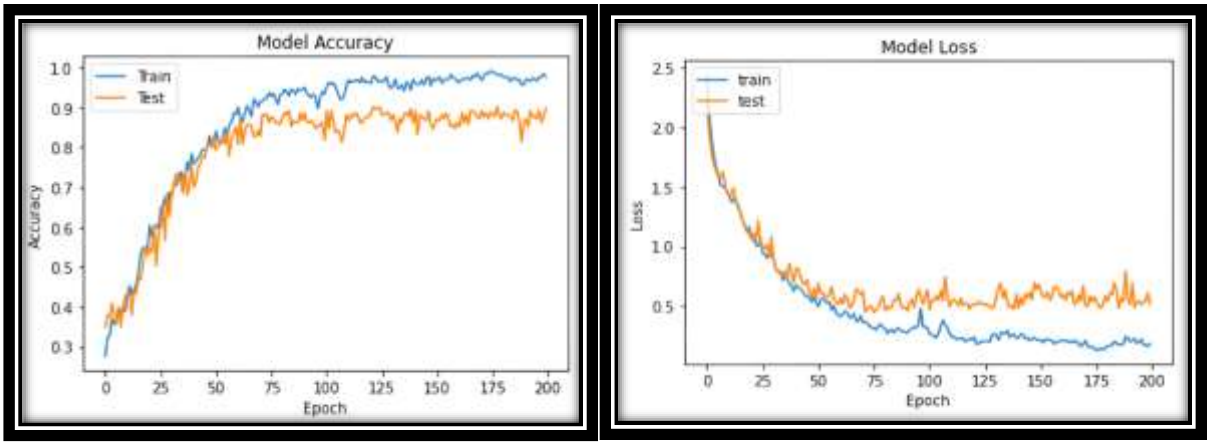


Figure 5. Model accuracy and the corresponding model loss for the RAVDESS dataset using CNN-LSTM classifier

For the Assamese dataset, the CNN-LSTM model trained with 338,724 parameters and with 100 epochs has shown an accuracy of 85.00%. The model accuracy with CNN-LSTM classifiers and the corresponding model loss is shown in Figure 6.

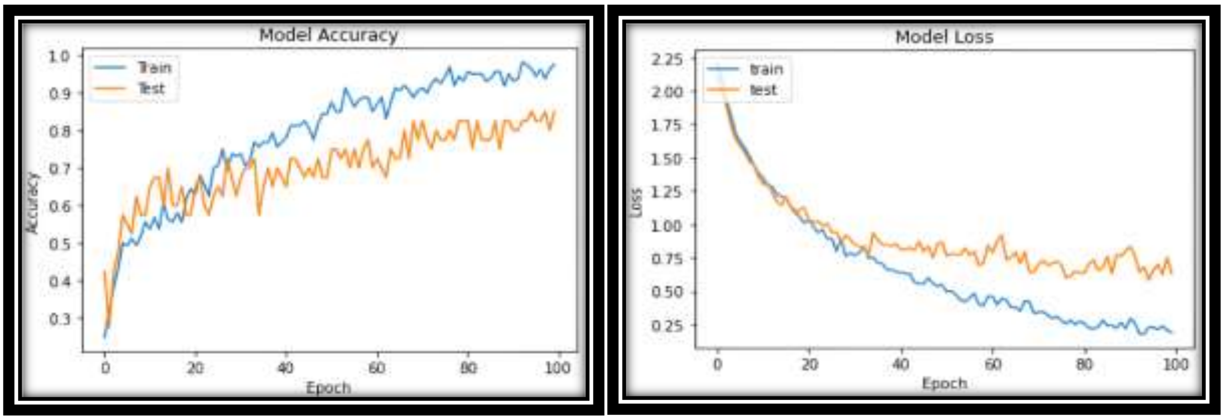


Figure 6. Model accuracy and the corresponding model loss for the Assamese dataset using CNN-LSTM classifier

