# PERFORMANCE IMPROVEMENT OF SPEECH EMOTION RECOGNITION IN BENGALI LANGUAGE USING DEEP LEARNING AND BLSTM NETWORKS

**BY**

**MD SHUMON MIA**

**ID: 212-15-14768**

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

## Supervised By

**Dr. S. M. Aminul Haque**
Professor and Associate Head
Department of CSE
Daffodil International University

## Co-Supervised by

**Md. Sazzadur Ahamed**
Assistant Professor
Department of CSE
Daffodil International University


**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JULY 2024**

# APPROVAL

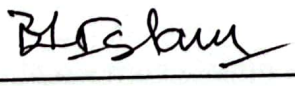
This Project titled **"Performance Improvement of Speech Emotion Recognition in Bengali Language Using Deep Learning and BLSTM Networks"**, submitted by **"Md Shumon Mia"**, ID No: **212-15-14768** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on July 16, 2024.

## BOARD OF EXAMINERS

**Dr. Sheak Rashed Haider Noori (SRH)** — **Board Chairman**
**Professor and Head**
Department of CSE
Daffodil International University

**Dr. Md. Zahid Hasan (ZH)** — **Internal Examiner 1**
**Associate Professor**
Department of CSE
Daffodil International University

**Mr. Abdus Sattar (AS)** — **Internal Examiner 2**
**Assistant Professor**
Department of CSE
Daffodil International University



**Dr. Md. Manowarul Islam (MI)** — **External Examiner**
**Associate Professor**
Department of Computer Science and Engineering
Jagannath University

## DECLARATION

I hereby declare that this study was completed under the supervision of **Dr. S. M. Aminul Haque, Professor and Associate Head, Department of CSE** Daffodil International University. I also declare that neither this research project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**

**Dr. S. M. Aminul Haque**
Professor and Associate Head
Department of CSE
Daffodil International University

**Co- Supervised by:**

**Mr. Md. Sazzadur Ahamed**
Assistant Professor
Department of CSE
Daffodil International University

**Submitted by:**

**Md Shumon Mia**
Id Number: 212-15-14768
Department of CSE
Daffodil International University

# ACKNOWLEDGEMENT

As a matter of first importance, we are so thankful and grateful to Almighty ALLAH for his kindness in enabling us to fulfill this research effectively.

Second, I would like to express sincere thankfulness to my supervisor **Dr. S. M. Aminul Haque, Professor and Associate Head,** Department of Computer Science and Engineering, Daffodil International University. I appreciate all his support and motivation through this research. This research could not have been completed without his inexhaustible patience, intellectual direction, continuous encouragement, constant and energetic supervision, constructive criticism, invaluable counsel, reviewing several substandard drafts and correcting them at every level.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor and Head, Department of CSE**, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I want to thank our parents who always trusted us, constantly supported and loved us, because without their love and support we would not be able to reach this level to complete our research. I also might want to thank the authority of Daffodil International University for giving us such a good research environment.

# ABSTRACT

Speech Emotion Recognition (SER) is a new field in artificial intelligence (AI) and Bengali signal processing that has the potential to improve targeted user interactions and enable more positive interactions with smart devices. The goal of this work is to improve SER for Bengali, a language with limited resources in the suggested domain. Additionally, a novel deep learning model (DCNN-BLSTM) is proposed, which aims to improve the accuracy of emotion recognition by combining 1D-CNN, TDF, BLSTM networks. In this article, a deep learning model is trained using the audio data's Mel-Frequency Cepstral Coefficients (MFCCs) to create a system that can identify audio signals almost exactly like a human auditory system. Mel Frequency Cepstral Coefficients (MFCCs) are obtained by decoding the audio signal before proceeding with local feature learning blocks (LFLBs), which create the feature values using one-dimensional convolutional neural networks (CNNs). Due to the temporal properties of audio signals, these feature values are then added to the Bi-LSTM layer, which helps to improve temporal learning. The TDF layers ensure that temporal dynamics are preserved throughout the processing stages, while the Dropout layer improves model generalization. Lastly, the procedures of categorization and prediction are carried out using fully connected layers. The Bi-LSTM model demonstrates that the recovered features by 1D CNN are well captured because of the time-series properties of speech signals, as can be observed from the experimental evaluation of the SUBESCO database. Additionally, this study employs five distinct data augmentation strategies, each of which helps to increase recognition accuracy. On the SUBESCO dataset, the suggested model produced promising accuracy of 88%, respectively. The results indicate that, in comparison to comparable studies in voice emotion recognition, the suggested approach obtains greater recognition rates.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1
# INTRODUCTION

## 1.1 Introduction

From its early focus on optimizing digital speech transmission to its current advanced applications of deep learning techniques, Speech Emotion Recognition (SER) has seen tremendous evolution. Due to advancements in computer hardware, the focus of SER research has shifted during the past 20 years from internet-based voice media to device-based applications. With the aim of developing systems that can precisely identify and categorize emotions expressed in human speech, speech recognition has emerged as a major field of study in voice media during the past ten years [1]. Numerous fields, including robotics [2], customer service [3], education [4], mental health [5], and human-computer interaction [6], have benefited from the broad and significant uses of SER. The two main steps of the SER job are usually feature extraction from the speech signal and emotion classification. For emotion recognition, existing approaches use a variety of signal processing and feature extraction techniques, then apply machine learning (ML) and deep learning (DL) techniques [7][8]. By examining characteristics including tone, intonation, and rhythm, automatic speech emotion detection uses artificial intelligence technology to detect emotional states like happy, sorrow, anger, and worry [9]. Although SER technology has been thoroughly examined for a number of major languages, there is still a dearth of materials for research on the Bengali language. Bengali is not well-represented in SER studies, although being spoken by over 230 million people globally [10], which makes it an interesting field for additional ML and DL research.

The use of speech emotion detection on portable devices is a recent development in SER. Speech has emerged as a vital mode of communication between users and smart mobile devices as a result of their widespread use. A terminal device can improve the user experience by providing more humanistic and personalized interactions, going beyond robotic and repetitive responses to offer more individualized services, if it can identify the user's emotional state through speech cues. SER technology does have a number of severe obstacles, though. Accurate recognition is impacted by the variety of speech signals produced in different languages by different people. Furthermore, interferences such as microphone distortion and ambient noise frequently affect voice signals, making recognition more difficult. The largest barrier to consistent and accurate emotion recognition among users is subjectivity, as different people may interpret and experience the same voice signal in various ways.

This research integrates bi-directional long short-term memory (BLSTM) networks, Time Distribution Flatten (TDF) layer, and convolutional neural networks (CNN) to build a more stable and accurate voice emotion recognition system. Mel-Frequency Cepstral Coefficients (MFCCs) are used in the method to extract features. A variety of data augmentation techniques, including pitch shifting, time stretching, time shifting, adding noise, speed adjustment, and combining augmentation techniques, are used to improve the resilience of the model. One-dimensional (1D) and two-dimensional (2D) convolutions are both part of the CNN architecture. 2D convolution computations are more complex than 1D convolutions since voice data must be transformed from the time domain to the frequency domain.

