

Article

Not peer-reviewed version

Bangla Speech Emotion Recognition Using Deep Learning-Based Ensemble Learning and Feature Fusion

Md. Shahid Ahammed Shakil , [Nitun Kumar Podder](#) , S.M. Hasan Sazzad Iqbal , [Abu Saleh Musa Miah](#) ^{*} , [Md Abdur Rahim](#) ^{*}

Posted Date: 25 March 2025

doi: 10.20944/preprints202503.1864.v1

Keywords: Speech-based emotion recognition (SER); Data Augmentation; Feature Extraction; Ensemble Learning; human-computer interaction (HCI); Deep Learning; LSTM; CNN; Handcrafted feature




Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Bangla Speech Emotion Recognition Using Deep Learning-Based Ensemble Learning and Feature Fusion

Md. Shahid Ahammed Shakil¹, Nitun Kumar Podder¹, S.M. Hasan Sazzad Iqbal¹, Abu Saleh Musa Miah²  and Md Abdur Rahim¹

¹ Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna 6600, Bangladesh

² Department of CSE, Bangladesh Army University of Science and Technology (BAUST), Nilphamari, Bangladesh; abusalehcse.ru@gmail.com

* Correspondence: Abu Saleh Musa Miah (musa@baust.edu.bd), Abdur Rahim (rahim@pust.ac.bd)

† These authors contributed equally to this work.

Abstract: Emotion recognition in speech is essential for enhancing human-computer interaction (HCI) systems. Despite progress in Bangla speech emotion recognition, challenges remain, including low accuracy, speaker dependency, and poor generalization across emotional expressions. Previous approaches often rely on traditional machine learning or basic deep learning models, struggling with robustness and accuracy in noisy or varied data. In this study, we propose a novel multi-stream deep learning feature fusion approach for Bangla speech emotion recognition, addressing the limitations of existing methods. Our approach begins with various data augmentation techniques applied to the training dataset, enhancing the model's robustness and generalization. We then extract a comprehensive set of handcrafted features, including Zero-Crossing Rate (ZCR), chromagram, spectral centroid, spectral roll-off, spectral contrast, spectral flatness, Mel-Frequency Cepstral Coefficients (MFCCs), Root Mean Square (RMS) energy, and Mel-spectrogram. These features capture key characteristics of the speech signal, providing valuable insights into the emotional content. Sequentially, we utilize a multi-stream deep learning architecture to automatically learn complex, hierarchical representations of the speech signal. This architecture consists of three distinct streams: the first stream uses 1D Convolutional Neural Networks (1D CNN), the second integrates 1D CNN with Long Short-Term Memory (LSTM), and the third combines 1D CNN with Bidirectional LSTM (Bi-LSTM). These models capture intricate emotional nuances that handcrafted features alone may not fully represent. For each of these models, we generate predicted scores, and then employ ensemble learning with a soft voting technique to produce the final prediction. This fusion of handcrafted features, deep learning-derived features, and ensemble voting enhances the accuracy and robustness of emotion identification across multiple datasets. Our method demonstrates the effectiveness of combining various learning models to improve emotion recognition in Bangla speech, providing a more comprehensive solution compared to existing methods. We utilize three primary datasets—SUBESCO, BanglaSER, and a merged version of both—as well as two external datasets, RAVDESS and EMODB, to assess the performance of our models. Our method achieves impressive results with accuracies of 92.90%, 85.20%, 90.63%, 67.71%, and 69.25% for the SUBESCO, BanglaSER, merged SUBESCO and BanglaSER, RAVDESS, and EMODB datasets, respectively. These results demonstrate the effectiveness of combining handcrafted features with deep learning-based features through ensemble learning for robust emotion recognition in Bangla speech.

Keywords: speech-based emotion recognition (SER); data augmentation; feature extraction; ensemble learning; human-computer interaction (HCI); deep learning; LSTM; CNN; handcrafted feature

1. Introduction

Speech-based emotion recognition (SER) has applications in natural language processing and human-computer interaction (HCI), among other areas [1]. It can contribute to the enhancement of HCI systems by enabling personalized and human-like interactions. SER is helpful in a number of areas, including marketing, education, mental health, speech synthesis, and customer satisfaction [1]. For instance, SER can enhance the user experience overall by recognizing disgruntled consumers and offering insights into user preferences and behaviour. To evaluate speech emotions, a variety of methods and strategies are used, including machine learning algorithms and statistical and probabilistic models [2]. Deep learning techniques have recently been central in this field [2–4]. Speech emotion identification has shown promise for deep learning approaches, including CNNs [5], DBNs, RNNs, and LSTMs [6]. Again, not much study or development has been done in the Bangla language to identify emotions. Therefore, there is a need and potential to create a speech-emotion recognition system for Bangla. Our study's main objective is to classify speech emotions using an ensemble learning approach that incorporates three types of deep learning algorithms: CNN, LSTM, BiLSTM, or combinations. Identifying emotions in spoken language, particularly in Bangla, is difficult because of varying linguistic usage, social and cultural factors, subjective experiences, and scant evidence [7] [8]. Once more, individual and cultural variations, together with the variety of emotional displays in tones, dialects, and speech rates, present major challenges to algorithms trying to identify emotions in Bangla speech [7,8]. Recently, many researchers employed traditional handcraft feature-based machine-learning approaches to recognise emotions from Bangla speech [6,9]. However, their performance accuracy is not satisfactory. More recently, some researchers employed a deep learning approach to improve the performance accuracy [6] [10]. However, the mentioned research work still faces challenges in achieving good performance accuracy and generalisation properties due to lacking effective features. To overcome the challenges, we proposed integrating the hand-created and multi-stream deep learning features to develop a Bangla speech-based emotion recognition system. Major contributions of the proposed model are given below:

1. **Fusion of Handcrafted and Deep Learning Features:** We combine handcrafted features with deep learning-derived representations to capture both explicit speech characteristics and complex emotional patterns. This fusion enhances the model's accuracy and robustness, improving generalization across different emotional expressions and speaker variations.
2. **Handcrafted Features:** We extract features such as Zero-Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCCs), spectral contrast, and Mel-spectrogram, which focus on key speech characteristics like pitch, tone, and energy fluctuations. These features provide valuable insights into emotional content, enhancing the model's ability to distinguish subtle emotional variations.
3. **Multi-Stream Deep Learning Architecture:** Our model employs three streams: 1D CNN, 1D CNN with Long Short-Term Memory (LSTM), and 1D CNN with Bidirectional LSTM (Bi-LSTM), which capture both local and global patterns in speech, providing a robust understanding of emotional nuances. The LSTM and Bi-LSTM streams improve the model's ability to recognize emotions in speech sequences.
4. **Ensemble Learning with Soft Voting:** We combine predictions from the three streams using an ensemble learning technique with soft voting, improving emotion classification by leveraging the strengths of each model.
5. **Improved Performance and Generalization:** Data augmentation techniques such as noise addition, pitch modification, and time stretching enhance the model's robustness and generalization, addressing challenges like speaker dependency and variability in emotional expressions. Our approach achieves impressive performance, with accuracies of 92.90%, 85.20%, 90.63%, 67.71%, and 69.25% for the SUBESCO, BanglaSER, merged SUBESCO and BanglaSER, RAVDESS, and EMODB datasets, respectively, demonstrating its superiority over traditional models.