Table 5. Evaluation measures for RAVDESS song dataset using CNN-LSTM classifier

Evaluation Measure/ Emotion	Precision	Recall	F-Measure
Neutral	1.00	1.00	1.00
Calm	0.85	1.00	0.92
Happy	0.97	0.90	0.94
Sad	0.82	0.82	0.82
Angry	0.93	0.93	0.93
Fearful	0.79	0.75	0.77

Table 6. Evaluation measures for Assamese song dataset using CNN-LSTM classifier

Evaluation Measure/ Emotion	Precision	Recall	F-Measure
Calm	0.91	0.67	0.77
Happy	0.60	0.86	0.71
Neutral	1.00	1.00	1.00
Sad	0.89	1.00	0.94

The emotion detection of song samples using the RAVDESS song dataset is shown in Figure 7, and the Assamese dataset is shown in Figure 8 using CNN-LSTM. The line graph for emotion detection with different classifiers using the RAVDESS song dataset is shown in Figure 9, and the Assamese dataset is shown in Figure 10. A comparative analysis is done with the proposed work and some of the research papers on music emotion recognition, tabulated in Table 7.

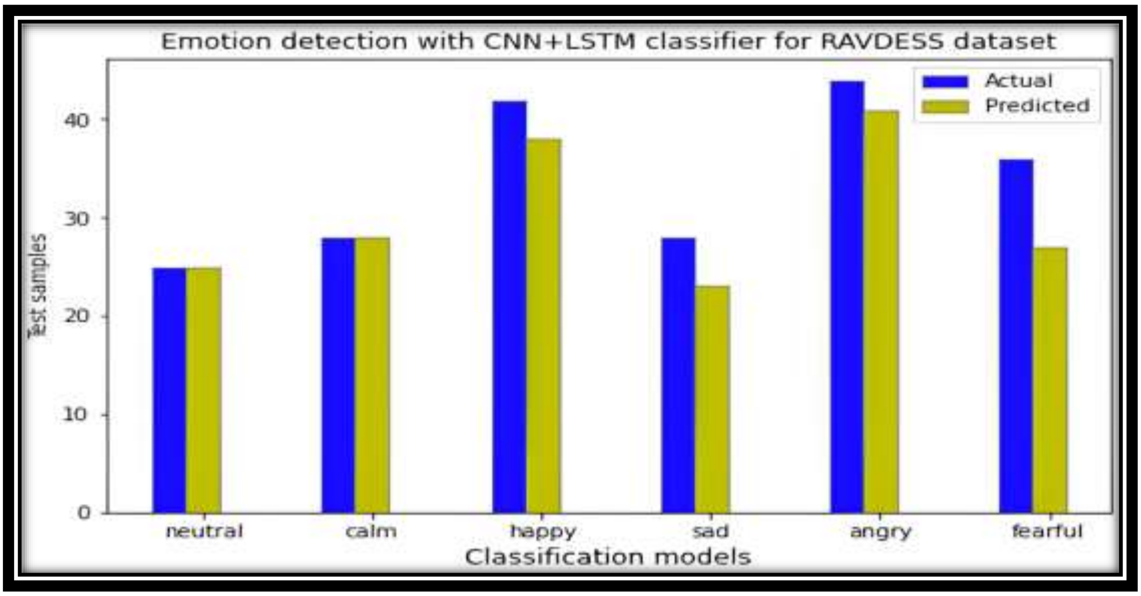


Figure 7. Emotion detection with CNN-LSTM classifier for RAVDESS dataset