This study takes into account the power and processing cost of portable smart devices and suggests a DL model for Bangla vocal emotion recognition based on a 1D CNN. The experimental results demonstrate that the proposed technique provides more generalizable data and a greater recognition rate when compared to other similar investigations. The examination of the SUBESCO database shows how effectively the model captures the dynamics of voice signals. Furthermore, by creating a Bengali dataset specifically for SER, comparative research with other regional languages ought to be made simpler, enhancing our knowledge and expertise in voice recognition for emotional expression. In conclusion, this study proposes significant enhancements to SER architectures, such as a new BLSTM model with MFCCs feature extraction and comprehensive data augmentation techniques. This model provides cutting-edge performance at reduced computational costs, setting the stage for more dependable and efficient SER systems.

## 1.2 Motivation

Voice contact has emerged as a crucial success factor for providing more individualized services with the advent of smart portable devices. The capacity to recognize and react to human emotions with precision is crucial as these gadgets become a more and more important part of our everyday lives. With over 230 million people speaking Bengali, there is an especially high demand for effective Speech Emotion Recognition (SER). Research and development for underrepresented languages like Bengali are still scarce, despite the global focus on SER. Enhancing user experience by enabling more natural and sympathetic interactions, better SER performance in Bengali would guarantee that technical gains in natural language processing are inclusive.

Improving SER in Bengali has important socioeconomic ramifications as well. It can spur innovation in a number of industries, such as education, healthcare, and telecommunications, by offering emotionally intelligent services that more effectively cater to Bengali-speaking customers' requirements. Applications like virtual assistants, customer support platforms, and mental health support systems will work more efficiently if Bengali emotion recognition is accurate. By investigating and putting into practice cutting-edge methods to enhance SER performance in the Bengali language, this thesis seeks to close the current research gap and eventually promote more inclusive and efficient voice interaction technologies.

## 1.3 Objective

Voice contact is now essential for providing tailored services due to the growing ubiquity of smart portable devices in daily life. But the efficacy of these exchanges depends on one's capacity for precise emotional recognition and interpretation. Speech Emotion Recognition (SER) for more widely spoken languages, such English and Mandarin, has advanced, but less so for less widely spoken languages, like Bengali. The underrepresentation of Bengali language in SER research hinders the voice-driven technologies' usability and accessibility for this large demographic. Add a dataset of Bengali speech to the problem. “SUBESCO” has been used to perform the analysis. Following the dataset's division into three sections, the study concentrated on test data accuracy and how we may more precisely predict the labels for emotions in Bengali voice-driven systems by using this accuracy.

## 1.4 Research Questions

1. What unique difficulties exist when interpreting Bengali speech for emotional content, and how could these difficulties be overcome?
2. How do various methods of data augmentation affect the Bengali SER performance?
3. In Bengali speech, how might deep learning models increase the accuracy of emotion recognition?

## 1.5 Report Layout

- Chapter 1: Outline the main research questions, motivation, goals, and research introduction.
- Chapter 2: provides a brief overview of the literature study, comparative analysis, and performance discussion.
- Chapter 3: The research approach, which includes feature extraction, dataset description, data augmentation strategy, and a proposed optimal model to increase Bangla SER performance.
- Chapter 4: provides a brief discussion of the findings and potential future directions.
- Reference

# CHAPTER 2
# BACKGROUND STUDY

## 2.1 Related Works

Though SER has been the focus of numerous studies for other languages, particularly English, very little has been done to create SER for Bangla. This section contains related articles in the topic of speech emotion recognition (SER). For the classification of acoustic sceneries, Paper [16] proposes to use LDA-based data augmentation to generate new samples and expand the dataset. By transforming unprocessed audio data into LDA space, this is accomplished. Pitch shift, white noise, and temporal stretch are some of the data augmentation techniques used in Paper [17] to expand the database. Spectral contrast, MFCCs, delta MFCCs, Mel spectrogram, Chromagram, and Tonnetz are the seven speech features that are extracted.

A ResNet-based approach for SER is presented in Paper [19]. Speech data is divided into frames, which are then clustered using K-means and replaced with cluster centroids. Utilizing the short-time Fourier transform, audio signals are transformed into a 2D spectrum. Time-series characteristics are represented by a bidirectional LSTM, and frequency domain data is utilized to train the ResNet 101 model. Accuracy values of 77.02% on RAVDESS, 85.57% on EMO-DB, and 72.25% on IEMOCAP are obtained with this approach. 1D and 2D CNN-LSTM models are compared in Paper [20], where the 1D model uses raw speech data while the 2D model uses Mel spectrograms. The 2D CNN-LSTM model obtains an accuracy of 95.89% on EMO-DB.

Paper [21] makes use of MFCCs, log-power Mel spectrogram, and chromagrams. To obtain 85.89% accuracy on RAVDESS and 95.93% accuracy on EMO-DB, it makes use of a residual bidirectional LSTM and an attention-based ML model. Paper [22] presents a novel short-term memory (SER) system based on deep learning. It combines a deep convolutional neural network (DCNN) and a bidirectional long-short term memory (BLSTM) network with a time-distributed flatten (TDF) layer. With weighted accuracies (WAs) of 86.9% on the Bangla SUBESCO dataset and 82.7% on the English RAVDESS dataset, this model exhibits cutting-edge perceptual efficiency. Furthermore, a CNN-LSTM model with attention-based BLSTM layers was described by Zhao et al. (2018), and it achieved a 68% WA on the IEMOCAP dataset [23]. Using a CNN-LSTM architecture, Etienne et al. (2018) obtained a 64.5% WA on the IEMOCAP dataset using different configurations of the CNN and BLSTM layers [24].

Using MFCC features on a small dataset of 200 words, Rahmanet al. (2018) proposed a Dynamic Time Warping-assisted SVM emotion classifier for Bangla words that achieved 86.08% accuracy [25]. A CNN with three convolutional layers and three fully connected (FC) layers was proposed by Badshah et al. (2017). It was trained using the Berlin Emotional Corpus and achieved 56% accuracy [26]. Using log-spectrograms as feature vectors, Satt et al. (2017) reported a model that experimented with convolution-only and convolution-LSTM architectures, obtaining 66% and 68% accuracy.

The DCNN (Convolutional, BLSTM and TDF) architecture from [22] shows the 2D CNN network outperformed the 1D CNN network for speech emotion recognition tasks where the Mel-Spectrogram feature has been used. The proposed model achieved a weighted accuracy (WA) of 81.56% for the SUBESCO dataset using the 1D CNN network and a weighted accuracy of 86.9% using the 2D CNN network.

For audio-only speech files in the RAVDESS dataset, the 2D CNN network achieved a weighted accuracy (WA) of 77.8%, whereas the 1D CNN network achieved a weighted accuracy (WA) of 66.43%. This research was enhanced in our study by integrating the MFCCs feature with five distinct data augmentation techniques, including 1D CNN, BLSTM, and TDF layer. When compared to the CNN-DNN and DCNN-BLSTM models, the DCNN-BLSTM model performed better (88% more accurate) with enhanced data.

### 2.2 Comparative Analysis and Summary

2.2.1 Data Augmentation Techniques:

- LDA-Based Augmentation (Paper [16]): This method creates new samples to increase the dataset by converting raw audio data into LDA space by using LDA.
- Conventional Augmentation Methods (Paper [17]): Enhances the dataset by applying pitch shifting, addition of white noise, and temporal stretch; features such MFCCs, spectral contrast, Mel spectrogram, chromagram, delta MFCCs, and tonnetz are extracted.