2. Related Works

Speech emotion recognition (SER) has been transformed by deep learning algorithms [11–13]; yet, research on SER in the Bangla language is scarce. A deep learning approach to speech emotion recognition was proposed in 2021 by Sadia Sultana, M. Zafar Iqbal, et al [14]. Using the Bangla audio-only dataset SUBESCO, they used bidirectional LSTM networks and deep convolutional neural networks with a time-distributed flatten layer for their investigation. They used cross-lingual and multilingual training testing sets in many tests, and their method with a TDF layer performed better than other cutting-edge CNN-based SER models. In order to conduct cross-lingual experiments in Bangla and English, they employed transfer learning and the SUBESCO and RAVDESS datasets for both cross- and multi-corpus training. The model achieved 86.9% (WA) for the SUBESCO dataset and 82.7% (WA) for the RAVDESS dataset. In 2018, Rahman, Md. Masudur, and associates presented a proposal for an automated voice recognition system for Bengali that included a support vector machine with dynamic time warping [15]. For the static features, they employed MFCCs, and for the dynamic features, MFCC derivatives. SVM with RBF was utilized for classification, and their modified DTW technique was utilized for feature matching. The system achieved 86.08% accuracy on 12 speakers. Using phase-based cepstral features, Chakraborty et al. suggested an Automatic Speech Emotion Recognition model in 2022 [8]. They used pre-processing methods with PBCC to extract phase-based information from speech data. The SUBESCO and BanglaSER datasets, which included a gradient-boosting machine-based classifier, were utilized to evaluate the model. In comparison to earlier methods, the findings demonstrated enhanced performance, with an average accuracy of 96% for both speaker-dependent and speaker-independent emotion detection tests. Using five characteristics taken from sound data and used as 1D CNN inputs, Dias Issa, M. Fatih Demirci, et al. created a unique approach for speech emotion identification in 2020 [16]. During testing on the RAVDESS, EMO-DB, and IEMOCAP databases, the model achieved good classification accuracy in speaker-independent audio. For instance, the model achieved 71.61% accuracy on the RAVDESS dataset with 8 classes, 86.1% accuracy on EMO-DB (535 samples) in 7 classes, 95.71% accuracy on EMO-DB (520 samples) in 7 classes, and 64.3% accuracy on IEMOCAP in 4 classes. Without the requirement for visual aids, the suggested model performed better than the majority of existing models. Using 2D and 1D CNN LSTM networks, Jianfeng Zhao, Xia Mao, et al. (2018) presented a deep learning method for speech emotion identification [17]. From speech and log-mel spectrograms, their algorithms retrieved features associated with both local and global emotions. They achieved recognition accuracy of 95.33% and 95.89% in speaker-dependent and speaker-independent trials on EmoDB, and 52.14% and 89.16% in speaker-independent and speaker-dependent testing on the IEMOCAP database, respectively, outperforming competing methods such as CNN and Deep Belief Network. A 1-D dilated CNN with hierarchical features learner blocks (HFLBs) and a bi-directional gated recurrent unit (BiGRU) for SER was suggested by Mustaqeem and Soonil Kwon in 2021 [18]. Through a layered 1-D dilated network (HFLBs), they employed spectral analysis to uncover unknown patterns from audio samples. After that, the characteristics were input into the BiGRU, which used a softmax layer to build the likelihood of speech emotions and learn temporal cues. On the IEMOCAP, EMO-DB, and RAVDESS datasets, their model yielded relative accuracy values of 72.75%, 91.14%, and 78.01%, respectively. In 2017, Badshah et al. proposed a model (CNN) with three convolutional and three FC layers [19]. The model was trained using the seven emotion classes found in the EMODB corpus. In order to train, they used spectrograms. The accuracy percentage of their study was 56%. In 2018, Etienne et al. developed the CNN-LSTM model for identifying speech emotions [20]. High-level features were extracted using convolutional layers, while long-term associations were gathered using recurrent layers. Their best-performing model, a convolutional (4 layers) and BLSTM combination (1 layer), produced 61.7% unweighted accuracy and 64.5% weighted accuracy for four emotions.

In 2020, Xusheng Ai, Victor S. Sheng, et al. proposed an ensemble learning approach for speech emotion identification using ACRNN [21]. Convolutional recurrent neural networks were utilized in conjunction with bagging and attention models to handle observation overlap problems. In order

to solve further issues, they also employed redagging and augmentation strategies. Their research made use of the Emo-DB and IEMOCAP databases. In 2020, Mustaqeem and Kwon presented a SER model based on spectrogram features that classified emotions using a CNN with deep strides (DSCNN) [22]. The average accuracy for the RAVDESS dataset was 79.5%, whereas the average accuracy for the IEMOCAP dataset was 81.75%. In 2015, Zheng et al. used log-spectrograms as input for their proposed DCNN model for speech emotion recognition [23]. They discovered that their model performed better than conventional models that depended on manually created features after using PCA to minimize dimensionality.

3. Datasets

By using the SUBESCO and BanglaSER datasets, which are collections of Bengali speech samples with annotations of expressed emotions, our work seeks to develop an SER system for the language. The RAVDESS and EMODB datasets, each with seven emotion categories, were also used to assess our models.

3.1. SUBESCO Dataset

SUBESCO, also known as the SUST Bangla Emotional Speech Corpus, is an audio-only dataset including about seven hours of emotional Bangla speech. It is made up of 7,000 words that were recorded by 20 native speakers who are split equally between the sexes. Each speaker recorded 10 lines, which correspond to seven different moods. Each audio clip underwent four rounds of verification by male and female raters after the dataset was examined by fifty college students [24]. It was developed to help in the development of Bangla SER systems and is publicly available.

3.2. BanglaSER Dataset

BanglaSER is a collection of speech-audio data from 34 speakers that reflect five fundamental emotional states, and it is used for SER in the Bangla language. Using computers and cell phones to capture spoken audio, 1467 recordings total of three phrases per recording were made, with an equal number of recordings in each category and an equal distribution of male and female performers [25].

3.3. SUBESCO and BanglaSER Merged Dataset

We conducted an inquiry to develop a multi-corpus classification model for Bangla SER by merging the SUBESCO and BanglaSER datasets. Making a more broadly applicable SER model was the primary goal. BanglaSER offers five different emotion courses: fear, neutral, sad, disgusted, and pleased. SUBESCO offers seven different emotional lessons.

3.4. RAVDESS Dataset

The RAVDESS collection is made up of audio and video clips of actors performing heartfelt songs and speeches. It is used for research and development in speech and audio processing, multimedia content analysis, and emotional analysis. It contains recordings of seven emotions sung by 24 performers. Three different forms of data are included in the dataset: audio-only, audio-video, and video-only [26]. We used audio-only data in our study.

3.5. EMODB Dataset

The Technical University of Berlin developed the free-to-use German emotional database EMODB, which has 535 statements expressing seven different emotions (angry, bored, anxious, fear, happy, disgust, sadness, and neutral) from 10 expert speakers [27].

4. Materials and Methods

We propose a novel multi-stream deep learning feature fusion approach for Bangla speech emotion recognition, addressing the limitations of existing methods, such as low accuracy, speaker dependency, and poor generalization across emotional expressions. Our approach begins with various

data augmentation techniques applied to the training dataset, improving the model’s robustness and generalization across different speakers and emotional expressions. We extract a comprehensive set of handcrafted features, including ZCR, chromagram, spectral centroid, spectral roll-off, spectral contrast, spectral flatness,MFCCs,RMS energy, and Mel-spectrogram. These handcrafted features capture essential characteristics of the speech signal, providing valuable insights into the emotional content, particularly in terms of pitch, tone, and energy variations. Subsequently, we utilize a multi-stream deep learning architecture to automatically learn complex, hierarchical representations of emotional content in speech. The architecture consists of three distinct streams: the first stream employs 1D CNN, the second integrates 1D CNN with LSTM, and the third stream combines 1D CNN withBi-LSTM. These deep learning models capture intricate emotional nuances, offering a deeper understanding of speech emotion than handcrafted features alone. We fuse both handcrafted and deep learning-derived features, combining traditional signal characteristics with learned representations of emotional content. We combine predictions from the three streams using an ensemble learning technique with soft voting, improving emotion classification by leveraging the strengths of each model. We assess our model using three primary datasets—SUBESCO, BanglaSER, and a merged version of both—as well as two external datasets, RAVDESS and EMODB. An overview of the proposed methodology is provided in Figure 1.

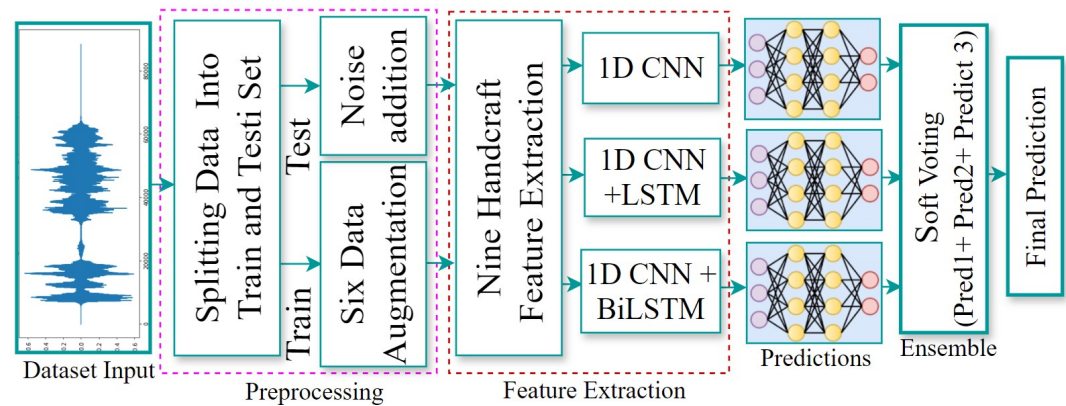


Figure 1. An overview of our proposed methodology.

