Islamic University of Technology (IUT)

Machine Learning Based Emotion Detection

Student Name & ID

1. Topy Khatun.....220032201

2. Kamal Hosen …....220032202

3. Md.Abusina.……220032203

4. Ikram Hossen.....…220032204

This thesis is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technical Education (BSc.TE) with specialization in Electrical and Electronics Engineering (EEE).

Department of Technical and Vocational Education (TVE)
Submission Date
October,2025

DECLARATION

This is to confirm that the work provided in this thesis is the result of analysis and experiments undertaken by Topy Khatun, Kamal Hosen, Md. Abusina, Ikram Hossen, under the supervision of Sheikh Munim Hussain, Lecturer, Department of Electrical and Electronic Engineering, Islamic University of Technology, Dhaka, Bangladesh. It also states the fact that no other place has presented this thesis or any portion of it in order to obtain any degree or diploma. The published and the unpublished work by other people has been taken into account in the text and a list of references has been provided.

Topy Khatun
Student ID: 220032201
Date: October 2025

Kamal Hosen
Student ID: 220032202
Date: October 2025

Md.Abusina
Student ID: 220032203
Date: October 2025

Ikram Hossen
Student ID: 220032204
Date: October 2025

Machine Learning Based Emotion Detection

Student Name & ID

1. Topy Khatun.....220032201

2. Kamal Hosen ….....220032202

3. Md.Abusina.……220032203

4. Ikram Hossen.....…220032204

Approval from supervisor

Sheikh Munim Hussain

Lecturer,

Department of Electrical and Electronic Engineering

Islamic University of Technology (IUT)

Acknowledgement

We are happy to offer our profoundest thanks to the Almighty Allah, who has given us the strength, patience, and perseverance to bring this thesis to a successful completion. It gives us the deepest sense of gratitude to our esteemed supervisor, Sheikh Munim Hussain, who has served as an invariable inspiration in this work of study through his guidance, encouragement, and useful suggestions. This thesis has been critical in his support.

We also want to thank the members of the professors and the personnel of the Department of Technical Education of the Islamic University of Technology deeply, and hope that their cooperation and support contributed to the success of the study. Our families and friends deserve some special attention as they supported, gave encouragement, and understanding throughout the research and the writing itself.

Finally, we would like to acknowledge all of you and all those who directly or indirectly assisted us in finishing this thesis to such a great success.

Abstract

The detection of emotion in speech has become relevant as social media becomes more popular and audio-based attacks become more frequent. This project aims to identify five emotions, including happiness, sadness, threat, panic, and neutrality, in the Bangla speech. The 1,106 five-second audio samples dataset was formed by using different sources and was also extended to 6,636 samples by inserting various types of noise. The MFCC, Melspectrogram, Spectral Centroid, and LPC features had been extracted and used by both machine learning and deep learning models.

Random Forest yielded the best results among conventional algorithms, with a 1D Convolutional Neural Network (CNN) performing better with the larger dataset, achieving 83 percent test accuracy and large ROC values. Gender-based testing was more accurate in female voices (88.07%) and male voices (81.2%). Further overfitting was suppressed and accuracy stabilized by the use of higher-order statistics, including the Teager Energy Operator.

Finally, the project was able to construct a Bangla speech sentiment database and create an effective CNN-based classifier. The findings prove to have potential uses in security, psychological evaluation, and social media surveillance.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

MFCC	Mel-Frequency Cepstral Coefficients
LPC	Linear Predictive Coding
AWGN	Additive White Gaussian Noise
LSTM	Long Short-Term Memory
BMO-DB	Berlin Emotional Speech Database
TEO	Teager Energy Operator
CNN	Convolutional Neural Network
RF	Random Forest
SVC	Support Vector Classifier
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic
KNN	K-Nearest Neighbor
ML	Machine Learning
RNN	Recurrent Neural Network
DL	Deep Learning

CHAPTER 1
INTRODUCTION

1.1 Background

Speech emotion recognition is one of the recently emerging ones most promising branches in the area of artificial intelligence and interaction between humans & computer. Today, speech is one of the most common mediums of interaction as social media, digital communication, and online interaction develop at light speed. As opposed to textual-based communication, a speech conveys not only a semantic message, but also emotional signs, which reveal the psychological state, mind, and intention of a speaker. Such emotional indicators can offer valuable insight It can be used for an extensive diversity of purposes, inclusive smart virtual assistants, medical assistance, customer care, learning, security, and entertainment.

Among the numerous uses of speech emotion detection, one of the most important is that it helps to recognize dangerous tones and panic cases. Because of the large-scale use of social media, verbal threats via audio channels have become frighteningly common. The threat of this trend has led to the attention of researchers in the creation of systems that can automatically identify emotions like anger, panic, or danger to prevent possible crimes, improve security, and ensure a safer environment during communication.

In the technological aspect, the speech recognition-emotion task involves the combination of several fields, such as digital signal processing, feature extraction, machine learning, and deep learning. Many features like Mel-Frequency Cepstral Coefficients (MFCC), chromagrams, spectral centroid, and spectral Rolloff have been extensively used over the years to encode emotional aspects of speech signals. As deep learning is developed, convolutional neural networks (CNNs) and other deep learning architectures have continued to improve performance on speech emotion detection tasks.

Therefore, the context in which this project is based is the convergence of increasing social needs and technological opportunities. Through the integration of contemporary methods of machine learning with knowledge of human speech, this study will offer a powerful model in identifying and categorizing emotions based on audio signals.

1.2 Motivation

The key driving force of the project is due to the growing applicability of emotion detection in practice. Emotions are an important part of human communication, and the fact that it might be possible to identify emotions in speech can give a more profound meaning than a sentiment analysis based only on texts.

The emergence of cyberbullying, online harassment, and threatening communication on digital platforms is one of the most significant issues in modern society. Threats that are based on audio are especially worrisome, since they are spontaneous and emotional in nature. Real-time detection of these threats can play an important role in law enforcement, social media surveillance, and safety systems.

Besides security, speech emotion recognition could also be important in healthcare and psychological assessment. Emotions such as sadness, panic, or stress are frequent factors that can be taken as early warning signs of mental health problems. A system, able to predict these emotional states reliably can assist psychologists and healthcare providers to provide interventions in time. Moreover, emotion-aware systems are also becoming a part of intelligent virtual assistants, customer care applications, and educational platforms, where the knowledge of the emotions of the user can significantly enhance the quality of the interaction.

Thus, the social and technological drive of this project is to overcome the acute necessity to identify dangerous and critical emotions in speech, as well as to make a contribution to the development of technology aware of emotions.

1.3 Objective

This research has the following objectives:

1. Threat Analysis - This is to analyze speech data to detect threatening tones and to offer a viable solution to the security and law enforcement applications.
2. Sentiment Detection- To design and realize a deep learning & machine learning algorithm that is capable of classifying emotion correctly and recognizing happiness, sadness, panic, threat, and neutral.
3. Psychological Assessment - To illustrate the potential of using speech emotion detection to track the presence of mental illness by detecting the emotions associated with stress, depression, or panic.
4. Dataset Development - In order to prepare and annotate a Bangla speech dataset with various emotion types and thus leave a valuable resource in later research, one will have to undertake a dataset development.
5. Performance Analysis - To compare and contrast the performance encompasses both conventional machine learning methods and deep learning, as well as the advantages and disadvantages of every strategy.