2.2.2 Feature Extraction Methods:

Common Features: Studies consistently use MFCCs, spectral contrast, Mel spectrogram, chromagram, delta MFCCs, and tonnetz for effective emotion classification.

2.2.3 Model Architectures and Performances:

1. ResNet-Based Approach (Paper [19]): Uses K-means clustering and ResNet 101, achieving accuracies of 77.02% (RAVDESS), 85.57% (EMO-DB), and 72.25% (IEMOCAP).

2. 1D vs. 2D CNN-LSTM Models (Paper [20]): The 2D CNN-LSTM model using Mel spectrograms achieves 95.89% accuracy on EMO-DB.

3. ABMD and RBLSTM (Paper [21]): Combines attention mechanisms with bidirectional LSTM, achieving 85.89% on RAVDESS and 95.93% on EMO-DB.

4. Deep Learning-Based Short-Term Memory System (Paper [22]): Combines BLSTM, TDF layer, and DCNN, achieving 86.9% on Bangla SUBESCO and 82.7% on RAVDESS.

5. CNN-LSTM Models: Zhao et al. (2018) and Etienne et al. (2018) achieve weighted accuracies of 68% and 64.5% on IEMOCAP.

6. Dynamic Time Warping-Assisted SVM (Rahman et al., 2018): Achieves 86.08% accuracy using MFCC features on a small Bangla dataset.

7. Convolutional Models: Badshah et al. (2017) and Satt et al. (2017) achieve accuracies ranging from 56% to 68%.

### 2.3 Performance Comparison and Conclusion

The comparative analysis highlights the effectiveness of advanced architectures and data augmentation techniques in enhancing SER performance. Notably, the DCNN architecture in Paper [22] demonstrates superior performance with the 2D CNN network achieving higher weighted accuracy (86.9% for Bangla SUBESCO, 77.8% for RAVDESS) compared to the 1D CNN network (81.56% for Bangla SUBESCO, 66.43% for RAVDESS). Moreover, the DCNN-BLSTM model, enhanced with data augmentation and MFCC features, outperforms other models, showing an 88% improvement in accuracy. This comparative analysis underscores the importance of innovative model architectures and data augmentation techniques in advancing SER research, particularly for underrepresented languages like Bangla.

# CHAPTER 3
# RESEARCH METHODOLOGY

## 3.1 Introduction

In Bengali, Speech Emotion Recognition (SER) is the process of recognizing and comprehending emotions expressed through speech, which is essential for improving user interaction with technology. When it comes to creating precise SER systems that take into account the linguistic and cultural subtleties of Bengali, there are particular opportunities and problems. Enhancing SER in Bengali has several uses in mind, from mental health assistance to virtual assistants, with the goal of making technology more sensitive to emotional cues in Bengali speakers. Researchers discussed various approaches and used various algorithms for their specific research goals in a number of review papers. We gathered a synthetic dataset for our study from the Kaggle source. Seven different categories of emotions are labeled on 7000 audio data points in the collection. To measure the accuracy, we employ two distinct deep learning models in this experiment. Rather than using CNN-DNN models, data augmentation strategies take advantage of the superior performance of DCNN-BLSTM models. The study is titled "Performance Enhancement of Bengali Speech Emotion Recognition (SER) Systems through the Integration of 1D CNN and TDF + BLSTM with Data Augmentation" and for experiment we use the Kaggle kernel as a platform and implement our technique in Python.

## 3.2 Dataset Description

This study made use of the "SUBESCO" dataset, an Open Access audio-only emotive speech corpus designed especially for the Bengali language. It was developed as a part of a doctoral research project at Shahjalal University of Science and Technology, in Sylhet, Bangladesh, in the Department of Computer Science and Engineering. With 7000 utterances over 7 hours, this is the largest emotional speech corpus accessible for Bengali. The exercise was recorded by twenty native speakers, ensuring a gender balance. In ten words, each participant imitated seven different emotions. The corpus was evaluated by fifty college students as well; every audio clip, with the exception of those that showed disgust, was validated four times by raters who were both male and female.

The database used in this paper's experiments for both model validation and training. To test the machine learning model's dependability, 10% of the dataset is randomly selected using the Scikit-Learn package, and the remaining 90% of the dataset is randomly split into 9:1 train and validation data sets.

### 3.3 Data Augmentation

Data augmentation, or ADA, is a technique that may be used to increase the size of an existing dataset. It is frequently employed in computer vision and machine learning to improve model performance. By using data augmentation, it is possible to increase the quantity of useful training and testing data, which can aid in the DL model's ability to learn additional features and prevent under-fitting. Some researchers compare different data augmentation techniques using the findings reported in [16, 17]. These evaluations lead them to select two comparatively straightforward data augmentation techniques that produce superior results: pitch shifting and noise addition.

For the first time, Bangla audio signals are methodically subjected to five distinct data augmentation approaches, which greatly increases the overall volume of audio data and the reliability of feature extraction procedures in deep learning applications. Using an add noise function with a noise factor of 0.035, random Gaussian noise is added to each audio signal, producing a variety of noise patterns that mimic differences in the actual world.

Stretch time causes signals to time-stretch at a rate of 1.1, enabling changes in signal speed without affecting pitch properties. In order for the model to learn and identify speech patterns over a broad variety of speaking rates, this characteristic is especially important. Furthermore, by introducing temporal displacements within audio segments, which are managed by a shift_max parameter of 0.2, the shift_time function improves the model's capacity to generalize temporal properties by simulating various starting positions. Shift_pitch modifies tonality to capture frequent speech changes by adjusting the signal's pitch by a pitch_factor of 0.7 while maintaining signal duration. Change_speed simultaneously adjusts the playback speed by a factor of 1.5 without changing pitch, which closely mimics situations in which speech happens at various speeds in the actual world. To extensively augment signals, the combine_augmentation function gradually incorporates time stretching, pitch shifting, time shifting, noise addition, and speed modification. By adding a variety of audio changes to the dataset, this method improves the model's robustness and performance in real-world scenarios.

The output signals from the data augmentation techniques used in this paper for an audio file are displayed in Figure 3.3