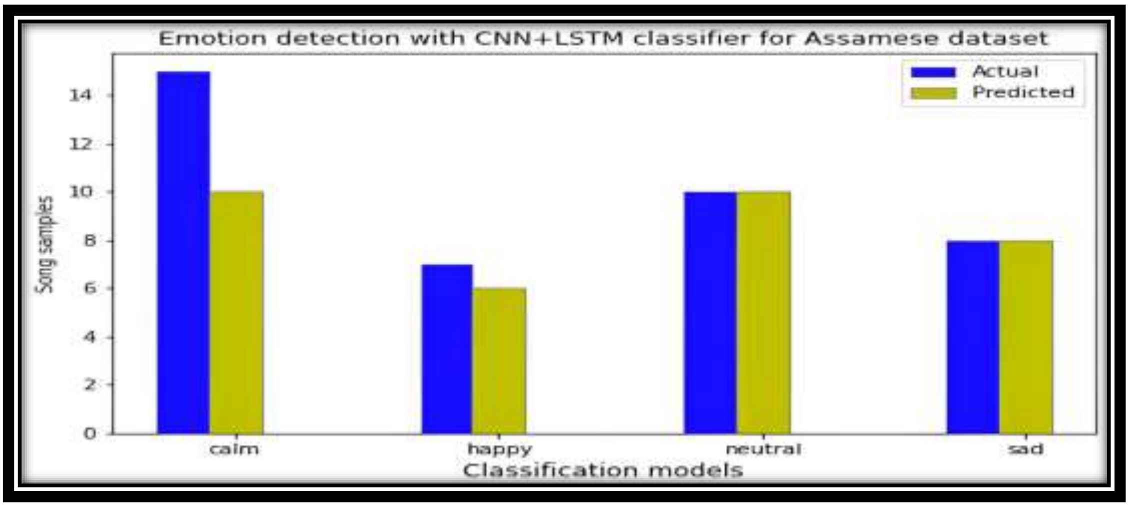


Figure 8. Emotion detection with CNN-LSTM classifier for Assamese dataset

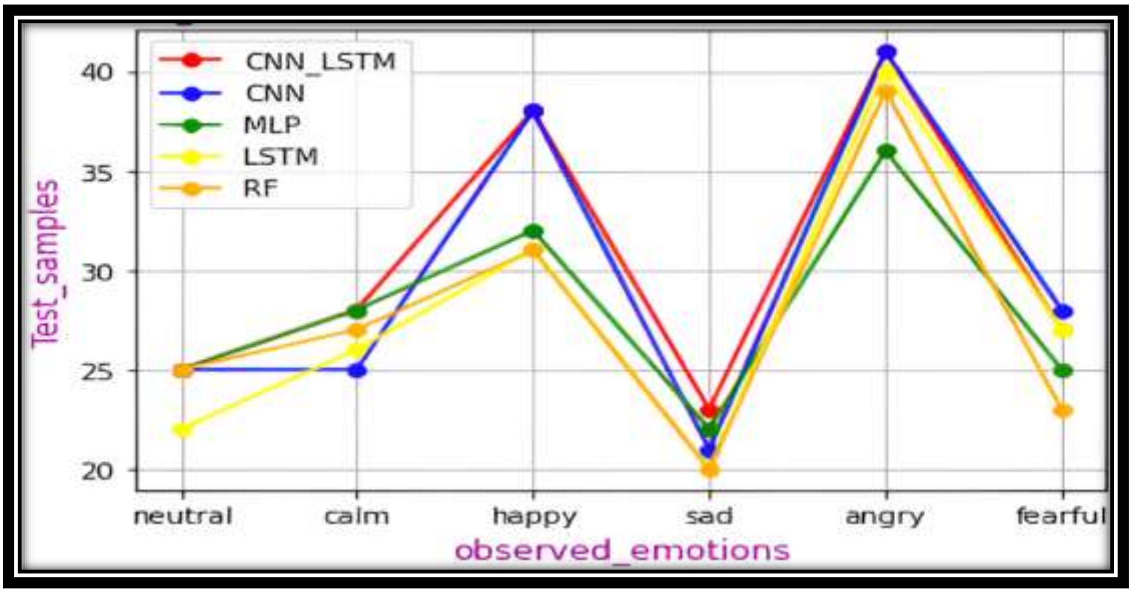


Figure 9. Emotion detection with different classifiers for the RAVDESS dataset

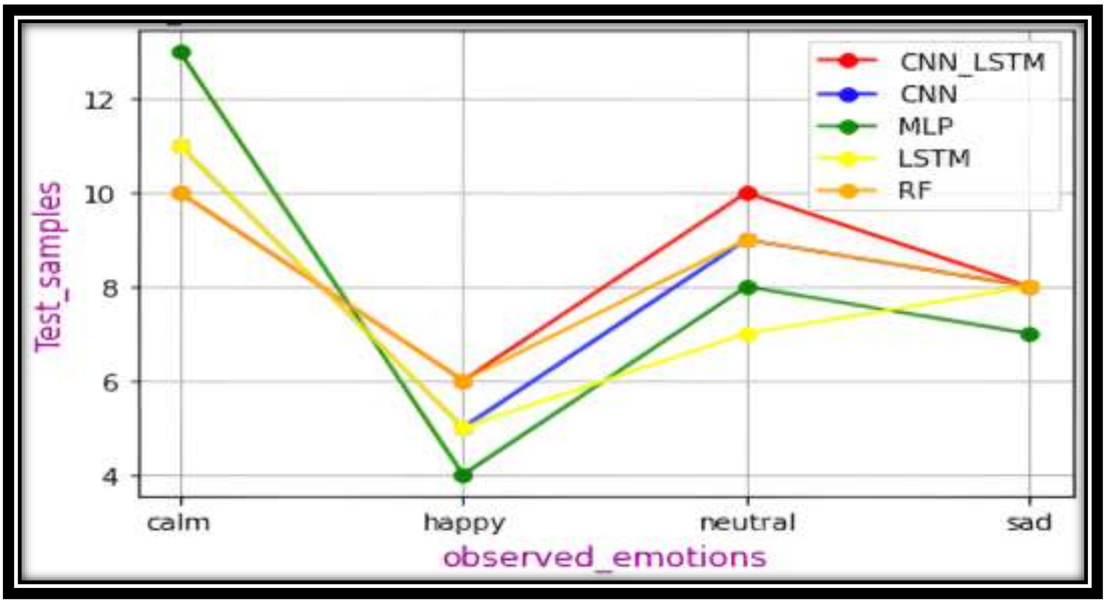


Figure 10. Emotion detection with different classifiers for the Assamese dataset

Table 7. Some of the significant works on music emotion recognition

Reference paper	Database	Emotion model	Number of emotions /approach	Methods (Features and Classifiers)	Best result (Accuracy in %)
Proposed MER model	RAVDEES song database	Categorical	neutral, calm, happy, sad, angry, fearful	MFCC, mel spectrogram, and chroma; decision tree, random forest, multi-layer perceptron (MLP), k-nearest neighbours, LSTM, CNN, CNN-LSTM	89.66
	Assamese song database	Categorical	calm, happy, neutral, sad		85.00
Han et al. (2023)	Soundtrack	Dimensional	valence and arousal	1D-CNN based on the optimized inception –GRU residual structure, SVM	84.27