Table 1. Overview of our train test samples.

Dataset Name	Total Samples	Train/Test Ratio	Train Samples	Test samples	Sam-
SUBESCO	7000	70/30	4900	2100	
BanglaSER	1467	80/20	1173	294	
SUBESCO and Ban- glaSER merged	8467	70/30	5926	2541	
RAVDESS (Audio-only)	1248	70/30	1008	432	
EMODB	535	80/20	428	107	

4.0.1. Data Augmentation

To assess the efficacy of a speech-emotion classification algorithm, the dataset is typically split into training and testing sets. In this study, we splitted the dataset into the training and testing in the ratio of 70/30 and 80/20 based on the dataset owner protocol . An overview of the train-test samples is provided in Table 1. Then we applied data augmentation approach on the training dataset which is described in Table 2. We used data augmentation technique to produce fresh synthetic training samples by making small adjustments to the original training dataset [9,28]. In this case, we applied six different audio data augmentations to improve the speech emotion detection models. It is important

to bear in mind that data augmentation needs to be limited to the training data once it has been split into train and test samples. On our training data, we applied polarity inversion, noise addition, time stretching, pitch change, sound shifting, and random gain [29–31]. To expand the amount of the test samples, we only added noise to our test data. All phases are canceled when the original signal and the phase-inverted signal are combined, producing silence [32]. We simply multiplied the signal by -1 to apply polarity inversion. A volume factor can be used to boost an audio signal’s loudness [33]. One can perceive a gain increase of 10 dB as twice as loud as one would a gain drop of 10 dB. Some examples of data augmentations are given in Figure 2.

Table 2. Data Augmentation Techniques and Their Descriptions

Augmentation Name	Description
Polarity Inversion	Reverses the phase of the audio signal by multiplying it by -1, effectively canceling the phase when combined with the original signal, resulting in silence [32].
Noise Addition	Adds random white noise to the audio data to enhance its variability and robustness [9].
Time Stretching	Alters the speed of the audio by stretching or compressing time series data, increasing or decreasing sound speed [9].
Pitch Change	Changes the pitch of the audio signal by adjusting the frequency of sound components, typically by resampling [34].
Sound Shifting	Randomly shifts the audio by a predefined number of seconds, introducing silence at the shifted location if necessary [9].
Random Gain	Alters the loudness of the audio signal using a volume factor, making it louder or softer [33].

4.1. Feature Extraction

We extracted nine audio features from the frequency, time, and time-frequency domains, including low- and mid-level features, to categorize emotions in speech. These features were extracted for both the original and augmented datasets, stacked horizontally, and arranged vertically for training and testing samples. Feature extraction was performed using the Python Librosa package [29]. The nine features name and other information are shown in the Table 3:

Table 3. Feature Extraction and Their Advantages

Feature Name	Description and Advantage
Zero-Crossing Rate (ZCR)	Counts the number of times the audio signal crosses the horizontal axis. It helps analyze signal smoothness and is effective for distinguishing voiced from unvoiced speech [35] [36].
Chromagram	Represents energy distribution over frequency bands corresponding to pitch classes in music. It captures harmonic and melodic features of the signal, useful for tonal analysis [37] [38].
Spectral Centroid	Indicates the "center of mass" of a sound's frequencies, providing insights into the brightness of the sound. It is useful for identifying timbral characteristics [39].
Spectral Roll-off	Measures the frequency below which a certain percentage of the spectral energy is contained. This feature helps in distinguishing harmonic from non-harmonic content [39].
Spectral Contrast	Measures the difference in energy between peaks and valleys in the spectrum, capturing timbral texture and distinguishing between different sound sources [40] [41].
Spectral Flatness	Quantifies how noise-like a sound is. A high spectral flatness value indicates noise-like sounds, while a low value indicates tonal sounds, useful for identifying the type of sound [42].
Mel-Frequency Cepstral Coefficients (MFCCs)	Captures spectral variations in speech, focusing on features most relevant to human hearing. It is widely used in speech recognition and enhances emotion recognition capabilities [40] [42].
Root Mean Square (RMS) Energy	Measures the loudness of the audio signal, offering insights into the energy of the sound, which is crucial for understanding the emotional intensity [43].
Mel-Spectrogram	Converts the frequencies of a spectrogram to the mel scale, representing the energy distribution in a perceptually relevant way, commonly used in speech and audio processing [44].

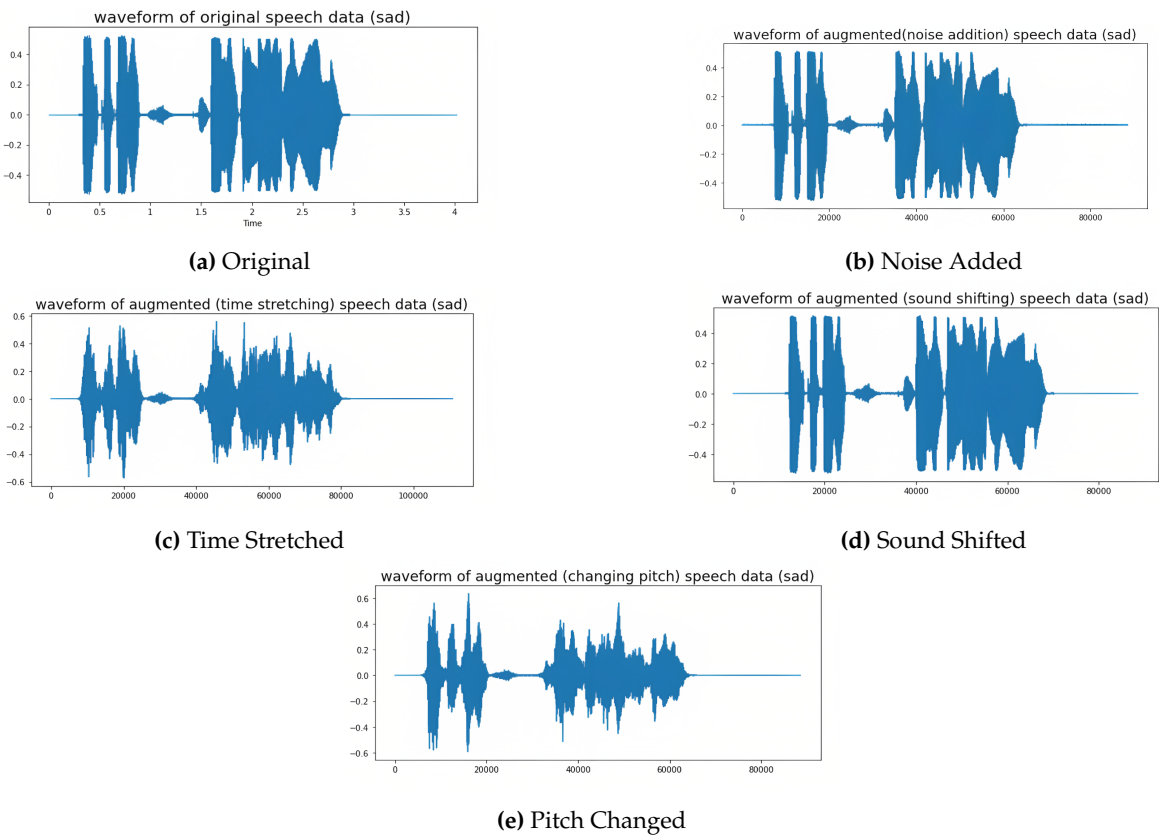


Figure 2. Examples of data augmentations

4.2. Deep Learning Model

After the data are prepared, the model has to be trained for classification. This is the process of giving the algorithm a large number of labeled samples to increase its accuracy. A loss function is used to minimize the discrepancy between the predicted and actual outputs. We were able to train three different models: a 1D CNN, a 1D CNN-LSTM, and a 1D CNN-Bidirectional LSTM.