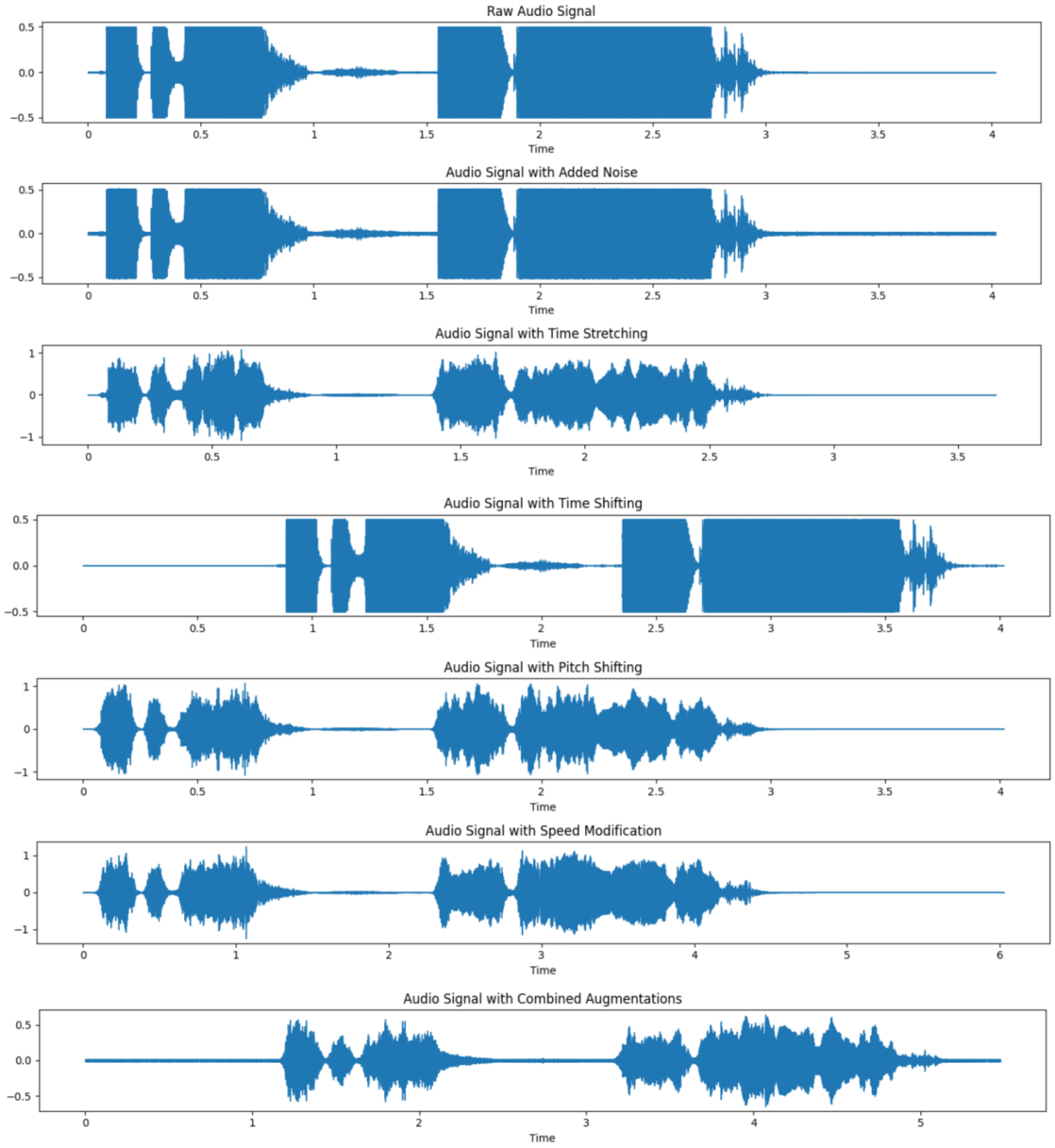



**Figure 3.3:** (a) Raw audio signal; and Augmented audio signal by adding (b) added noise; (c) time stretching; (d) time shifting; (e) pitch shifting; (e) Audio signal with speech modification (f) combined augmentation

### 3.4 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) are the most commonly employed speech feature values for speech recognition and speaker identification. The database's raw audio data will be transformed into MFCCs for this study. During training, these MFCCs will subsequently be incorporated into the neural network model. Audio signals are converted into a set of feature values that replicate how the human auditory system reacts to various frequency components in spoken sounds by extracting MFCCs. Figure 3.4 shows the MFCC extraction procedure.

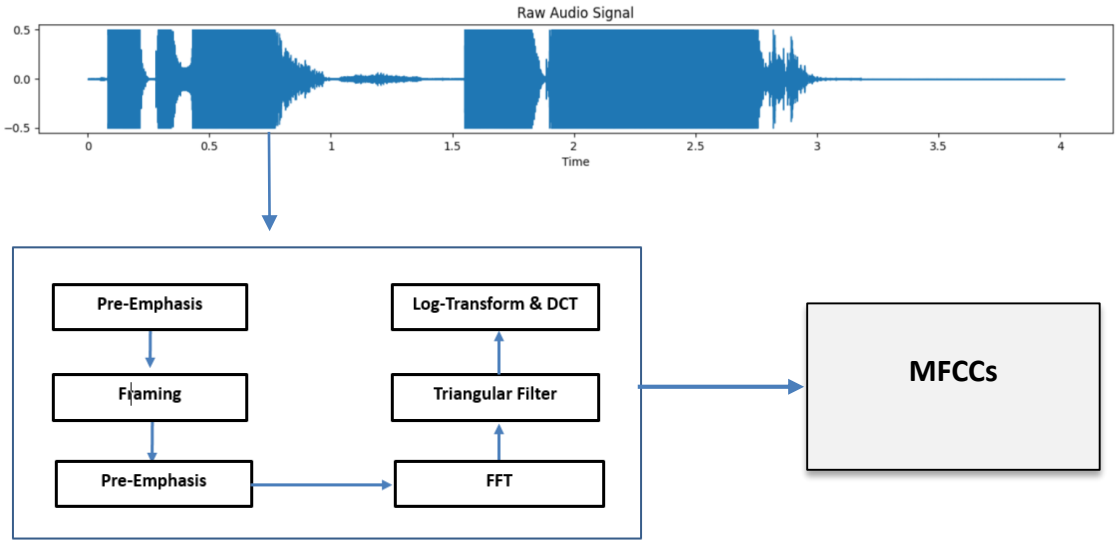



**Figure 3.4:** Extracting MFCCs feature for audio signals.

To extract Mel-frequency Cepstral Coefficients (MFCCs) from audio signals, several steps are involved, which are briefly explained mathematically based on the provided code:

1. **Pre-emphasis**: Enhance high frequencies to balance the frequency spectrum. Emphasized signal[n] = signal[n] - α·signal [n-1], where α is the pre-emphasis coefficient (default α =0.95)
2. **Frame Blocking:** Divide the signal into overlapping frames. Frames [i, t] = signal [t × frame_step+i], where frame_step and frame_length determine the overlap and length of each frame respectively.
3. **Hamming Windowing:** Apply a Hamming window to each frame to reduce spectral leakage. Hamming_frames[i, t] = frames[i, t]·(0.54-0.46 cos (2πt/ fame_length - 1))

4. **Fast Fourier Transform (FFT)**: Compute the magnitude of the FFT for each frame.
   mag_frames[i, k]= |FFT(hamming_frames[i])|
5. **Power Spectrum**: Calculate the power spectrum.
   pow_frames[i, k]=NFFT1(mag_frames[i, k])^2
6. **Mel Filter Banks**: Convert the power spectrum to Mel-scale. Where fbank is a matrix of triangular Mel Filter Bank Coefficients.

$$filter - banks[i,m] = \sum_{K=0}^{NFFT/2} pow_frames[i,k] \odot fbank[m,k]$$

7. **Logarithm:** Covert filter bank outputs to decibels (db)
   Filter_banks [i,m] = 20 · log10 (filter_banks [I,m]
8. **Discrete Cosine Transform (DCT):** Extract MFCCs from the filter bank energies.

$$mfccs[i,c] = \sum_{m=1}^{nfilt} filter - banks[i,m] \odot cos[\frac{\pi c}{nfilt}(m - \frac{1}{2})]$$

   Where c represents the MFCC index and nfilt is the number of Mel Filters.
   In this paper, after extracting MFCCs with a dimensionality of 13.