With the accomplishment of these goals, this project will make contributions to the study in the Speech Emotion Recognition Region, both theoretical and practical.

CHAPTER 2
LITERATURE REVIEW & METHODOLOGY

2.1 Literature Review

2.1.1 Theoretical Background

Research in the multidisciplinary topic of speech emotion recognition (SER) principles of which are based on the foundations of speech processing, psychology, AI and machine learning. The gist of the matter is that speech also entails not only linguistic information but also paralinguistic features, as they also disclose the emotional status of a speaker. Different studies have indicated that the elements of pitch, energy, formants, and spectral properties have a great correlation with human emotions[1].

Emotion recognition theories based on speech are directly related to the human sense of hearing and signal analysis. Conventional methods were very dependent upon hand-crafted characteristics of Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and chromograms that tried to reflect speech characteristics in the frequency and temporal domains. With the advent of machine learning, scholars started employing machine learning classifiers such as Support Vector Machines (SVM), Decision Trees, and Random Forests to infer these features into emotion categories.

The past few years have seen an increase in performance due to the use of deep learning, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) that can automatically discover discriminative features on raw or lightly processed speech signals[2]. This kind of model is more accurate and stronger, but normally requires large datasets to be trained successfully.

2.1.2 Previous Research

Initial research in the 2000s was conducted on English and other common languages and generated benchmark datasets, including EMO-DB (Berlin Emotional Speech Database).

Nevertheless, Relatively little study has been done on speech emotion recognition of Bengali (Bangla).[3] investigated emotion recognition of Bengali speech based on the RNN-based modulation categorization, and[1] presented a method of entropy-based recognition based on the improved wavelet coefficients. The two articles showed that SER was feasible in Bangla but cited limitations like insufficient datasets and changes in speech quality.

Other, more recent literature has proposed noise augmentation methods and higher-order statistics to enhance resilience. As an example, spectral features, when applied with the Teager Energy Operator (TEO),[4] have been realized to provide better results in accuracy by boosting the patterns of energy that are pertinent to emotions.

2.1.3 Gaps in Literature

The review of literature indicates that there are a number of gaps:

- There is a dearth of annotated Bangla emotional speech datasets of high quality.
- Excessive dependence on small-scale experimentation, which limits generalization.
- Absence of methodical contrast of old machine learning frameworks and deep learning systems in Bangla SER.
- Lack of focus on gender-specific variations in how emotions are recognized and expressed.

The gaps in this project are filled by building a big Bangla dataset, performing noise augmentation, experimenting with machine learning and deep learning, and comparing the outcomes with the gender-based subdivisions.

2.2 Methodology

2.2.1 Dataset Preparation

The data sample in this project is 1106 original speech samples that are 5 seconds in length and sampled at 44.1 kHz. The audio files were in mono mode and in .wav format. The samples of speech were taken at various sources including Bangla series, movies, dramas, talk shows, interviews, news broadcasting, and real-life conversations.

In order to test the reliability, 72 volunteers were involved in annotating the dataset into five emotion categories, including happiness, sadness, threat, panic, and neutral. This dataset was then separated into training (80 %), validation (10 %), and testing sets (10 %). In order to address the limitation of datasets, noise augmentation was used by adding five noise types (bird, cricket, train, wind, and AWGN). This added to the sample size to 6,653 samples, which made the models more robust in real-life scenarios.

2.2.2 Feature Extraction

Extracting features is a very critical aspect in SER. The following features were obtained from audio signals in this project:

- The frequency spectrum's center of gravity is indicated by the term spectral centroid.
- Spectral Rolloff - This shows how frequently at which a given spectral energy amount is enclosed.
- Chromagram- the strength of 12 semitone pitch classes.
- MFCC (Mel-Frequency Cepstral Coefficients)- very popular in speech analysis to represent perceptually relevant features.
- Mel Spectrogram - a time-frequency analysis that is consistent with human auditory perception.
- LPC (Linear Predictive Coding)- approximates the spectral envelope of speech signals.

Also, spectral centroid and Rolloff features were subjected to the Teager Energy Operator (TEO) in order to optimize the energy distribution, which marginally increased the classification accuracy.

2.2.3 Algorithms Used

Deep learning combined with conventional machine learning were both applied:

Machine Learning Models:

- Support Vector Classifier (SVC)
- K-Nearest Neighbors (KNN)
- Decision Tree Classifier
- Random Forest Classifier
- AdaBoost Classifier

Among the Random Forest Classifier's performance on the original dataset.

Deep Learning Model:

Convolutional Neural Network (CNN): A 1D CNN was used, using dropout regularization to minimize overfitting. Noise-augmented data gave CNN a higher accuracy of 83.41% and ROC values that were closer to 1.

2.2.4 Strengths and Weaknesses of the Approach

Strengths:

- Constructed one of the initial large-scale Bangla emotion speech datasets.
- Applied classic ML and contemporary DL techniques, which allow them to be compared.
- Applied noise augmentation, which enhanced generalization and strength.
- Gender-specific testing gave a more detailed understanding of the variability of the datasets.

Weaknesses:

- A small number of classes of emotions (just five), which is a weak model of expressiveness.
- CNN also had the problem of overfitting when it was trained on smaller datasets.
- The quality of the data set was influenced by the real-life recording, as opposed to the studio datasets.
- The size of the dataset did not allow the test of recurrent architectures (RNN, LSTM).

2.3 Hypothesis

According to the literature and research approach, the study's hypothesis is:

The application of deep learning models, especially CNNs with noise-augmented Bangla datasets and improved higher-order statistical features, will be more effective compared to conventional techniques for machine learning in the recognition of speech emotions with higher accuracy and strong robustness in practical scenarios.

CHAPTER 3
MAIN WORK

This chapter contains, the key practical work of the research (The suggested Speech Emotion Detection system's design, implementation, testing, and simulation) is described. It also gives a detailed discussion of the theoretical analysis and the experimental results associated with it. This chapter has a goal of demonstrating how the model was constructed, trained, and assessed in order to detect emotions in Bangla speech correctly.

3.1 Design

This project design was created through the literature review results and hypothesis. The overall workflow model of the proposed model is depicted below:

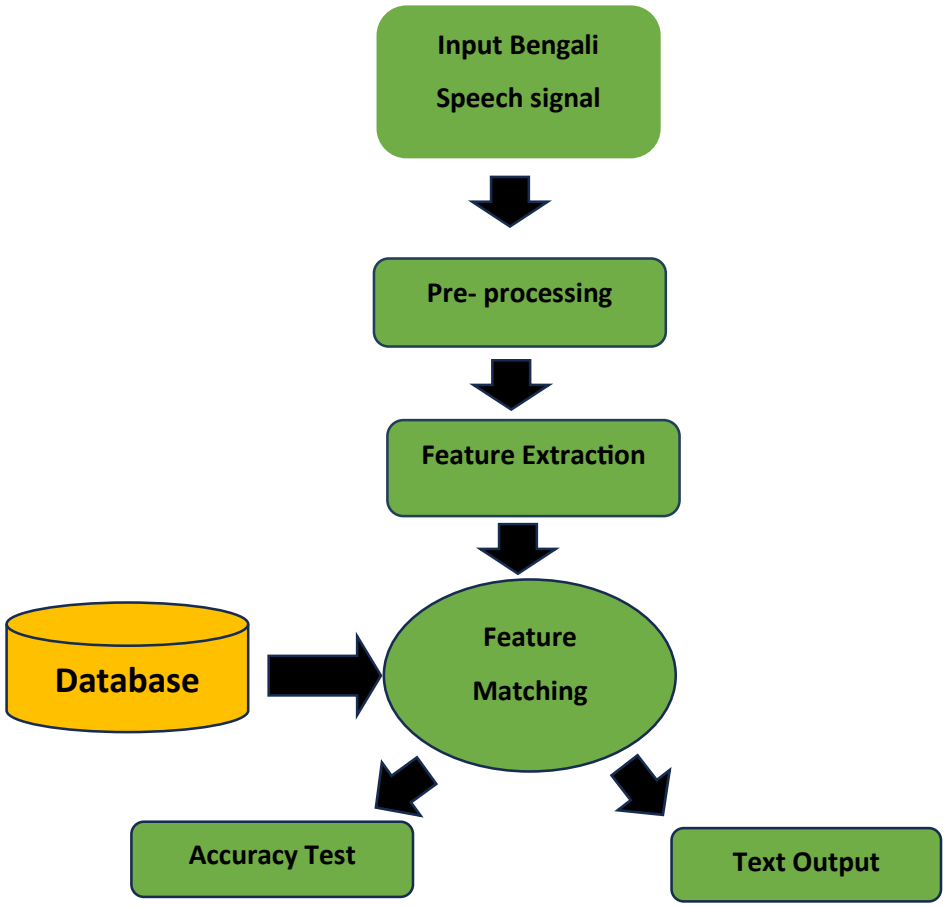


Figure 1: Workflow diagram of proposed system

The system was broken down into a number of stages:

1. **Collection and Annotation of Data:**

   Only 1,106 samples of the Bangla speech were taken, originating in different natural media, including TV dramas, movies, news, talk shows, and real-life conversations. The length of each sample was 5 seconds and was recorded at 44.1 kHz and in mono in the .wav format. The volunteer group members and 72 other volunteers manually annotated the data into five emotion categories (Happiness, Sadness, Threat, Panic, and Neutral).

2. **Data Splitting**

   The data set was separated into three sections:

   - Training: 80%
   - Validation: 10%
   - Testing: 10%

3. **Noise Augmentation**

   To enhance the strength of the dataset, they added five classes of background noises: Bird, Cricket, Train, Wind, and Additive White Gaussian Noise. This made the data set larger (6,653 samples) and provided better diversity and stability during model training.

4. **Feature Extraction**

   The features extracted were made to represent both the frequency and the temporal features of speech.

   - Spectral Centroid
   - Spectral Rolloff
   - Chromagram
   - MFCC (Mel-Frequency Cepstral Coefficients)
   - Mel Spectrogram
   - Linear Predictive Coding (LPC)

Subsequently, these characteristics were increased through Teager Energy Operator (TEO) to increase the power of significant spectral components.

5. **Algorithm Design**

Emotion classification was done with both Machine Learning (ML) and Deep Learning (DL) algorithms.

- ML: Support Vector Classifier (SVC), KNN, Decision Tree, Random Forest, and AdaBoost.
- DL: A Convolutional Neural Network (CNN) of 1D audio sequences was created with dropout regularization.

3.2 Implementation

It was implemented with Python and the TensorFlow, Keras, Librosa, and Scikit-learn libraries. The general implementation procedures were:

Step 1: Preprocessing and loading Audio.

All the wav files were normalized, trimmed, and resampled to 44.1 kHz. There were fewer parts in the silence that were eliminated.

Step 2: Extraction of features with the help of Librosa.

Each 5-second sample was calculated to give MFCC, Chromagram, and LPC features. The spectral centroid, as well as the Rolloff features, were subjected to the TEO operator.

Step 3: Model Training

Scikit-learn was used to train machine learning models, and Keras was used in order to put deep learning models into practice. Performance was better with the Random Forest Classifier of the ML techniques.

When it comes to deep learning, a 1D CNN was used, and the layers were as follows:

- Combination of Convolutional and ReLU layers
- Pooling Layer
- Dropout Layer (to stop overfitting)
- Fully Connected Dense Layer
- SoftMax Output Layer

Step 4: Optimization, Regularization.