He and Ferguson (2022)	PMEmo and AllMusic	Dimensional	valence and arousal	Bidirectional-LSTM (BiLSTM)	79.01 (valence) 83.62 (arousal)
Er and Esin (2021)	Turkish music dataset	Categorical	Happy, sad, angry, relax	SVM, KNN, and ANN; RMS energy, tempo, spectrum centroid, spectral entropy, ZCR, MFCC, chroma	79.30
Yang (2021)	MediaEval Emotion in Music (MEM)	Dimensional	valence and arousal	Artificial Bee Colony (ABC) algorithm; RMSE, MAE, and R^2	MAE: Valence=0.8872 Arousal=0.9156
Hizlisoy et al. (2021)	Turkish Emotional Music Database	Dimensional	valence and arousal	KNN, RF, SVM; LSTM+DNN	99.19
De Benito-Gorron et al. (2019)	Google Audioset	-	-	Mel spectrum/mel spectrogram, CNN-LSTM	84.20
Zhang et al. (2016)	APM music database	Categorical	happy, sad, fear, relax	MFCC, RMS energy, zero crossing rate (ZCR), fundamental frequency F0; random forest	83.29
Jamdar et al. (2015)	last.fm website (a popular social musical discovery website)	Categorical	calm, energetic, dance, happy, sad, romantic, seductive, hopeful, angry	Enery, tempo, danceability; k-nearest neighbors	83.40