### 3.5 Proposed Methodology and Experimental Setup

The proposed methodology enhances Speech Emotion Recognition (SER) for Bengali using an innovative DCNN-BLSTM model that integrates Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and Deep Neural Networks (DNN). Initially, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from audio signals and processed by Local Feature Learning Blocks (LFLBs), which employ 1D CNNs to capture local patterns. These features are then passed into the Bi-LSTM layer, which learns temporal dynamics, followed by a Dropout layer for better generalization. The Time-Distributed Flattening (TDF) layer preserves temporal structure, while fully linked layers are employed for classification and prediction. Furthermore, five data augmentation techniques—pitch shifting, white noise addition, temporal stretching, time shifting, change speed and combination augmentation—are used to improve recognition accuracy.

This paper introduces two models for voice emotion recognition tasks: the CNN-DNN model and the 1D DCNN-BLSTM model with TDF and BLSTM layers. Multiple experiments were conducted using an Intel i5-8600 processor for data augmentation, MFCC extraction, and deep learning training and testing, with the numerical results

published in the study. Compared to previous techniques in Bengali SER, our strategy demonstrated considerable gains, achieving a remarkable 88% accuracy on the SUBESCO dataset for detecting seven emotions: happy, sad, disgust, angry, fear, neutral, and surprised. Table 1 shows the experimental setup.

**Table 3.5:** Experimental Environment.

| Experimental Environment |
|---|
| CPU-Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz-1.90 GHz, RAM-8GB |
| IDE- Kaggle Notebook Editor |

### 3.5.1 CNN-DNN Model

The architecture and training process of the CNN-DNN model used in this research are illustrated in Figure 3.5.1 (A). Following data enhancement, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the audio data and fed into the model. The model constructs feature maps using five Local Feature Learning Blocks (LFLBs), each consisting of one-dimensional CNN layers and a hidden layer for feature extraction. The output layer employs the softmax activation function to generate the classification results.

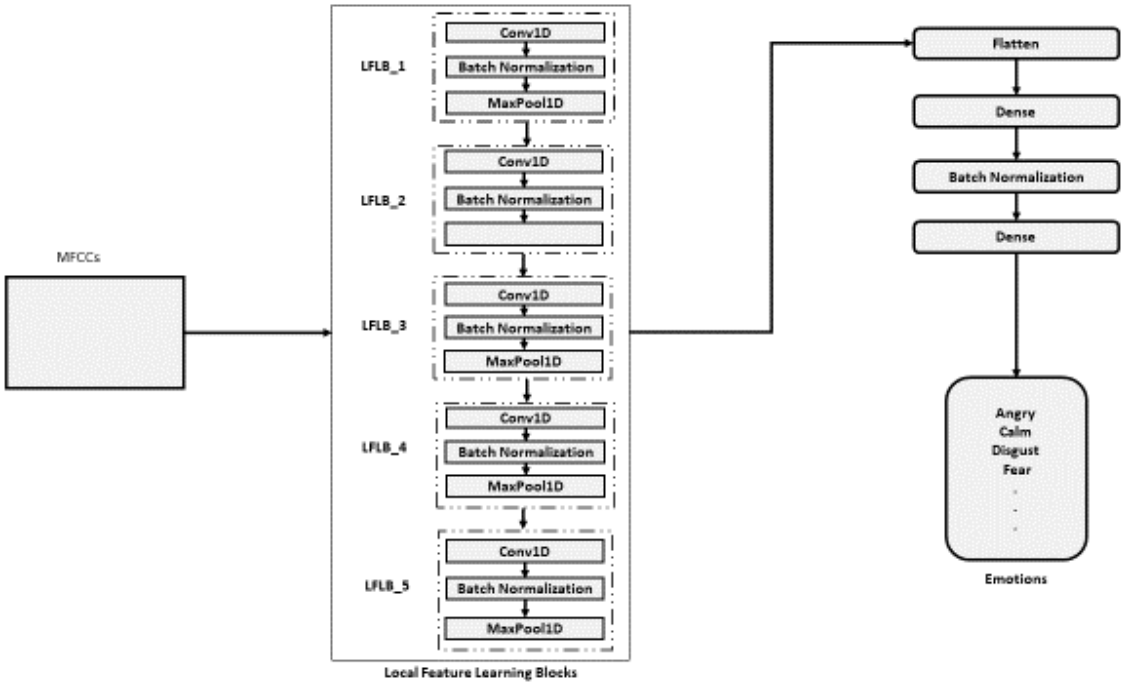



**Figure 3.5.1 (A):** CNN-DNN model architecture and training procedure using MFCC characteristics.

The experiments utilize the Rectified Linear Unit (ReLU) activation function and apply a zero-padding approach to all convolutional layers. Each convolutional layer also includes a max pooling layer and a batch normalization layer to aid in feature value convergence and prevent gradient vanishing. Table 3.5.1 provides a detailed breakdown of the CNN-DNN model parameters used in this study.

In order to evaluate the effects of data augmentation, this study contrasted the results of raw audio data with augmented audio data using model architectures that were equivalent and had the same parameters applied to the SUBESCO training datasets. Cross-validation was not used in the model's performance evaluation. The average accuracy obtained with raw data, using the SUBESCO database, was 75%. Following data augmentation, there was an 8% improvement in the average accuracy, which rose to 83%. These findings demonstrate how the CNN-DNN model's performance is highly influenced by the caliber of the training set. Furthermore, the study's data augmentation methods shown efficacy in improving model training. Figures 43.5.1 (B) and 3.5.1 (C) illustrate the confusion matrices that show the experimental findings for the CNN-DNN model trained on augmented data and raw audio data, respectively.

**Table 3.5.1.** CNN-DNN model Parameters used in this paper

| | LFLB_1(input) | LFLB_2 | LFLB_3 | LFLB_4 | LFLB_5 |
|---|---|---|---|---|---|
| **Layer** | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D |
| **Layer value** | filters = 256, kernel size = 5, strides = 1, pool size = 5, strides = 2 | filters = 128, kernel size = 5, strides = 1, pool size = 5, strides = 2 | filters = 128, kernel size = 5, strides = 1, pool size = 5, strides = 2 | filters = 64, kernel size = 3, strides = 1, pool size = 5, strides = 2 | filters = 64, kernel size = 3, strides = 1, pool size = 3, strides = 2 |

| Common Layer for all LFLB | Respective Value |
|---|---|
| Flatten layer | |
| Dense layer | Units= 256, Activation = “relu” |
| Batch_Normalization | |
| Dense (Output) | Activation = “Softmax”) |