The loss was Adam optimizer and Categorical Cross-Entropy. To minimize overfitting and enhance generalization, dropout and batch normalization were proposed. The model architecture of CNN is based on the architecture of a standard computer. The CNN model architecture is a standard computer architecture.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 160, 128)          768
_________________________________________________________________
activation (Activation)      (None, 160, 128)          0
_________________________________________________________________
dropout (Dropout)            (None, 160, 128)          0
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 20, 128)           0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 20, 128)           82048
_________________________________________________________________
activation_1 (Activation)    (None, 20, 128)           0
_________________________________________________________________
dropout_1 (Dropout)          (None, 20, 128)           0
_________________________________________________________________
flatten (Flatten)            (None, 2560)              0
_________________________________________________________________
dense (Dense)                (None, 5)                 12805
_________________________________________________________________
activation_2 (Activation)    (None, 5)                 0
=================================================================
Total params: 95,621
Trainable params: 95,621
Non-trainable params: 0
```

Figure 2:1D convolutional approach

3.3 Experimentation

Various experiments had been carried out in order to evaluate the efficiency of various models in various conditions.

3.3.1 Experience-1 Base Dataset (1106 sample)

We used a base dataset consisting of 1106 samples. Only the original dataset (1106 samples) was used in the first experiment.

Results showed:

- CNN Validation Accuracy: ~40–44%
- Large overfitting and noisy validation accuracy.
- At this stage, Random Forest did better than other ML models.

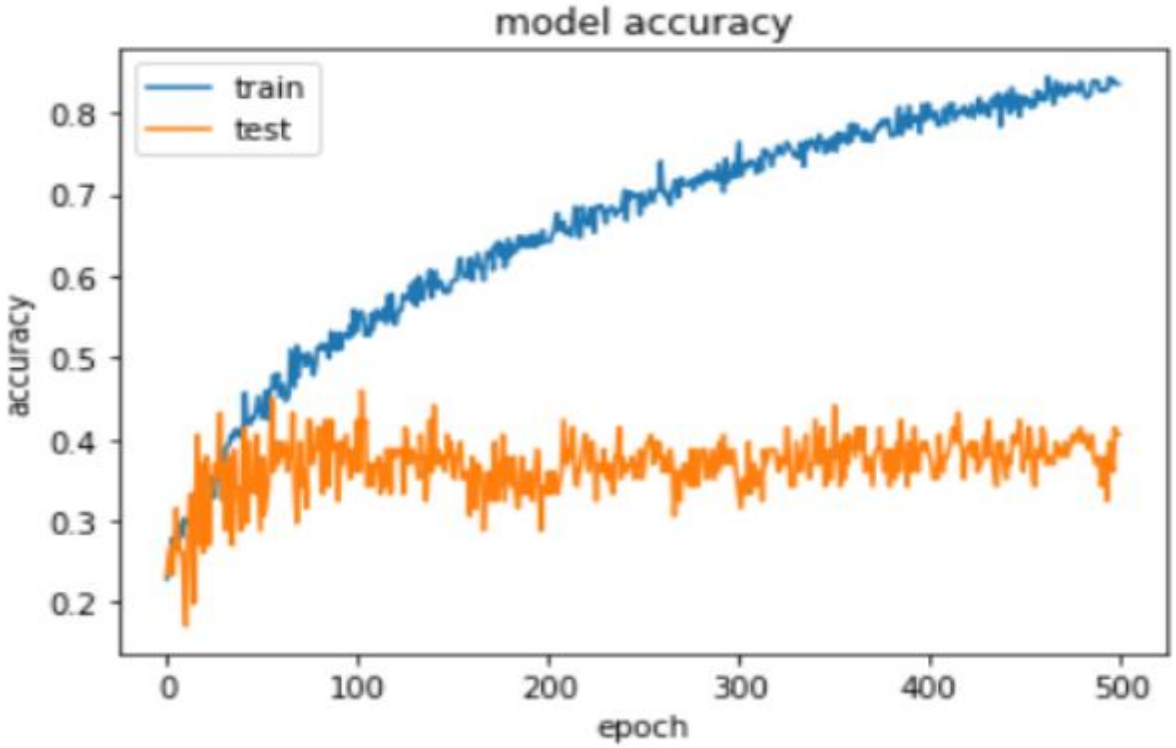


Figure 3: CNN performance on 1106 samples

3.3.2 Experiment 2 – Noise-Augmented Dataset (6636 Samples)

Once five forms of noise were included, the dataset size was 6636.

In the CNN model, the improvement was remarkable:

- There was a consistent improvement in accuracy during training and validation.
- Much less overfitting than in the first experiment.
- Test Accuracy was 83% and the ROC curve area was close to 1.0.

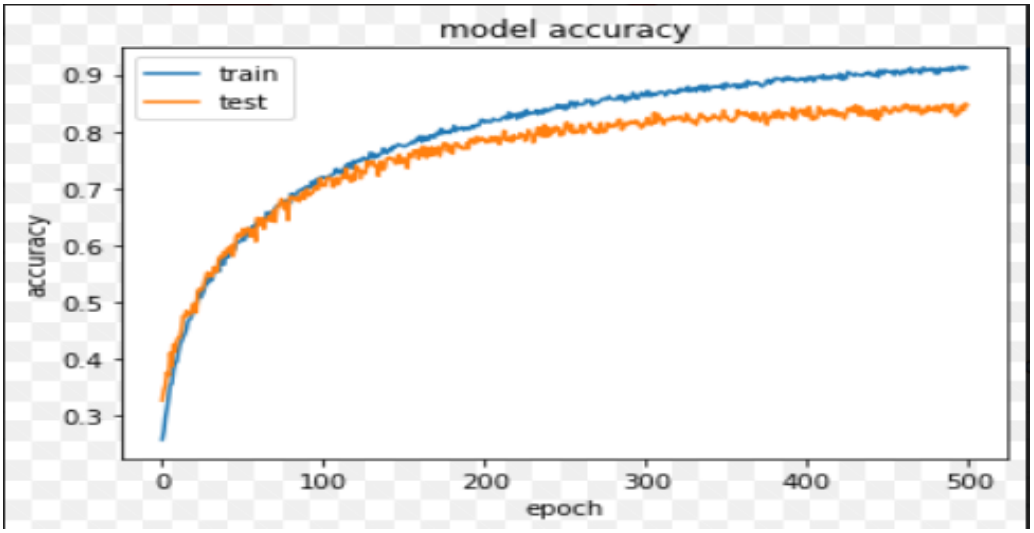


Figure 4: Performance on improving the data set(CNN)t

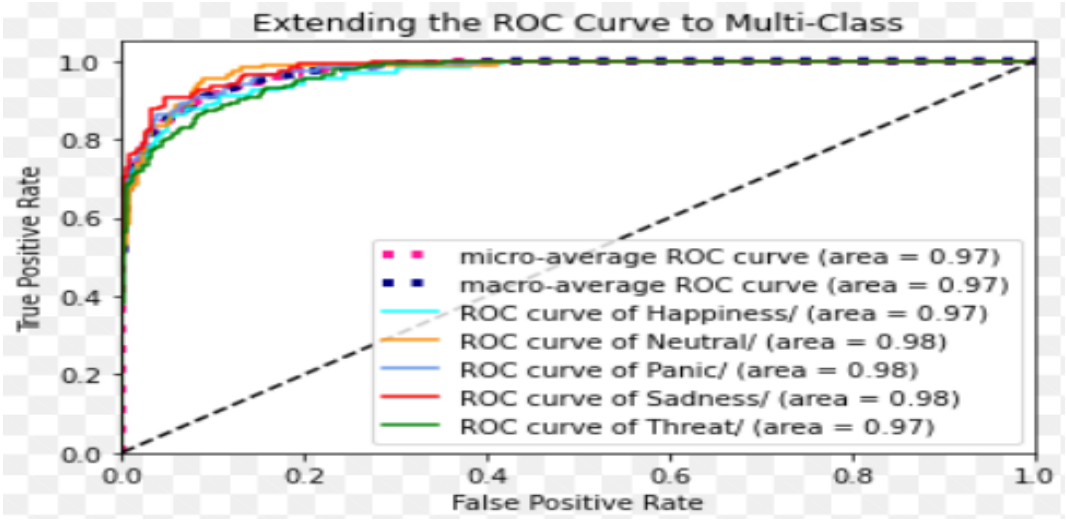


Figure 5: The ROC curve for all classes

3.3.3 Experiment 3 – Gender-Based Evaluation

To examine the gender effects on emotion recognition, the data were divided into male and female data**.**

- Male voices: 4194 samples (63%)
- Female voices: 2442 samples (37%)

Model results:

- CNN Accuracy for Female data: 88.07%
- CNN Accuracy for Male data: 81.2%

This experiment demonstrated that female voice data presented higher recognition performance, which might be because of more distinct pitch differences.

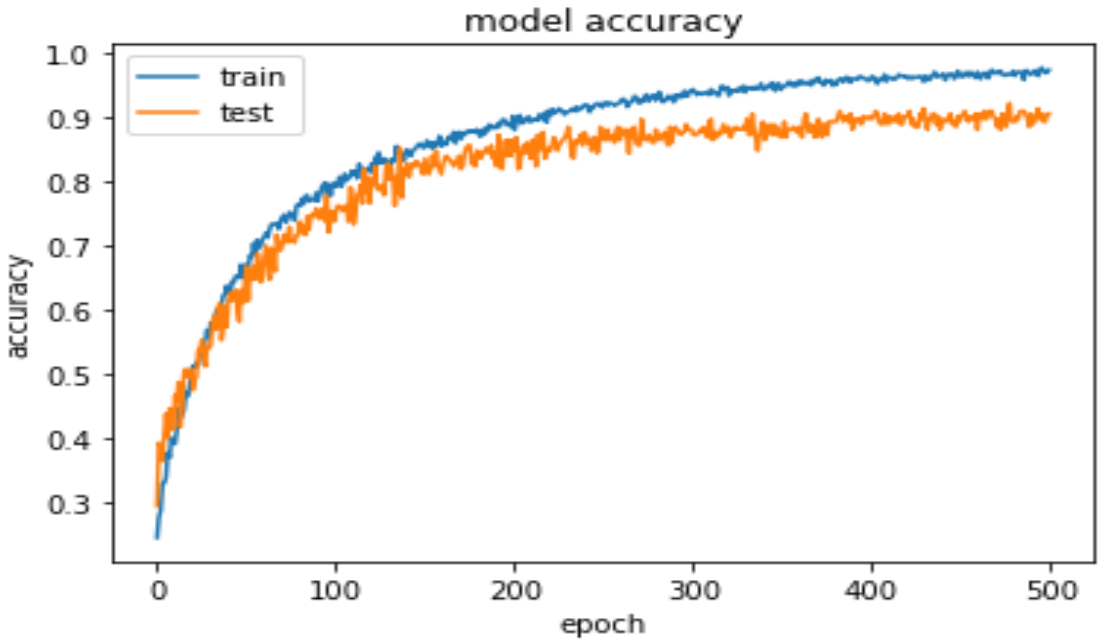


Figure 6: Accuracy on Female data with CNN architecture-88.07%

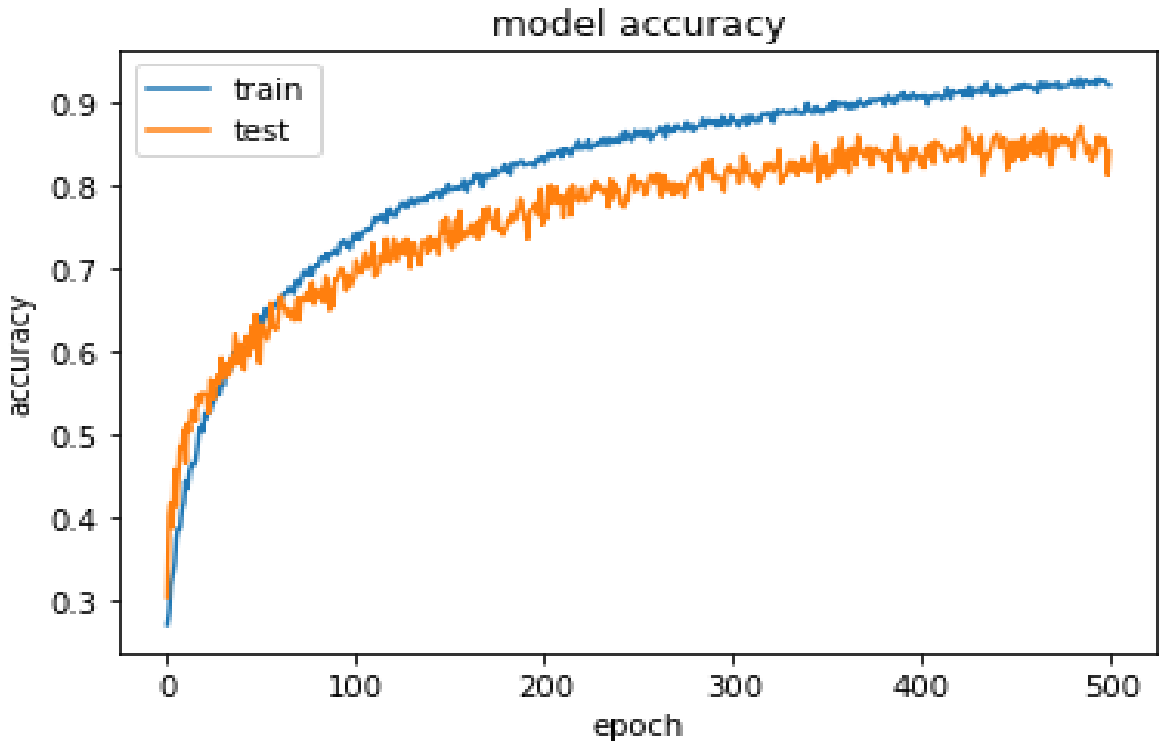


Figure 7:Accuracy on male data with CNN architecture-81.2%

3.3.4 Experiment 4 – Effect of Higher Order Statistics

In order to enhance the spectral feature energy, the Teager Energy Operator (TEO) was used on the spectral centroid and spectral rolloff.

The TEO formula is:

$$TE(x[n]) = x[n]^2 - x[n-1].x[n+1],$$

This procedure enriched some of the important emotional signals in the speech signal. Test accuracy after using TEO was a little higher (≈82%), and there was less overfitting. The optimized version of the given model (TEO + CNN without LPC) obtained 83.41% accuracy.