CONCLUSIONS

This research work proposed a hybrid deep learning CNN-LSTM model to identify music emotion and compare its performance with other classifiers to predict the music emotion in Assamese and RAVDESS song databases. This research first classifies the features of music in a combined form for emotion classification, and a total of 180 features are extracted using MFCC, mel spectrogram, and chroma features. Secondly, combined 1D-CNN and bidirectional LSTM neural network models are constructed to verify the recognition results of the databases. Finally, the evaluation measure is performed in terms of precision, recall, and f-measure, and it is observed that, in terms of performance evaluation, the proposed CNN-LSTM model can be employed to achieve better results in recognizing music emotions for both the Assamese and RAVDESS song databases compared to other models. In this study, introducing a new Assamese song dataset deep learning algorithm and identifying relevant features of MER enhance the theoretical contributions to the MER literature. Regarding practical implications, the MER studies highlight its ability to recognize and respond to a listener's emotional state, offering significant support for both music recommender systems and mental health applications. In the future, the model's accuracy can be further enhanced by choosing relevant features and investigating new classification models that can perform effectively in all evaluation metrics. We will also study the performance of the proposed model with different emotional music databases.

Author Contributions: Conceptualization, J.D. and D.C.; Methodology, J.D.; Software, J.D.; Validation, J.D. and D.C.; Formal Analysis, J.D. and D.C.; Investigation, J.D.; Resources, J.D. and D.C.; Data Curation, J.D., Writing – Original Draft Preparation J.D.; Writing – Review & Editing, J.D. and D.C.; Visualization, J.D.; Supervision, D.C.; Project Administration, J.D.; Funding Acquisition, J.D. and D.C. Author has read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the research does not deal with vulnerable groups or sensitive issues.

Funding: The authors received no direct funding for this research.

Acknowledgments: Not Applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The RAVDESS data described in this research paper are openly available at https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-song-audio. The Assamese data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Agga, A., Abbou, A., Labbadi, M., Houm, Y. E., & Ali, I. H. O. (2022). CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production. *Electric Power Systems Research*, *208*, 107908. https://doi.org/10.1016/j.epsr.2022.107908

Aziz, M. N. (2020). A Review on Artificial Neural Networks and its' Applicability. *Bangladesh Journal of Multidisciplinary Scientific Research*, *2*(1), 48-51. https://doi.org/10.46281/bjmsr.v2i1.609

Aksan, F., Li, Y., Suresh, V., & Janik, P. (2023). CNN-LSTM vs. LSTM-CNN to Predict Power Flow Direction: A Case Study of the High-Voltage Subnet of Northeast Germany. *Sensors*, *23*(2), 901. https://doi.org/10.3390/s23020901

Aljanaki, A., Yang, Y. H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS One*, *12*(3), e0173392. https://doi.org/10.1371/journal.pone.0173392

Bhatkar, A. P., & Kharat, G. U. (2015, December). Detection of diabetic retinopathy in retinal images using MLP classifier. In *2015 IEEE international symposium on nanoelectronic and information systems* (pp. 331-335). IEEE. https://doi.org/10.1109/inis.2015.30

Chen, Y. A., Yang, Y. H., Wang, J. C., & Chen, H. (2015, April). The AMG1608 dataset for music emotion recognition. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 693-697). IEEE. https://doi.org/10.1109/icassp.2015.7178058