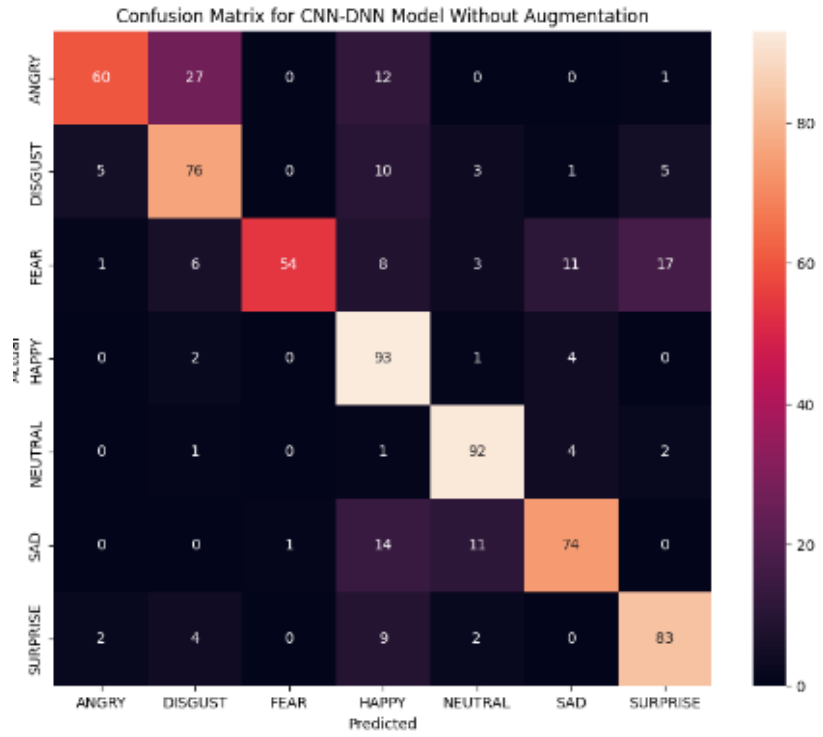



**Figure 3.5.1 (B):** Confusion matrix of CNN-DNN model accuracy using raw audio data

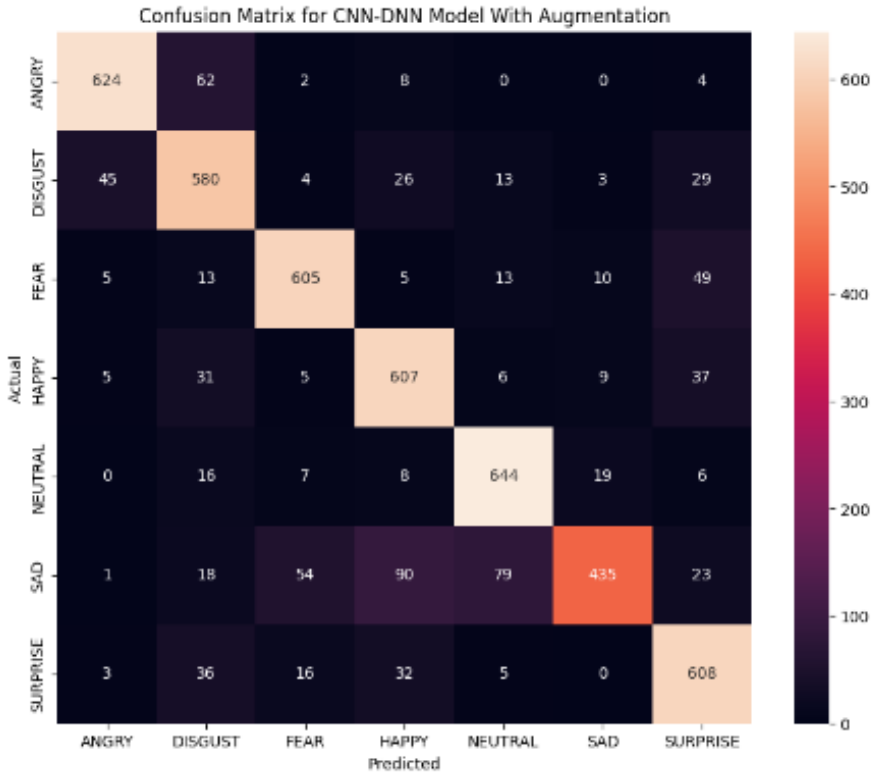



**Figure 3.5.1 (C):** Confusion matrix of CNN-DNN model accuracy using augmented audio data

### 3.5.2 DCNN-BLSTM Model

The primary goal of convolutional neural networks (CNNs) is to learn local features. However, training with recurrent neural networks (RNNs) can aid the model in learning features over time, as audio is temporal sequence data. As a result, this article suggests a DCNN-BLSTM model that combines Time-Distributed Fully Connected (TDF) layers between the convolutional layers and hidden layers with a bidirectional long short-term memory (BiLSTM) network. The relationship between feature values and their temporal variations throughout the training process may now be examined by the model. Figure 3.5.2 (A) depicts the model's architecture and training procedure. The architecture preserves temporal information across the network and avoids overfitting by combining BiLSTM, TDF, and Dropout layers into the DCNN-BLSTM model. This allows the architecture to accurately represent the temporal dependencies present in audio data.

For the purpose of identifying emotions in speech, the BiLSTM network aids in comprehending the timing and sequence of features in both forward and backward directions. The Dropout layer enhances model generalization, while the TDF layers guarantee that temporal dynamics are maintained throughout the processing phases. With these improvements, the model is more reliable and efficient for tasks involving the recognition of speech emotions.

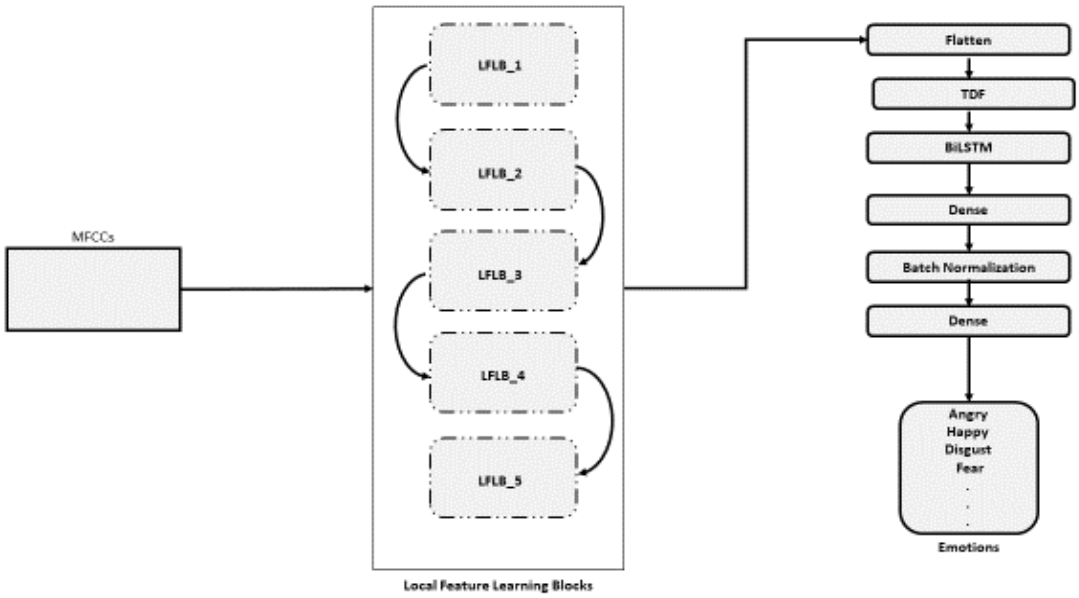



**Figure 3.5.2 (A):** DCNN-BLSTM model's architecture and training procedure using MFCC characters

In this paper, the BLSTM layer's output size is set to 128 and dropout value is set to 0.3.