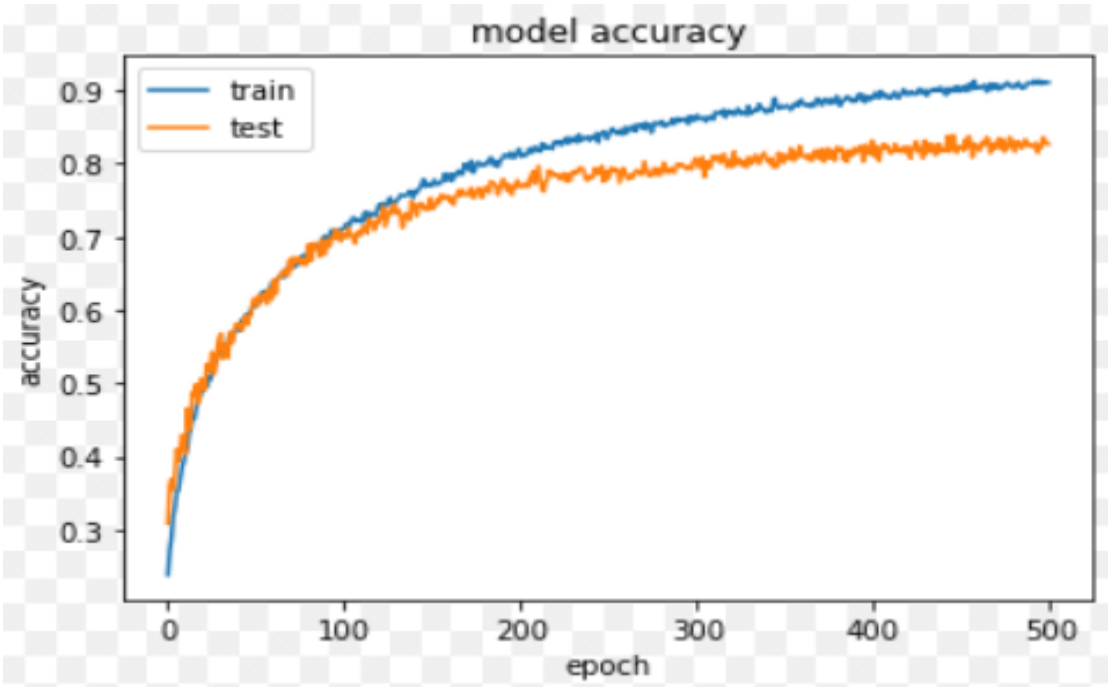


Figure 8: Performance comparison with TEO+LPC

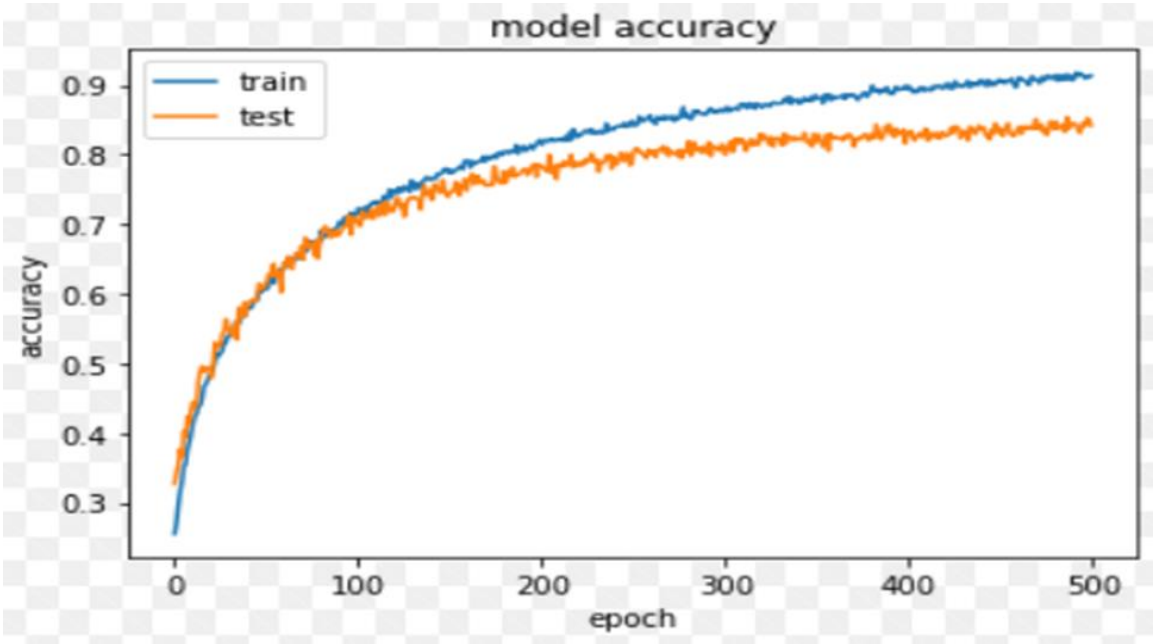


Figure 9:: Performance comparison without LPC

3.4 Simulation and Analysis

Training and validation accuracy, both during simulation, were plotted, in a series of epochs. The CNN model exhibited smooth convergence with a small overfitting of the model towards later epochs. The ROC curve area of each emotion type was close to 1.0, indicating a good fit in emotion recognition. These simulation findings supported the reality that exists proposed device can actually detecting emotions in Bengali speech in noisy conditions.

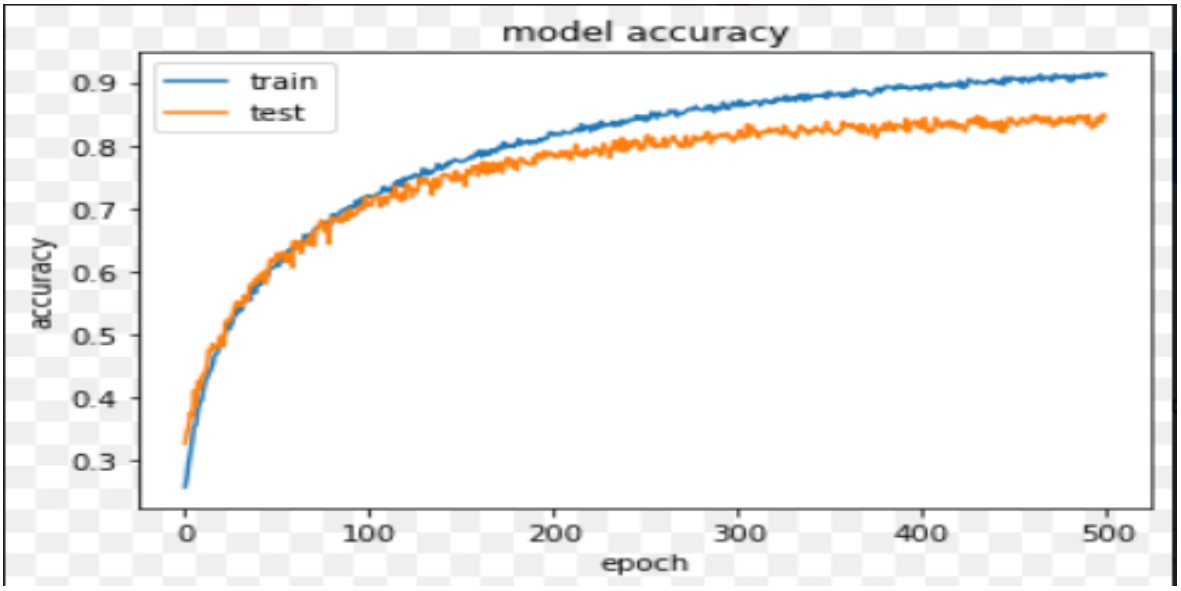


Figure 10: Training and Validation Accuracy vs. Epochs

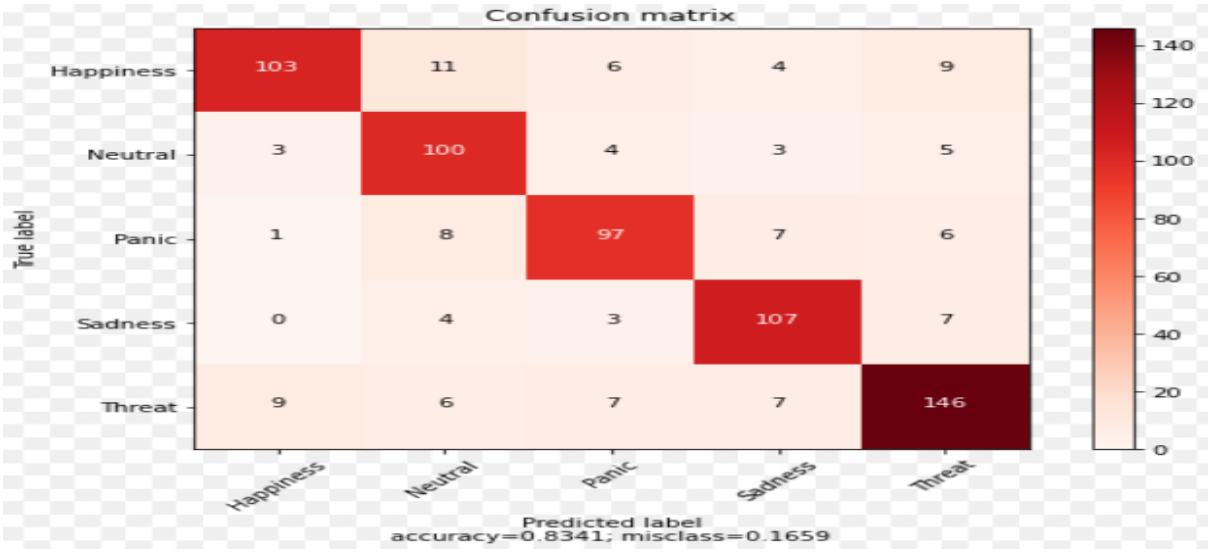


Figure 11: Confusion Matrix for Final CNN Model

3.5 Explanation

Using the experiments and simulations, it is evident that deep learning models, in particular CNN, are more effective than classic ML classifiers in cases where there is enough data used, as well as other augmentation methods. The approach's resilience and generalizability were enhanced by the addition of noise and higher-order feature operations (TEO). Nevertheless, it still had certain drawbacks, such as fewer classes of emotions and inconsistent audio quality that compromised the accuracy.

Irrespective of these, the ultimate CNN model was successful in reaching a test accuracy of 83.41, which confirms that the proposed technique is useful in Bangla Speech Emotion Recognition.

3.6 Summary

The design, implementation, experimentation, and simulation of the proposed Speech Emotion Detection system were described in this chapter.

Practical use of theoretical concepts of earlier studies created a strong CNN-based model. The hypothesis that deep learning techniques with data augmentation can perform better than traditional classifiers in recognizing emotions using Bangla Speech was proven correct by the experimental results.

CHAPTER 4
RESULTS & DISCUSSIONS

This section is a report among the experimental findings of the research work and discusses them critically with references to the aims set above. The results are analyzed in terms of the data set, algorithms, feature extraction methods, and performance differences in various experimental conditions. Tables and figures are given to explain some of the important observations, and results are juxtaposed with available literature to substantiate their relevance.

4.1 Dataset and Baseline Performance

This research started with the development of a set of data that could be used to sufficiently cover the five chosen categories of emotions: happiness, sadness, threat, panic, and neutral. The collected audio samples amounted to 1,106 samples, about 220 samples per class, providing a class distribution that was close to balanced. The audio clips were specifically chosen to represent a variety of natural speaking situations. The samples were 5 seconds each, in .wav format, and were recorded in mono format at a common sampling rate of 44.1 kHz, commonly used in speech processing[5].

The sound sources were carefully varied and included Bengali films, dramas, talk shows, interviews, news programs, and real-life conversations. This variety was predetermined to represent real-world emotion recognition tasks with widely varied speakers, recording conditions, and background conditions. Reliability was achieved by annotating the dataset with 72 volunteers, including the project members. Several annotators were used to reduce labeling bias and to make sure that the perceived emotional content was related to the desired class. The data set was then divided into training (80%), validation (10%), and test (10%) sets, which is customary in a machine learning experiment.