Christy, A., Vaithyasubramanian, S., Jesudoss, A., & Praveena, M. D. A. (2020). Multimodal speech emotion recognition and classification using convolutional neural network techniques. *International Journal of Speech Technology*, *23*(2), 381–388. https://doi.org/10.1007/s10772-020-09713-y

De Benito-Gorron, D., Lozano-Diez, A., Toledano, D. T., & Gonzalez-Rodriguez, J. (2019). Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. *EURASIP Journal on Audio, Speech and Music Processing*, *2019*(1), 1-18. https://doi.org/10.1186/s13636-019-0152-1

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M. (2018). Music mood detection based on audio and lyrics with deep neural net. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1809.07276

Deng, J. J., & Leung, C. H. C. (2012). Music emotion retrieval based on acoustic features. In *Lecture notes in electrical engineering* (pp. 169–177). https://doi.org/10.1007/978-3-642-28744-2_22

Domínguez-Jiménez, J., Campo-Landines, K., Martínez-Santos, J., Delahoz, E., & Contreras-Ortiz, S. (2020). A machine learning model for emotion recognition from physiological signals. *Biomedical Signal Processing and Control*, *55*, 101646. https://doi.org/10.1016/j.bspc.2019.101646

Ellis, D. P., & Poliner, G. E. (2007, April). Identifyingcover songs' with chroma features and dynamic programming beat tracking. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 4, pp. IV-1429). IEEE. https://doi.org/10.1109/icassp.2007.367348

Er, M. B., & Esin, E. M. (2021). Music emotion recognition with machine learning based on audio features. *Computer Science*, *6*(3), 133-144. https://doi.org/10.53070/bbd.945894

Farooq, M., Hussain, F., Baloch, N. K., Raja, F. R., Yu, H., & Zikria, Y. B. (2020). Impact of feature selection Algorithm on speech emotion Recognition using deep convolutional neural Network. *Sensors*, *20*(21), 6008. https://doi.org/10.3390/s20216008

Han, X., Chen, F., & Ban, J. (2023). Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure. *Electronics*, *12*(4), 978. https://doi.org/10.3390/electronics12040978

He, N., & Ferguson, S. (2022). Music emotion recognition based on segment-level two-stage learning. *International Journal of Multimedia Information Retrieval*, *11*(3), 383–394. https://doi.org/10.1007/s13735-022-00230-z

Hizlisoy, S., Yildirim, S., & Tufekci, Z. (2021). Music emotion recognition using convolutional long short term memory deep neural networks. *Engineering Science and Technology, an International Journal*, *24*(3), 760–767. https://doi.org/10.1016/j.jestch.2020.10.009

Iversen, A., Taylor, N., & Brown, K. (2006). Classification and verification through the combination of the multi-layer perceptron and auto-association neural networks. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. https://doi.org/10.1109/ijcnn.2005.1556018

Jamdar, A., Abraham, J., Khanna, K., & Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence and Applications*, *6*(3), 35–50. https://doi.org/10.5121/ijaia.2015.6304

Jitendra, M., & Radhika, Y. (2021). An automated music recommendation system based on listener preferences. In *Recent Trends in Intensive Computing* (pp. 80-87). IOS Press. https://doi.org/10.3233/apc210182

Kaya, E. M., Huang, N., & Elhilali, M. (2020). Pitch, timbre and intensity interdependently modulate neural responses to salient sounds. *Neuroscience*, *440*, 1–14. https://doi.org/10.1016/j.neuroscience.2020.05.018

Kim, Y., Lee, H., & Provost, E. M. (2013, May). Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3687-3691). IEEE. https://doi.org/10.1109/icassp.2013.6638346

Liu, H., Mi, X. W., & Li, Y. F. (2018). Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. *Energy Conversion and Management*, *156*, 498–514. https://doi.org/10.1016/j.enconman.2017.11.053

Liu, X., Chen, Q., Wu, X., Liu, Y., & Yang, L. (2017). CNN based music emotion classification. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1704.05665