4.2.1. 1D-CNN Approach

First, we use a 1D-CNN, which has the same architecture across all five datasets, as our emotion classification model. The model includes several layers for one-dimensional convolution, pooling, batch normalization, dropout, activation, flattening, and fully connected layers [45] [46]. The convolution layers extract deep features and produce feature maps [46]. Max-pooling is used by the pooling layers, which down-sample and optimize the spatial size of the feature maps [46] [47]. The model with ReLU incorporates non-linearity through activation layers [46]. While batch normalization layers hasten deep neural network training, dropout layers prevent overfitting [48] [49]. One dimension is extracted from the input by the Flatten layer [50]. The fully connected layer uses a Softmax function to create a probability distribution across the classes to produce predictions based on data from previous layers [46] [51]. With a learning rate of 0.00001, Adam was employed as the optimizer [52]. The architecture of our 1D CNN model is given in Figure 3.

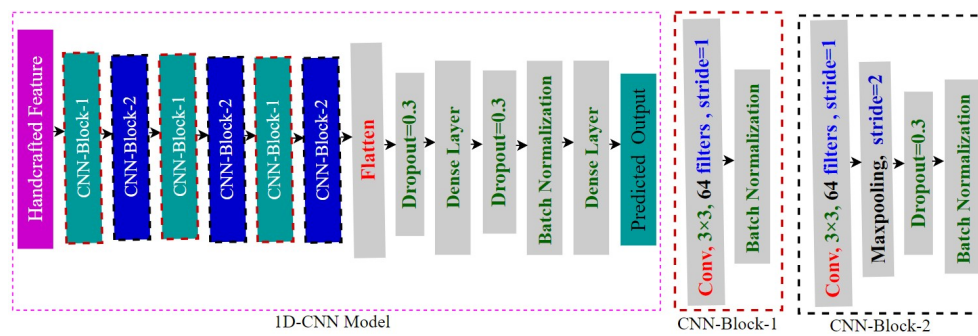


Figure 3. Architecture of our 1D CNN model.

4.2.2. Integration of 1D-CNN and LSTM Approach

For our second classification model, we merged an LSTM with a 1D CNN in an attempt to use both models' advantages. Using a CNN component for feature extraction, the model's architecture is the same for all five types of datasets. After processing the CNN architecture's output, a TimeDistributed Flatten layer sends it to an LSTM layer for sequential analysis and a fully connected layer for predictions. To make sure that the same weights and biases are given to each temporal timestep of the layer, a wrapper called the TimeDistributed Flatten method was used [14]. A Softmax function is used for probability distribution over emotion categories [51]. The Adam optimizer was used with a 0.000001 learning rate [52]. An architecture of our 1D CNN-LSTM model is given in Figure 4.

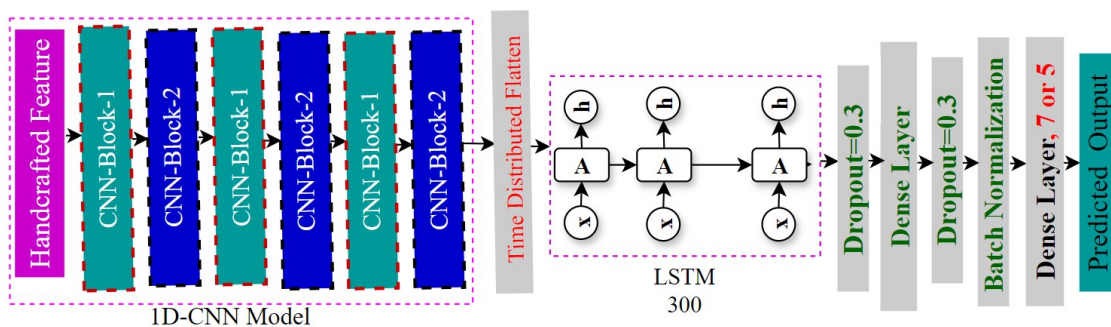


Figure 4. Architecture of our 1D CNN LSTM Model.

4.2.3. Integration of 1D-CNN and Bi-LSTM Approach

Our third classification model extracts contextual and deep characteristics from the input data by combining two bidirectional LSTM layers with a 1D CNN. Similar to our second model, the CNN output is sent to the LSTM layers via a TimeDistributed Flatten layer; however, this time, the layers are bidirectional, allowing for the acquisition of information from both past and future data points [14]. The input is categorized into one of seven or five emotion categories by the model using a Softmax layer [51]. Utilizing a 0.000001 learning rate, the Adam optimizer was used [52]. The architecture of our 1D CNN-BiLSTM model is given in Figure 5.

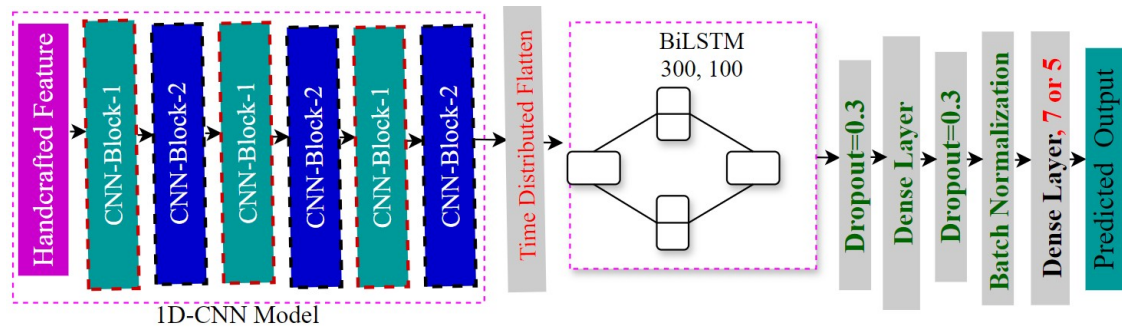


Figure 5. Architecture of our 1D CNN BiLSTM Model.

4.3. Soft Voting Ensemble Learning

We trained three classification models and then applied an ensemble learning technique. Using the sum rule ensemble technique, the three classification models (1D CNN, 1D CNN LSTM, and 1D CNN BiLSTM) trained in the prior stage were combined for better emotion categorization. Sum rule ensemble learning (a form of soft voting) aggregates predictions from various models for each class label and predicts the class label based on the highest summed probability, as shown in Figure 1[53]. In this approach, the overall accuracy is improved by focusing the models on different data points and combining their predictions. The procedure involves generating predictions from each model and aggregating them to form a final prediction. The sum rule is given by the Equation 1:

$$\hat{y}_{\text{final}} = \arg \max \left(\sum_{i=1}^n P(y|x, \theta_i) \right) \quad (1)$$

where: \hat{y}_{final} is the predicted class label. $P(y|x, \theta_i)$ is the probability prediction for class y from model i for input x , with parameters θ_i . n is the total number of models in the ensemble. The prediction is based on the highest summed probability across all models. After generating these summed predictions, we apply the $\arg \max$ function to convert continuous values into class labels. Finally, we compare the predicted class labels with the ground truth labels to assess the ensemble's accuracy. An accuracy metric from `scikit-learn`, such as `accuracy_score`, is used for this evaluation.

5. Results

Three classification models and an ensemble learning strategy were tested on each of the five data sets that we used. Across the datasets, we saw notable variations in the model's performance, and our ensemble technique continuously increased accuracy.

5.1. Ablation Study

Table 4 displays the ablation study where we show the performance for each of our suggested models. Our investigation revealed that the 1D CNN-LSTM model outperformed other base models on the SUBESCO and BanglaSER datasets. Conversely, the 1D CNN model outperformed the others for the SUBESCO and BanglaSER merged, RAVDESS, and EMOB datasets.

Table 4. Summary of performances (accuracy %) of our proposed models.