4.1.1 Baseline Performance with Traditional Machine Learning Models

The dataset was turned on the traditional machine learning algorithms, including Decision Tree (DT), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), AdaBoost, & Random Forest (RF) after feature extraction (MFCC, LPC, spectral centroid, spectral roll-off, chromagram, & mel-spectrograms).

The findings showed that the Random Forest always was the most accurate of all the algorithms tested. This is possibly explained by the nature of Random Forest, which put together decisions of many decision trees to minimize the variance & enhance predictive consistency. Random Forest does not overfit to noise and feature redundancy as much as single-tree models do[6]. Also, since Random Forest does its own internal feature selection at training time, it is adapted to high-dimensional data like speech signals, where correlated features (e.g., spectral centroid and spectral roll-off) would otherwise hurt performance.

In comparison, Support Vector Classifier suffered an adequate level of performance because of the comparatively tiny amount of its input dataset and limited number features. Although SVCs typically work well in high-dimensional feature space, they are highly dependent on both the type of kernel and regularization parameters, which may be difficult to optimize when using limited training data[7]. On the same note, the Decision Trees were overfitting, meaning that they segmented the training data so finely and failed to perform well when applied to the validation and test data.

4.1.2 Comparison with Existing Literature

The improved output of random forest on original data corresponds to the general results of the speech emotion recognition. Indicatively, using Random Forest, as opposed to other classifiers like Naïve Bayes and KNN for emotion recognition tasks on small and noisy data sets, has been established to be superior[8]. Similarly,[9] have found that ensemble approaches are especially effective in solving problems where there is variability in the recording conditions and speaker features in the dataset.

Additionally, low-resource language speech datasets such as Bengali usually have the disadvantage of limited purposefully labeled data available, so dense deep learning architectures have a high likelihood of overfitting. Under these circumstances, machine learning methods based on ensembles, such as Random Forest, can be considered as a resilient and robust baseline[10]. This justifies the application of the random forest to the initial step in this study and provides a reference point against which deep learning models will be compared later.

4.1.3 Critical Reflection on Dataset and Baseline Limitations

Although the design of the dataset provided coverage of a wide range of contexts, a few limitations should be mentioned:

Size Constraint: Since the dataset consists of approximately 220 samples per class, it is quite small in comparison to other popular speech databases such as RAVDESS or Emo-DB, which can have thousands of samples.

Natural Variability: Since audio was recorded based on actual and media sources, the quality of audio was also different. Background noise, overlapping conversation, and imprecise recorders are also possible sources of variability that influenced the performance of the models.

Emotion Subtlety: There are more difficult categories, and the most difficult of these were threat and panic, because the acoustic cues associated with these categories were more similar, and thus could be confused by both human annotators and machine models.

Nonetheless, the dataset had a special contribution as it was considering the Bangla speech emotion recognition, which is a comparatively understudied field. It was also a realistic test bench based on non-ideal conditions, and it represents the real world where emotion recognition systems will probably be implemented.