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS One*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Masood, S., Nayal, J. S., & Jain, R. K. (2016, July). Singer identification in Indian Hindi songs using MFCC and spectral features. In *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)* (pp. 1-5). IEEE. https://doi.org/10.1109/icpeices.2016.7853641

Modran, H. A., Chamunorwa, T., Ursuțiu, D., Samoilă, C., & Hedeșiu, H. (2023). Using deep learning to recognize therapeutic effects of music based on emotions. *Sensors*, *23*(2), 986. https://doi.org/10.3390/s23020986

Murthy, Y. V., Jeshventh, T. K. R., Zoeb, M., Saumyadip, M., & Shashidhar, G. K. (2018, August). Singer identification from smaller snippets of audio clips using acoustic features and DNNs. In *2018 eleventh international conference on contemporary computing (IC3)* (pp. 1-6). IEEE. https://doi.org/10.1109/ic3.2018.8530602

Mustaqeem, N., & Kwon, S. (2019). A CNN-Assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, *20*(1), 183. https://doi.org/10.3390/s20010183

Panda, R., Malheiro, R., & Paiva, R. P. (2020). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, *11*(4), 614–626. https://doi.org/10.1109/taffc.2018.2820691

Panda, R., Rocha, B., & Paiva, R. P. (2015). Music Emotion Recognition with Standard and Melodic Audio Features. *Applied Artificial Intelligence*, *29*(4), 313–334. https://doi.org/10.1080/08839514.2015.1016389

Patra, B. G., Das, D., & Bandyopadhyay, S. (2016). Labeling data and developing supervised framework for Hindi music mood analysis. *Journal of Intelligent Information Systems*, *48*(3), 633–651. https://doi.org/10.1007/s10844-016-0436-1

Patra, B. G., Das, D., & Bandyopadhyay, S. (2018). Multimodal mood classification of Hindi and Western songs. *Journal of Intelligent Information Systems*, *51*(3), 579–596. https://doi.org/10.1007/s10844-018-0497-4

Qing, X., & Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, *148*, 461–468. https://doi.org/10.1016/j.energy.2018.01.177

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010, March). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the international conference on Multimedia information retrieval* (pp. 267-274). https://doi.org/10.1145/1743384.1743431

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C. Y., & Yang, Y. H. (2013, October). 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia* (pp. 1-6). https://doi.org/10.1145/2506364.2506365

Tasdelen, A., & Sen, B. (2021). A hybrid CNN-LSTM model for pre-miRNA classification. *Scientific reports*, *11*(1), 14125. https://doi.org/10.1038/s41598-021-93656-0

Weninger, F., Eyben, F., & Schuller, B. (2014, May). On-line continuous-time music mood regression with deep recurrent neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5412-5416). IEEE. https://doi.org/10.1109/icassp.2014.6854637

Yang, J. (2021). A novel music emotion recognition model using neural network technology. *Frontiers in psychology*, *12*, 760060. https://doi.org/10.3389/fpsyg.2021.760060

Zhang, F., Meng, H., & Li, M. (2016, August). Emotion extraction and recognition from music. In *2016 12th international conference on natural computation, fuzzy systems and knowledge discovery (icnc-fskd)* (pp. 1728-1733). IEEE. https://doi.org/10.1109/fskd.2016.7603438

Zhou, Q., Shan, J., Ding, W., Wang, C., Yuan, S., Sun, F., ... & Fang, B. (2021). Cough recognition based on mel-spectrogram and convolutional neural network. *Frontiers in Robotics and AI*, *8*, 580080. https://doi.org/10.3389/frobt.2021.580080

Zuber, S., & Vidhya, K. (2022, July). Detection and analysis of emotion recognition from speech signals using Decision Tree and comparing with Support Vector Machine. In 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1-5). IEEE. https://doi.org/10.1109/icses55317.2022.9914046

Publisher's Note: CRIBFB stays neutral with regard to jurisdictional claims in published maps and institutional affiliations