Dataset	1D CNN	1D CNN LSTM	1D CNN BiLSTM	Ensemble Learning
SUBESCO	90.93%	90.98%	90.50%	92.90%
BanglaSER	83.67%	84.52%	81.97%	85.20%
SUBESCO + BanglaSER	88.92%	88.61%	87.56%	90.63%
RAVDESS	65.63%	64.93%	60.76%	67.71%
EMODB	69.57%	67.39%	65.84%	69.25%

5.2. Outcomes of the Models for SUBESCO Dataset

For the SUBESCO Bangla Speech Emotion Corpus, our proposed models performed well. Over 99% accuracy was achieved by each model in learning the training set of data. The three 1D CNN classification models, 1D CNN LSTM, 1D CNN BiLSTM, and 1D CNN, have testing accuracy scores of 90.93%, 90.98%, and 90.50%, respectively. Ultimately, the accuracy of the ensemble learning approach that combined the three models was 92.90%. This indicates a significant increase in accuracy. A summary of all these findings is given in Table 5. In Table 6, the accuracy of the Ensemble Learning approach for each emotion class of the SUBESCO dataset is given, and the performance metrics are given in Table 7.

Table 5. Model accuracy on the SUBESCO dataset

Model	Accuracy
1D CNN	90.93%
1D CNN LSTM	90.98%
1D CNN BiLSTM	90.50%
Ensemble Learning	92.90%

In Figure 6, the learning curves (accuracy and loss) of our trained models for the SUBESCO dataset are given.

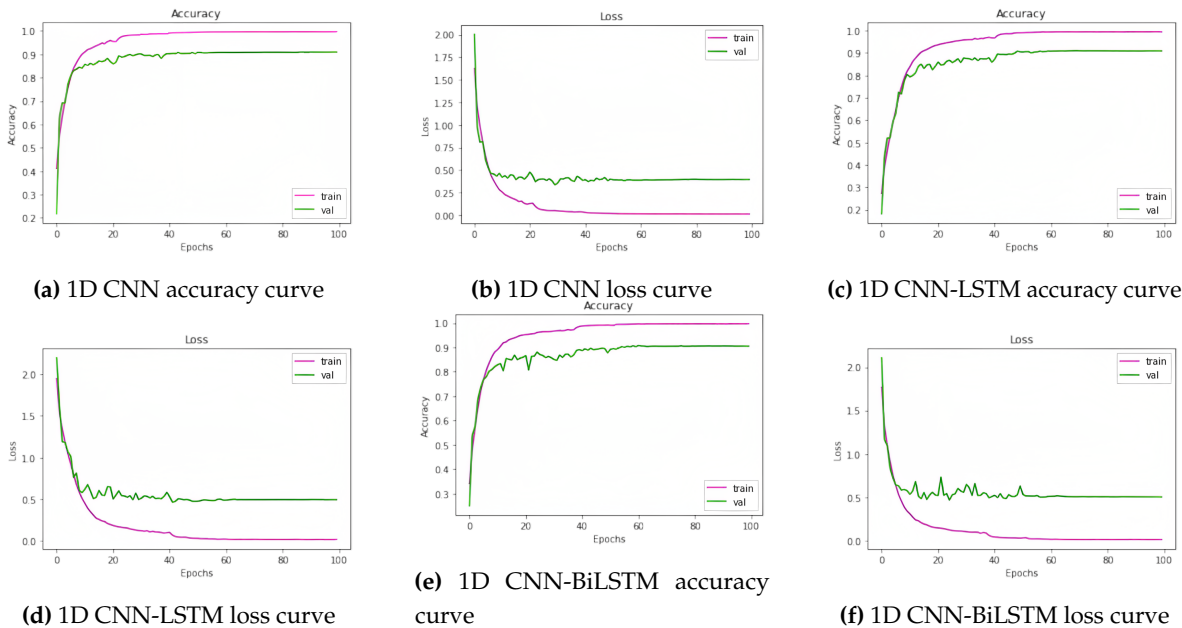


Figure 6. Learning curves (accuracy and loss) of the models (1d cnn, 1d cnn-lstm, and 1d cnn-bilstm, respectively) for the SUBESCO dataset.

Table 6. Accuracy of the Ensemble Learning approach for each emotion class of SUBESCO dataset

Emotion	Accuracy (%)
Angry	93.93
Disgust	84.98
Fear	93.61
Happy	94.98
Neutral	98.69
Sad	92.32
Surprise	92.05

Table 7. Performance metrics (Precision, Recall, and F1-Score) of the Ensemble Learning approach for each emotion class of SUBESCO dataset

Class	Precision	Recall	F1-Score
Angry	0.94	0.94	0.94
Disgust	0.92	0.85	0.88
Fear	0.93	0.93	0.93
Happy	0.91	0.95	0.93
Neutral	0.94	0.99	0.96
Sad	0.95	0.92	0.94
Surprise	0.91	0.92	0.92
Macro Average	0.93	0.93	0.93
Weighted Average	0.93	0.93	0.93
Accuracy = 0.93			

5.3. Outcomes of the Models for BanglaSER Dataset

On the BanglaSER dataset, the accuracy of the three classification models 1D CNN, 1D CNN LSTM, and 1D CNN BiLSTM was 83.67%, 84.52%, and 81.97%, in that order. Ultimately, the accuracy of the ensemble learning method that integrates the three models was 85.20%. This indicates an improvement in accuracy. A summary of all these findings is given in Table 8. In Table 9, the accuracy of the Ensemble Learning approach for each emotion class of the BanglaSER dataset is given, and the performance metrics are given in Table 10. The learning curves (accuracy and loss) of the models (1d cnn, 1d cnn-lstm, and 1d cnn-bilstm, respectively) for the BanglaSER dataset are given in Figure 7.

Table 8. Model accuracy on the BanglaSER dataset

Model	Accuracy
1D CNN	83.67%
1D CNN LSTM	84.52%
1D CNN BiLSTM	81.97%
Ensemble Learning	85.20%

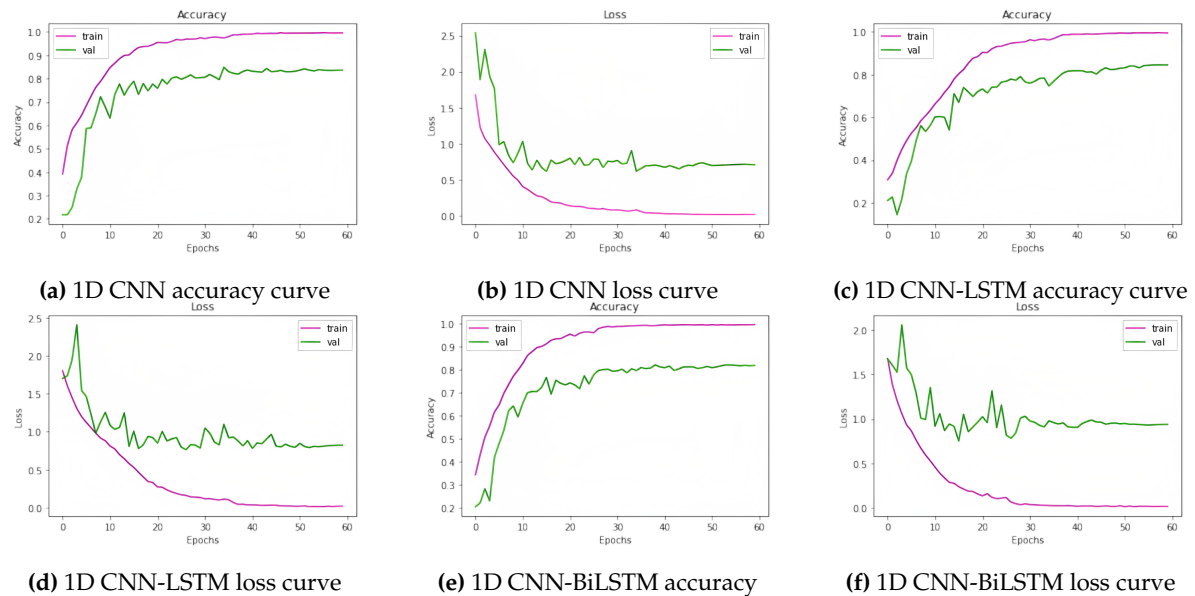


Figure 7. Learning curves (accuracy and loss) of the models (1d cnn, 1d cnn-lstm, and 1d cnn-bilstm, respectively) for the BanglaSER dataset.