**Table 3.5.2:** DCNN-BLSTM model Parameters used in this paper

| | LFLB_1(input) | LFLB_2 | LFLB_3 | LFLB_4 | LFLB_5 |
|---|---|---|---|---|---|
| **Layer** | Conv1D Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D | Conv1D_Batch Normalization MaxPooling1D |
| **Layer value** | filters = 256, kernel size = 5, strides = 1, pool size = 5, strides = 2 | filters = 128, kernel size = 5, strides = 1, pool size = 5, strides = 2 | filters = 128, kernel size = 5, strides = 1, pool size = 5, strides = 2 | filters = 64, kernel size = 3, strides = 1, pool size = 5, strides = 2 | filters = 64, kernel size = 3, strides = 1, pool size = 3, strides = 2 |

| Common Layer for all LFLB | Respective Value |
|---|---|
| TDF layer | |
| BLSTM<br>Dropout (0.3) | Units= 128 |
| Flatten layer | |
| Dense layer | Units= 256, Activation = “relu” |
| Batch_Normalization<br>Dropout (0.3) | |
| Dense (Output) | Activation = “Softmax”) |

Similarly, the SUBESCO database was utilized to train the DCNN-BLSTM model to evaluate the impact of data augmentation on model training. Initially, the model achieved an average accuracy of 80% using unprocessed audio data. After applying data augmentation, with a batch size of 32 and 16 epochs, the model's average accuracy increased to 88%, representing an 8% improvement.

Figures 3.5.2 (B) and 3.5.2 (C) display the confusion matrices of the experimental results on the SUBESCO database before and after data augmentation, respectively.

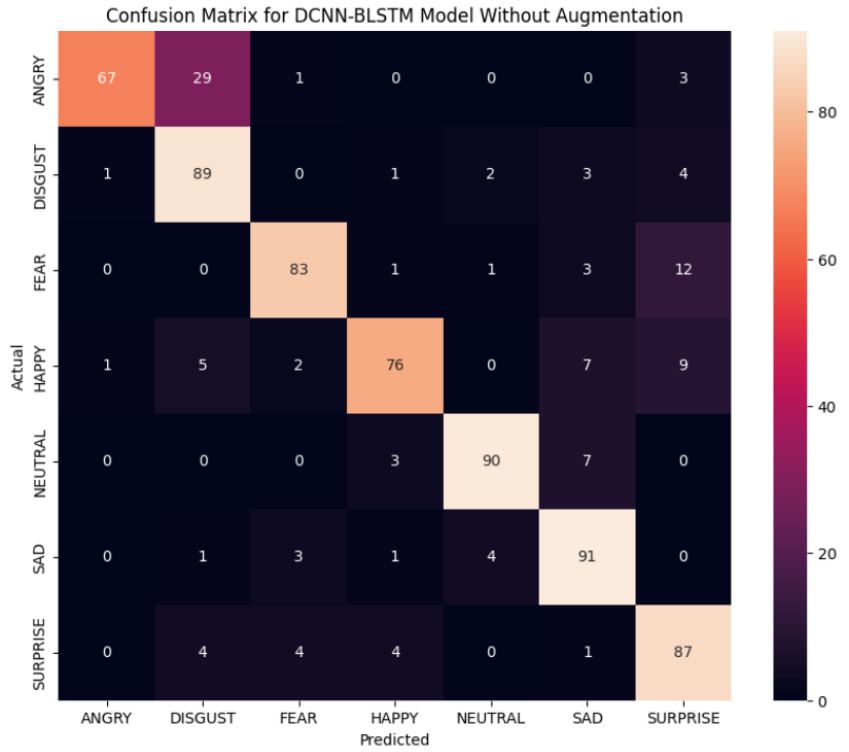



Figure 3.5.2 (B): Confusion matrix of DCNN-BLSTM model accuracy using raw audio data

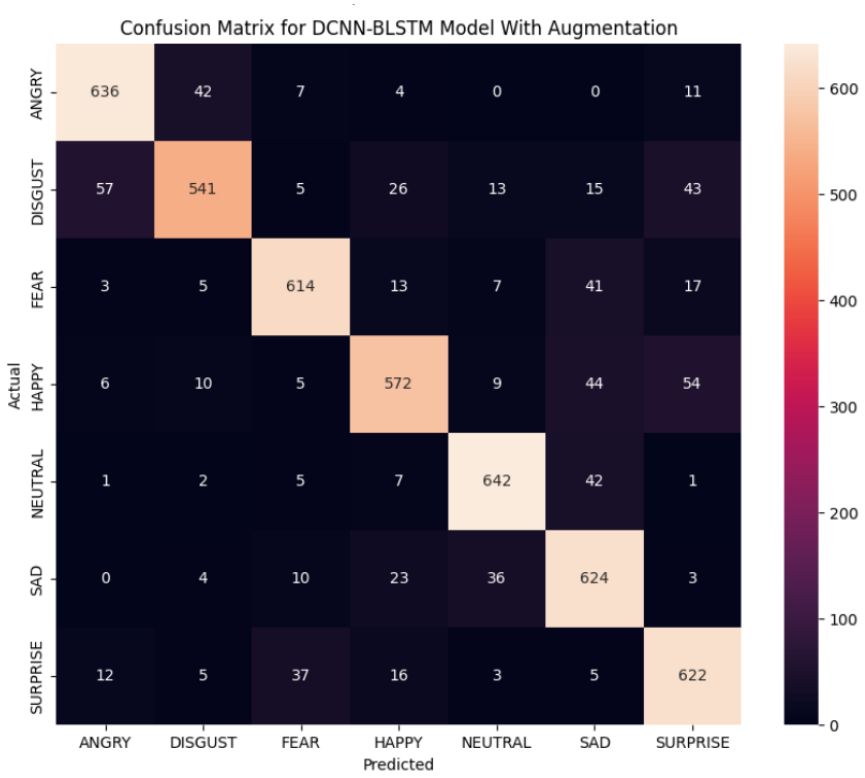



Figure 3.5.2 (C): Confusion matrix of DCNN-BLSTM model accuracy using augmented audio data

# CHAPTER 4
# CONCLUSION AND FUTURE SCOPE

## 4.1 Results and Discussion

Comparing the training outcomes of the two models (CNN-DNN and DCNN-BLSTM), the research finds that the DCNN-BLSTM model has an average accuracy of 80% without data augmentation, which is higher than that of the CNN-DNN model. Moreover, after data augmentation, the average accuracy of the DCNN-BLSTM model increases by 8%, which is similar to the improvement of the CNN-DNN model. This implies that adding TDF and BLSTM layers can enhance training performance by effectively learning features over time periods. The average accuracy of the CNN-DNN and DCNN-BLSTM models applied to both augmented and raw audio data is displayed in Table 4.1.

**Table 4.1:** Model Camparison

| Model | Raw_Audio_Data | Augmented_Data |
|---|---|---|
| **CNN-DNN** | **75%** | **83%** |
| **DCNN-BLSTM** | **80%** | **88%** |

## 4.2 Conclusions and Future Work

The benefits of employing a 1D CNN model for speech emotion recognition (SER) are demonstrated in this work, especially for small and light devices. Local patterns in acoustic waves are captured by the 1D CNN, which effectively analyzes sequential data and extracts features to determine emotions. Real-time emotion recognition on devices with constrained processing power is made possible by its computational efficiency, which is manifested by fewer parameters and a reduced memory requirement. Furthermore, 1D CNNs are perfect for SER systems on embedded and mobile devices because they can adjust to changing audio sample rates, are noise-resistant, and work well in a variety of settings. In Bangla speech emotion recognition, the suggested 1D DCNN-BLSTM model—which combines CNN, TDF, BLSTM—showed better recognition rates. To increase the size of the training and testing datasets, data augmentation techniques such as noise introduction, time stretching, time shifting, pitch shifting, speed changing, and combination augmentation were used. When MFCCs were utilized as training data for the DCNN-BLSTM model, the model performed better than in earlier research.