4.2 Convolutional Neural Network (CNN) Performance on Original Dataset

Convolutional Neural Networks (CNNs) have considered among the greatest efficient algorithms for deep machine learning to execute audio classification tasks such as speech emotion recognition (SER). Their advantage is that they can automatically learn hierarchical

representations of spectral and temporal features of raw or pre-processed speech signals, with little or no hand-crafted feature engineering. In this experiment, a 1D CNN architecture was trained using the initial data of 1,106 samples to evaluate its effectiveness in the recognition of emotional behavior in Bangla speech.

Although the CNNs promise high performance, the performance on the original dataset was poor, having validation accuracy that stalled at about 40 percent and test accuracy of just 44 percent. This result means that deep learning on small and heterogeneous data is difficult to apply. The subsections that follow examine the performance in more detail.

4.2.1 CNN Architecture and Training Setup

The applied CNN involved:

Input layer: sequential-frame speech features (MFCC, LPC, spectral features) are represented as input features.

Convolutional layers: 1D kernel used to represent local time-dependent relationalities.

Layers: max-pooling in order to make the dimensions smaller and be able to retain the largest features.

Layers with full connections: to combine high-level trained representations.

SoftMax output layer: to classify into five emotions.

The model was trained with the Adam optimizer on the categorical cross-entropy loss. Early stopping and 50 epochs of training were conducted to ensure unnecessary overfitting.

4.2.2 Training and Validation Performance

The training and validation curves showed that instability and divergence occurred within 15 to 20 epochs. Training accuracy steadily rose, whereas validation accuracy rose and peaked at about 40 percent. This overfitting is reflected in the discrepancy between observed and model values because the model is memorizing training data, but not predicting unobserved samples.

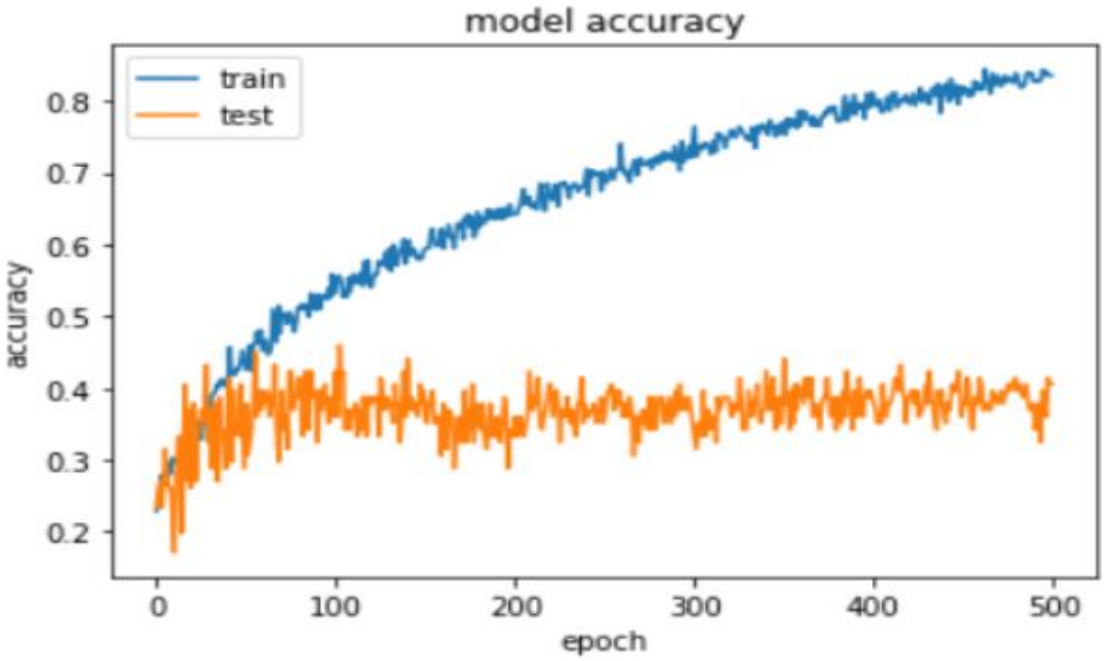


Figure 12: Train and validation accuracy

Table 1: CNN Performance on Original Dataset

Matrix	**Value**
Training Accuracy (final)	90%
Validation Accuracy	40%
Test Accuracy	44%
ROC-AUC (average)	0.70

4.2.3 Factors Affecting CNN Performance

A number of factors led to low performance:

Dataset Size Constraint:

The CNN did not have enough training examples with its own features and enough samples per class to learn strong hierarchies of features. Deep networks have the advantage of needing large databases to prevent overfitting, which is frequently mentioned in SER studies[11].

High Intra-Class Variation:

The data set featured speech of different origins (movies, interviews, and real-life conversations) and therefore had different recording conditions. Therefore, there was extensive acoustic variability within the same emotion category (e.g., panic), which also increases the difficulty of CNN filters to be generalized.

Unbalanced Acoustical Properties:

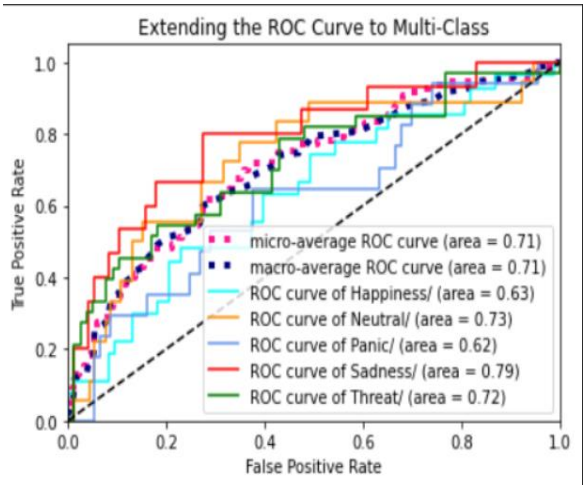
There is a significant difference in pitch, timbre, and spectral distribution between male and female voices[10] The difference could have caused bias, which decreases classification performance without stratified normalization.

Regularization Deficiency:

Early CNN training used dropout layers and batch normalization with limited usage. Little regularization resulted in the model rapidly learning the training set, as indicated by the steep rise in training accuracy relative to validation performance.