Table 9. Accuracy of the Ensemble Learning approach for each emotion class of BanglaSER dataset

Emotion	Accuracy (%)
Angry	93.07
Happy	76.61
Neutral	93.02
Sad	83.59
Surprise	81.67

Table 10. Performance metrics (Precision, Recall, and F1-Score) of the Ensemble Learning approach for each emotion class of BanglaSER dataset

Class	Precision	Recall	F1-Score
Angry	0.91	0.93	0.92
Happy	0.83	0.77	0.80
Neutral	0.89	0.93	0.91
Sad	0.85	0.84	0.84
Surprise	0.78	0.82	0.80
Macro Average	0.85	0.85	0.85
Weighted Average	0.85	0.85	0.85
Accuracy = 0.85			

5.4. Outcomes of the Models for SUBESCO and BanglaSER Merged Dataset

In order to evaluate the three classification models (1D CNN, 1D CNN LSTM, and 1D CNN BiLSTM), we combined two Bangla speech emotion datasets. On the combined dataset, the models' accuracy was 88.92%, 88.61%, and 87.56%, in that order. Nonetheless, we obtained an accuracy of 90.63% with an ensemble learning approach that included all three models, suggesting an improvement in accuracy. A summary of model accuracies for the SUBESCO and BanglaSER Merged Dataset is given in Table 11. In Table 12, the accuracy of the Ensemble Learning approach for each emotion class of the SUBESCO and BanglaSER merged dataset is given, and the performance metrics are given in

Table 13. The learning curves (accuracy and loss) of the models for the SUBESCO and BanglaSER merged dataset are given in Figure 8.

Table 11. Model accuracy on the SUBESCO and BanglaSER merged dataset

Model	Accuracy
1D CNN	88.92%
1D CNN LSTM	88.61%
1D CNN BiLSTM	87.56%
Ensemble Learning	90.63%

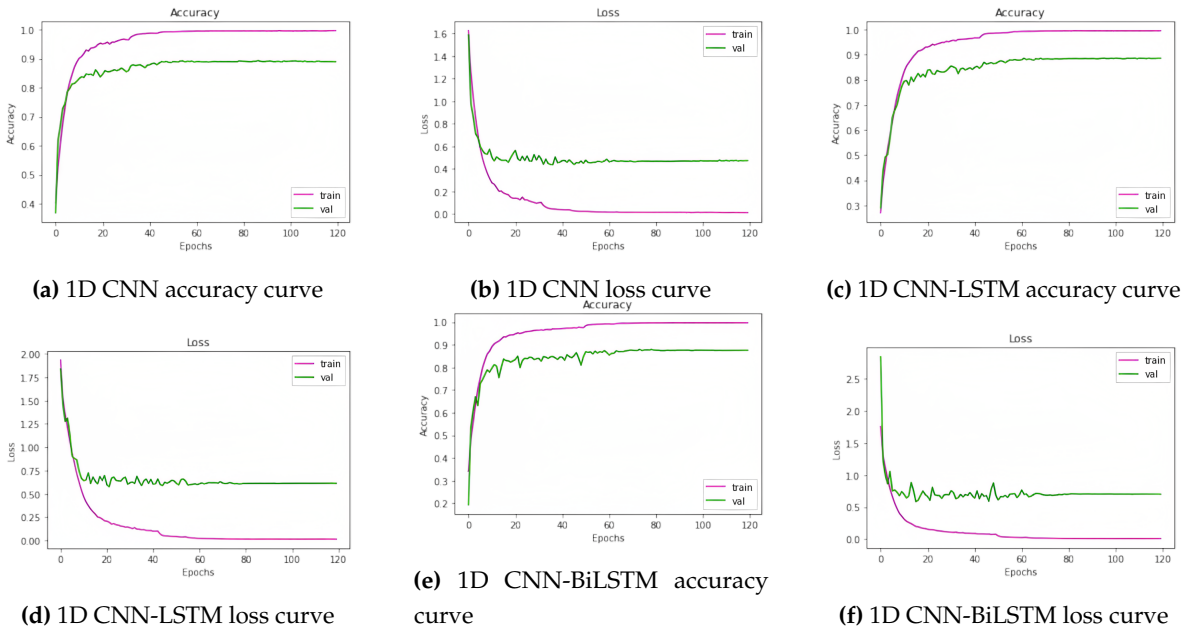


Figure 8. Learning curves (accuracy and loss) of the models (1d cnn, 1d cnn-lstm, and 1d cnn-bilstm, respectively) for the SUBESCO and BanglaSER merged dataset.

Table 12. Accuracy of the Ensemble Learning approach for each emotion class of SUBESCO and BanglaSER merged dataset.

Emotion	Accuracy (%)
Angry	94.84
Disgust	86.19
Fear	92.69
Happy	89.03
Neutral	96.74
Sad	87.53
Surprise	87.15

Table 13. Performance metrics (Precision, Recall, and F1-Score) of the Ensemble Learning approach for each emotion class of SUBESCO and BanglaSER merged dataset.

Class	Precision	Recall	F1-Score
Angry	0.93	0.95	0.94
Disgust	0.91	0.86	0.88
Fear	0.92	0.93	0.92
Happy	0.89	0.89	0.89
Neutral	0.90	0.97	0.93
Sad	0.89	0.88	0.88
Surprise	0.90	0.87	0.89
Macro Average	0.91	0.91	0.91
Weighted Average	0.91	0.91	0.91
Accuracy = 0.91			

For all of our datasets except EMOB, the ensemble learning approach showed improved classification accuracy. Figure 9 shows the confusion matrices of the Ensemble Learning approach for the SUBESCO, BanglaSER, and SUBESCO-BanglaSER merged datasets, where each matrix provides a detailed view of the classification outcomes of the ensemble learning approach for each of the datasets. By comparing these matrices, we can evaluate the performance of our ensemble learning approach, which provided improved identification of emotion classes and handled class imbalance and misclassification across all the datasets.

5.5. Outcomes of the Models for RAVDESS and EMOB Datasets

The models on the RAVDESS and EMOB datasets were trained using the identical model setups as the previous datasets. The models' accuracy on these datasets was mediocre; however, our ensemble learning model's accuracy increased. The accuracy of our models for the RAVDESS and EMOB datasets is given in Tables 14 and 15, respectively.

Table 14. Model accuracy on the RAVDESS dataset

Model	Accuracy
1D CNN	65.63%
1D CNN LSTM	64.93%
1D CNN BiLSTM	60.76%
Ensemble Learning	67.71%

Table 15. Model accuracy on the EMOB dataset

Model	Accuracy
1D CNN	69.57%
1D CNN LSTM	67.39%
1D CNN BiLSTM	65.84%
Ensemble Learning	69.25%