However, in order to progress SER for the Bangla language, two important issues must be resolved. First, as actors record the majority of databases now in existence, gathering databases of native speakers' recordings of Bangla speech emotions is crucial. Second, there are a lot of limitations on computing capacity, hardware expenses, and power consumption that small portable devices must contend with. Subsequent investigations ought to concentrate on tackling these obstacles by creating more profound neural networks and cutting down on hardware expenses and energy usage by effectively utilizing 1D CNNs. This strategy is anticipated to further the field of speech emotion recognition in Bangla by improving the generality and stability of SER models for small, portable devices.

# Reference:


[1] S. Monisha and S. Sultana, “A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages,” Advances in Human-Computer Interaction, vol. 2022, pp. 1–11, 12 2022.

[2] J. G. R´azuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, “Speech emotion recognition in emotional feedback for human-robot interaction,” International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015.

[3] M. U. Islam and B. M. Chaudhry, “Learnability assessment of speech based intelligent personal assistants by older adults,” in Human Aspects of IT for the Aged Population, Q. Gao and J. Zhou, Eds. Cham: Springer Nature Switzerland, 2023, pp. 321–347.

[4] Q. Luo and H. Tan, “Facial and speech recognition emotion in distance education system,” in The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007). IEEE, 2007, pp. 483–486.

[5] M. Lech and L. He, “Stress and emotion recognition using acoustic speech analysis,” in Mental Health Informatics. Springer, 2013, pp. 163–184.

[6] M. U. Islam and B. M. Chaudhry, “A framework to enhance user experience of older adults with speech-based intelligent personal assistants,” IEEE Access, vol. 11, pp. 16 683–16 699, 2023.

[7] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Speech Communication, vol. 116. 2020, doi: 10.1016/j.specom.2019.12.001.

[8] J. de Lope and M. Graña, “An ongoing review of speech emotion recognition,” Neurocomputing, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.

[9] Abbaschian, B.J.; Sosa, D.S.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensor 2021, 21, 1249. [CrossRef] [PubMed]

[10]"Bengali language," Britannica, 2023. [Online]. Available: https://www.britannica.com/topic/Bengali-language. [Accessed: 01 May-2023].

[11] Zhang, R.; Yin, Z.; Wu, Z.; Zhou, S. A novel automatic modulation classification method using attention mechanism and hybrid parallel neural network. Appl. Sci. 2021, 11, 1327. [CrossRef]

[12] Kulin, M.; Kazaz, T.; Poorter, E.D.; Moerman, I. A survey on machine learning-based performance improvement of wireless networks: PHY, MAC and network layer. Electronics 2021, 10, 318. [CrossRef].

[13] Mirmozaffari, M.; Yazdani, M.; Boskabadi, A.; Dolatsara, H.A.; Kabirifar, K.; Golilarz, N.A. A novel machine learning approach combined with optimization models for eco-efficiency evaluation. Appl. Sci. 2020, 10, 5210. [CrossRef]

[14] Noroznia, H.; Gandomkar, M.; Nikoukar, J.; Aranizadeh, A.; Mirmozaffari, M. A Novel Pipeline Age Evaluation: Considering Overall Condition Index and Neural Network Based on Measured Data. Mach. Learn. Knowl. Extr. 2023, 5, 252–268. [CrossRef]

[15] Han, X.; Chen, S.; Chen, M.; Yang, J. Radar specific emitter identification based on open-selective kernel residual network. Digit. Signal Process. 2023, 134, 103913. [CrossRef]

[16] Leng, Y.; Zhao, W.; Lin, C.; Sun, C.; Wang, R.; Yuan, Q.; Li, D. LDA-Based Data Augmentation Algorithm for Acoustic Scene Classification. Knowl. Based Syst. 2022, 195, 105600. [CrossRef]

[17] Jahangir, R.; Teh, Y.W.; Mujtaba, G.; Alroobaea, R.; Shaikh, Z.H.; Ali, I. Convolutional Neural Network-based Cross-corpus Speech Emotion Recognition with Data Augmentation and Features Fusion. Mach. Vis. Appl. 2022, 33, 41. [CrossRef]

[18] Chen, Z.Q.; Pan, S.T. Integration of Speech and Consecutive Facial Image for Emotion Recognition Based on Deep Learning. Master's Thesis, National University of Kaohsiung, Kaohsiung, Taiwan, 2021.

[19] Sajjad, M.; Kwon, S. Clustering Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access 2020, 8, 79861–79875.

[20] Zhao, J.; Mao, X.; Chen, L. Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks. Biomed. Signal Process Control 2019, 47, 312–323.

[21] Kakuba, S.; Poulose, A.; Han, D.S. Attention-Based Multi-Learning Approach for Speech Emotion Recognition with Dilated Convolution. IEEE Access 2022, 10, 122302–122313. [CrossRef]

[22] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid and M. S. Rahman, "Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks," in IEEE Access, vol. 10, pp. 564-578, 2022, doi: 10.1109/ACCESS.2021.3136251

[23] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang and C. Li, "Deep spectrum feature representations for speech emotion recognition", Proc. Joint Workshop 4th Workshop Affect. Social Multimedia Comput. 1st Multi-Modal Affect. Comput. Large-Scale Multimedia Data, pp. 27-33, 2018.

[24] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers and B. Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation" in arXiv:1802.05630, 2018.

[25] M. M. Rahman, D. R. Dipta and M. M. Hasan, "Dynamic time warping assisted SVM classifier for Bangla speech recognition", Proc. Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. (IC4ME2), pp. 1-6, Feb. 2018.

[26] A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network", Proc. Int. Conf. Platform Technol. Service (PlatCon), pp. 1-5, Feb. 2017.

[27] A. Satt, S. Rozenberg and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms", Proc. Interspeech, pp. 1089-1093, Aug. 2017.

# Performance Improvement of Speech Emotion Recognition in Bengali Language Using Deep Learning and BLSTM Networks

ORIGINALITY REPORT

| 10% | 7% | 4% | 4% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to Daffodil International University<br>Student Paper | 4% |
| 2 | Md. Riadul Islam, M. A. H. Akhand, Md Abdus Samad Kamal, Kou Yamada. "Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning", Electronics, 2022<br>Publication | 1% |
| 3 | univagora.ro<br>Internet Source | 1% |
| 4 | www2.mdpi.com<br>Internet Source | 1% |
| 5 | "Neural Information Processing", Springer Science and Business Media LLC, 2017<br>Publication | 1% |
| 6 | Ton Duc Thang University<br>Publication | 1% |
| 7 | era.ed.ac.uk<br>Internet Source | 1% |

| | | |
|---|---|---|
| 8 | www.mdpi.com<br>Internet Source | 1 % |
| 9 | www.scilit.net<br>Internet Source | 1 % |
| 10 | Shing-Tai Pan, Han-Jui Wu. "Performance Improvement of Speech Emotion Recognition Systems by Combining 1D CNN and LSTM with Data Augmentation", Electronics, 2023<br>Publication | <1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | On | | |