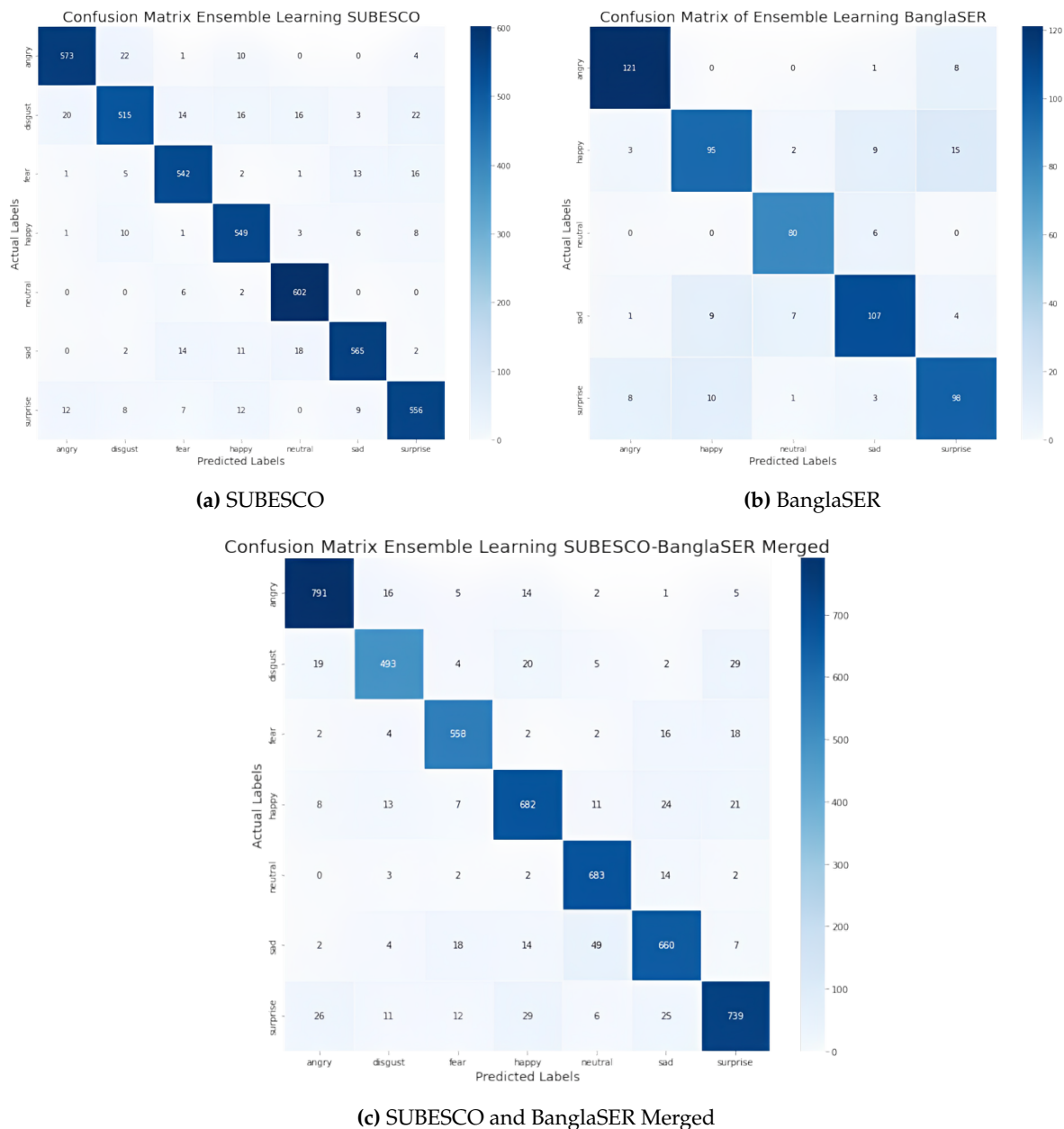


Figure 9. Confusion matrices of the Ensemble Learning approach for different datasets.

5.6. State of art Comparison

The Table 16 presents a comparison of various state-of-the-art models in emotion recognition using speech datasets. Each study employs distinct features and models, with a variety of accuracy metrics for different datasets. For instance, Sultana et al. [14] achieved an accuracy of 86.9% on the SUBESCO dataset using Mel-Spectrogram and Deep CNN combined with BLSTM, while Rahman et al. [15] used MFCCs and their derivatives with SVM and DTW, achieving 86.08% accuracy on Bangla 12 speakers. Chakraborty et al. [8] utilized PBCC features with a Gradient Boosting Machine, reaching an impressive 96% accuracy on their dataset. Notably, the proposed model demonstrates superior performance, with 92.90% accuracy on the SUBESCO dataset, and notable results on RAVDESS, EMO-DB, BanglaSER, and a combined dataset (SUBESCO + BanglaSER) with 90.63% accuracy. This highlights the robustness and effectiveness of the proposed ensemble model.

Table 16. State of the art comparison of the proposed model

Research	Features used	Model	Accuracy for 5 Datasets (%)				
			SUBESCO (Bangla),	RAVDESS (American English)	EMO-DB	BanglaSER (SUBESCO + BanglaSER)	
Sultana et al. [14]	Mel-Spectrogram	Deep CNN and BLSTM	86.9	82.7			
Rahman et al. (2018) [15]	MFCCs, MFCC derivatives	SVM with RBF, DTW	86.08 (Bangla 12 speakers)	-			
Chakraborty et al. (2022) [8]	PBCC	P Gradient Boosting Machine		96			96
Issa, et al. [16].	MFCCs, Mel-spectrogram, Chroma-gram, Spectral contrast feature, Tonnetz representation	1D CNN	64.3% (IEMO-CAP, 4 classes)	71.61 (8 classes)	95.71		
Zhao, et al. 2019 [17].	Log mel spectrogram	1D CNN LSTM, 2D CNN LSTM	89.16 (IEMO-CAP dependent)	52.14% (IEMO-CAP independent)		95.33 (Emo-Db dependent)	95.89% (independent)
Mustaqeem et al. [18]	Spectral analysis	1D Dilated CNN with BiGRU	72.75	78.01	91.14	-	-
Badshah et al. (2017) [19]	Spectrograms	CNN (3 convolutional, 3 FC layers)			56%		
Etienne, et al. (2018) [20]	High-level features, log-mel Spectrogram	CNN-LSTM (4 conv + 1 BLSTM layer)	61.7% (Un-weighted), 64.5% (Weighted)				
Proposed	ZCR, chroma-gram, RMS, spectral centroid, spectral roll-off, spectral contrast, spectral flatness, mel spectrogram, and MFCCs	Ensemble of 1D CNN, 1D CNN LSTM, and 1D CNN BiLSTM	92.90 % (SUBESCO)	67.71% (RAVDESS)	69.25% (EMODB)	85.20% (BanglaSER),	90.63% (SUBESCO + BanglaSER)

5.7. Discussion

Table 4 shows the ablation study, and 16 shows the state of the art comparison model. For every dataset except EMO-DB, the accuracy of the ensemble learning strategy was better than that of other base models. Establishing a method for improved speech emotion categorization from Bangla speech was the main goal of our research. Our method demonstrates how the classification accuracy of

Bangla speech may be increased by combining the right features with data augmentations, followed by an ensemble of trained models. In contrast to Sadia Sultana, M. Zafar Iqbal, et al. (2021), who used CNN and the BiLSTM model to get 86.9% accuracy on the SUBESCO dataset [14], our method achieved 92.90% accuracy on the SUBESCO dataset, 85.20% accuracy on the BanglaSER dataset, and 90.63% accuracy on the combined dataset of SUBESCO and BanglaSER. Regarding the classification of emotions from speech, the ensemble learning approach performs better than most of the models. One main limitation of our research is that our experiment has been conducted on an ACT dataset only. Our experiment was limited to the acted dataset, which is one of the key limitations of our study. In real-world situations, our model might not function well because of variances in speech data caused by environments, languages, and cultural differences. It's also important to take into account the various Bangla dialects. In a subsequent study, we may expand to incorporate natural speech data rather than just-acted data to more accurately detect speech patterns and emotional expressions in real-world situations. Moreover, speech data from multiple regional dialects of the language or multilingual datasets may be integrated to create a more robust system that can accurately recognize emotions in a range of languages and cultural contexts. We may look into several sets of data augmentation and feature extraction strategies to find out which combinations perform best together for voice emotion recognition. To improve the ensemble learning method, we may eventually try incorporating more categorization models and experimenting with other combinations. Our ensemble learning strategy demonstrated constant superiority over the other models on all datasets. Additionally, our method's efficacy with Bengali datasets is demonstrated by the results obtained. With good accuracy levels on most datasets, our method demonstrated the overall effectiveness of ensemble learning for speech-emotion recognition.

6. Conclusion

In this study, we propose a novel multi-stream deep learning feature fusion approach for Bangla speech emotion recognition, which effectively addresses the challenges of low accuracy, speaker dependency, and poor generalization. By combining handcrafted features with deep learning-derived features and employing an ensemble learning technique, our method significantly enhances the robustness and accuracy of emotion recognition across diverse datasets. We combine predictions from the three streams using an ensemble learning technique with soft voting, improving emotion classification by leveraging the strengths of each model. The results demonstrate the ability to capture intricate emotional nuances in Bangla speech, achieving high performance in both primary and external datasets. This approach holds broader potential for expanding emotion recognition systems to other languages and emotional categories, paving the way for more universal and accurate speech-emotion recognition systems. Looking ahead, we plan to extend this research by applying the proposed approach to other underrepresented languages and exploring additional emotional categories. We aim to enhance the generalization capability of the model by incorporating more diverse datasets and refining the feature extraction techniques. Furthermore, future work will focus on integrating this emotion recognition model into real-time systems, such as virtual assistants and conversational AI platforms, to improve their responsiveness and empathy in human-machine interactions.

Abbreviations

PBCC	Phase-Based Cepstral Coefficients
DTW	Dynamic Time Warping
RMS	Root Means Square
MFCCS	Mel-frequency cepstral coefficients

References

1. Bashari Rad, B.; Moradhaseli, M. Speech emotion recognition methods: A literature review. *AIP Conference Proceedings* **2017**, *1891*, 020105–1.

2. Ahlam Hashem, M.A.; Alghamdi, M. Speech emotion recognition approaches: A systematic review. *Speech Communication* **2023**, *154*, 102974. <https://doi.org/https://doi.org/10.1016/j.specom.2023.102974>.
3. Muntaqim, M.Z.; Smrity, T.A.; Miah, A.S.M.; Kafi, H.M.; Tamanna, T.; Farid, F.A.; Rahim, M.A.; Karim, H.A.; Mansor, S. Eye Disease Detection Enhancement Using a Multi-Stage Deep Learning Approach. *IEEE Access* **2024**, pp. 1–1. <https://doi.org/10.1109/ACCESS.2024.3476412>.
4. Hossain, M.M.; Chowdhury, Z.R.; Akib, S.M.R.H.; Ahmed, M.S.; Hossain, M.M.; Miah, A.S.M. Crime Text Classification and Drug Modeling from Bengali News Articles: A Transformer Network-Based Deep Learning Approach. In Proceedings of the 2023 26th International Conference on Computer and Information Technology (ICCIT). IEEE, 2023, pp. 1–6.
5. Rahim, M.A.; Farid, F.A.; Miah, A.S.M.; Puza, A.K.; Alam, M.N.; Hossain, M.N.; Karim, H.A. An Enhanced Hybrid Model Based on CNN and BiLSTM for Identifying Individuals via Handwriting Analysis. *CMES-Computer Modeling in Engineering and Sciences* **2024**, *140*.
6. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* **2019**, *7*, 117327–117345.
7. Saad, F.; Mahmud, H.; Shaheen, M.; Hasan, M.K.; Farastu, P. Is Speech Emotion Recognition Language-Independent? Analysis of English and Bangla Languages using Language-Independent Vocal Features. *Computing Research Repository (CoRR)* **2021**.
8. Chakraborty, C.; Dash*, T.K.; Panda, G.; Solanki, S.S. Phase-based Cepstral features for Automatic Speech Emotion Recognition of Low Resource Indian languages. *Transactions on Asian and Low-Resource Language Information Processing* **2022**.
9. Ma, E. Data Augmentation for Audio. *Medium* **2019**.
10. RINTALA, J. *Speech Emotion Recognition from Raw Audio using Deep Learning*; School of Electrical Engineering and Computer Science Royal Institute of Technology (KTH), 2020.
11. Tusher, M. M. R., F.F.A.K.H.M.M.A.S.M.R.S.R.I.M.R.M.A.M.S.K.H.A. BanTrafficNet: Bangladeshi Traffic Sign Recognition Using A Lightweight Deep Learning Approach. *Computer Vision and Pattern Recognition*.
12. Siddiqua, A.; Hasan, R.; Rahman, A.; Miah, A.S.M. Computer-Aided Osteoporosis Diagnosis Using Transfer Learning with Enhanced Features from Stacked Deep Learning Modules. *arXiv preprint arXiv:2412.09330* **2024**.
13. Md. Mahbubur Rahman Tusher, Fahmid Al Farid, M.A.H.A.S.M.M.S.R.R.M.H.J.S.M.M.A.R.H.A.K. Development of a Lightweight Model for Handwritten Dataset Recognition: Bangladeshi City Names in Bangla Script. *Computers, Materials & Continua* **2024**, *80*, 2633–2656.
14. Sultana, S.; Iqbal, M.Z.; Selim, M.R.; Rashid, M.M.; Rahman, M.S. Bangla speech emotion recognition and cross-lingual study using deep CNN and BLSTM networks. *IEEE Access* **2021**, *10*, 564–578.
15. Rahman, M.M.; Dipta, D.R.; Hasan, M.M. Dynamic time warping assisted SVM classifier for Bangla speech recognition. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). IEEE, 2018, pp. 1–6.
16. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* **2020**, *59*, 101894.
17. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control* **2019**, *47*, 312–323.
18. Mustaqeem.; Kwon, S. 1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features. *CMC-Computers Materials & Continua* **2021**, *67*, 4039–4059.
19. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 international conference on platform technology and service (PlatCon). IEEE, 2017, pp. 1–5.
20. Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L.; Schmauch, B. Cnn+ lstm architecture for speech emotion recognition with data augmentation. *arXiv preprint arXiv:1802.05630* **2018**.
21. Ai, X.; Sheng, V.S.; Fang, W.; Ling, C.X.; Li, C. Ensemble learning with attention-integrated convolutional recurrent neural network for imbalanced speech emotion recognition. *IEEE Access* **2020**, *8*, 199909–199919.
22. Kwon, S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2019**, *20*, 183.
23. Zheng, W.; Yu, J.; Zou, Y. An experimental study of speech emotion recognition based on deep convolutional neural networks. In Proceedings of the 2015 international conference on affective computing and intelligent interaction (ACII). IEEE, 2015, pp. 827–831.

24. Sultana, S.; Rahman, M.S.; Selim, M.R.; Iqbal, M.Z. SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *Plos one* **2021**, *16*, e0250173.
25. Das, R.K.; Islam, N.; Ahmed, M.R.; Islam, S.; Shatabda, S.; Islam, A.M. BanglaSER: A speech emotion recognition dataset for the Bangla language. *Data in Brief* **2022**, *42*, 108091.
26. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391.
27. Agnihotri, P. Berlin Database of Emotional Speech (EmoDB) Dataset. *Kaggle* **2020**. Accessed: December 2022.
28. Ippolito, P.P. Data Augmentation Guide [2023 edition], 2023.
29. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; Battenberg, E.; Nieto, O.; Dieleman, S.; Tokunaga, H.; McQuin, P.; NumPy.; et al. librosa/librosa: 0.10.1. <https://zenodo.org/records/8252662>, 2023. <https://doi.org/10.5281/zenodo.8252662>.
30. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the Proceedings of the 14th python in science conference, 2015, Vol. 8, pp. 18–25.
31. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
32. Jordal, I. Polarity inversion, 2018.
33. Jordal, I. Gain, 2018.
34. Titeux, N. Everything you need to know about pitch shifting: Nicolas Titeux, 2023.
35. Kedem, B. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE* **1986**, *74*, 1477–1493. <https://doi.org/10.1109/PROC.1986.13663>.
36. Bourtsoulatz, A. Audio Signal Feature Extraction for Analysis. *Medium* **2020**.
37. Shah, A.; Kattel, M.; Nepal, A.; Shrestha, D. Chroma Feature Extraction. In Proceedings of the Chroma Feature Extraction using Fourier Transform, 01 2019.
38. Zaheer, N. Audio Signal Processing: How Machines Understand Audio Signals, 2023.
39. Behera, K. Feature extraction from audio, 2020.
40. Daehnhardt, E. Audio Signal Processing with python's librosa, 2023.
41. West, K.; Cox, S. Finding An Optimal Segmentation for Audio Genre Classification. 01 2005, pp. 680–685.
42. Peeters, G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project **2004**.
43. Fabien, M. Sound Feature Extraction, 2020.
44. Saranga-K-Mahanta-google.; Arvindpdmn. Audio Feature Extraction. *Devopedia* **2021**.
45. Wu, J. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China* **2017**, *5*, 495.
46. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal Of Big Data* **2021**, *8*. <https://doi.org/10.1186/s40537-021-00444-8>.
47. Zafar, A.; Aamir, M.; Mohd Nawi, N.; Arshad, A.; Riaz, S.; Alruban, A.; Dutta, A.; Alaybani, S. A Comparison of Pooling Methods for Convolutional Neural Networks. *Applied Sciences* **2022**, *12*, 8643. <https://doi.org/10.3390/app12178643>.
48. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* **2015**.
49. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **2014**, *15*, 1929–1958.
50. GeeksforGeeks. What is a Neural Network Flatten Layer?, 2024.
51. Nwankpa, C.; Ijomah, W.; Gachagan, A.; Marshall, S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning, 2018, [[arXiv:cs.LG/1811.03378](https://arxiv.org/abs/cs.LG/1811.03378)].
52. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization, 2017, [[arXiv:cs.LG/1412.6980](https://arxiv.org/abs/cs.LG/1412.6980)].
53. Mohammed, A.; Kora, R. A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges. *Journal of King Saud University - Computer and Information Sciences* **2023**, *35*. <https://doi.org/10.1016/j.jksuci.2023.01.014>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.