4.2.4 ROC Curve and Class-wise Analysis

The five classes of emotion formed a Receiver Operating Characteristic (ROC) curve. Although the area under the curve (AUC) was reasonably good at around 0.70, some classes, like neutral, were more well distinguished than the others, like panic and threat, where there was considerable overlap in the probability distributions.



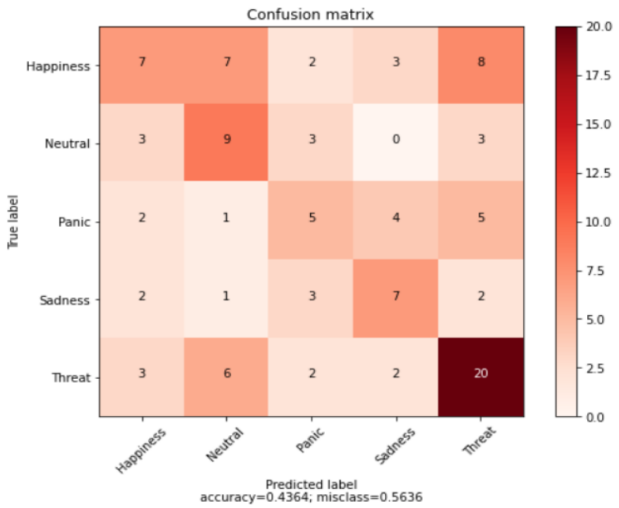


Figure 13: ROC curves for five emotion classes (CNN original Dataset)

Table 2: Class-wise Accuracy (CNN on Original Dataset)

Emotion	Accuracy (%)
Happiness	48.2
Sadness	46.7
Neutral	52.1
Panic	39.5
Threat	34.8

This asymmetry demonstrates the challenge of categorizing emotions with a small acoustic difference, especially when there are limitations on data.

4.2.5 Comparison with Traditional Models

compared to the Random Forest baseline, which performed better in generalization, the poor performance of CNN supports the argument that deep models are not universally better. In cases where:

The data is small (as it is here). Two powerful discriminatory features (MFCC, LPC) are already available with feature engineering. The resources of computation are scarce.[12] report similar findings and state that in the case of under-resourced languages and low-volume corpora, the Deep learning and conventional ML models are on an equal footing.

4.2.6 Critical Reflection

The results highlight the CNN's reliance on the scale and quality of the datasets. Although CNNs learn features automatically, they are weaker when presented with small, noisy, or non-homogeneous data. This is why convergence and plateauing validation accuracy are unstable.

Nevertheless, those findings would be a valuable starting point: they prove that CNNs need augmentation measures, be it the expansion of the dataset, introduction of noise, or advanced regularization, to be able to reach their full potential. These are discussed in Section 4.3, where data augmentation resulted in a substantial increase in accuracy.

4.3 Data Augmentation with Noise

A major issue with utilizing deep learning for speech emotion recognition (SER) is the absence of substantial, labeled datasets, especially while utilizing specific languages with limited resources like Bangla. Neural networks need a lot of training data in order to acquire useful feature representations and prevent overfitting. To counter this, data augmentation techniques were used in this work to artificially increase the data size to enhance the generalization of the model to previously unseen speech.

4.3.1 Noise Types and Dataset Expansion

The initial data consisted of 1,106 audio samples and five emotion categories. Five forms of environmental noise were introduced to augment the volume of data and its variety.

- Bird sounds
- Cricket sounds
- Train background noise
- Wind noise
- AWGN (Additive White Gaussian Noise)

These noises were mixed with each original audio sample at different levels, thus forming five new versions of the audio sample. This growth raised the number of samples to 6,636, and each emotion category was equally represented.

This additive noise augmentation method has been extensively tested in speech recognition studies[13] because it does not only increase the size of the dataset, but also adds variability that is present in the real world, such as background noise that makes models more robust to external perturbations, often present in natural settings (social media recordings, phone calls, live interviews, etc.).

4.3.2 CNN Performance on Augmented Dataset

The augmented data was used to retrain the Convolutional Neural Network (CNN). This input augmentation effect was immediately seen in the training and validation curves, which converged more smoothly, with much less overfitting than training on original dataset.

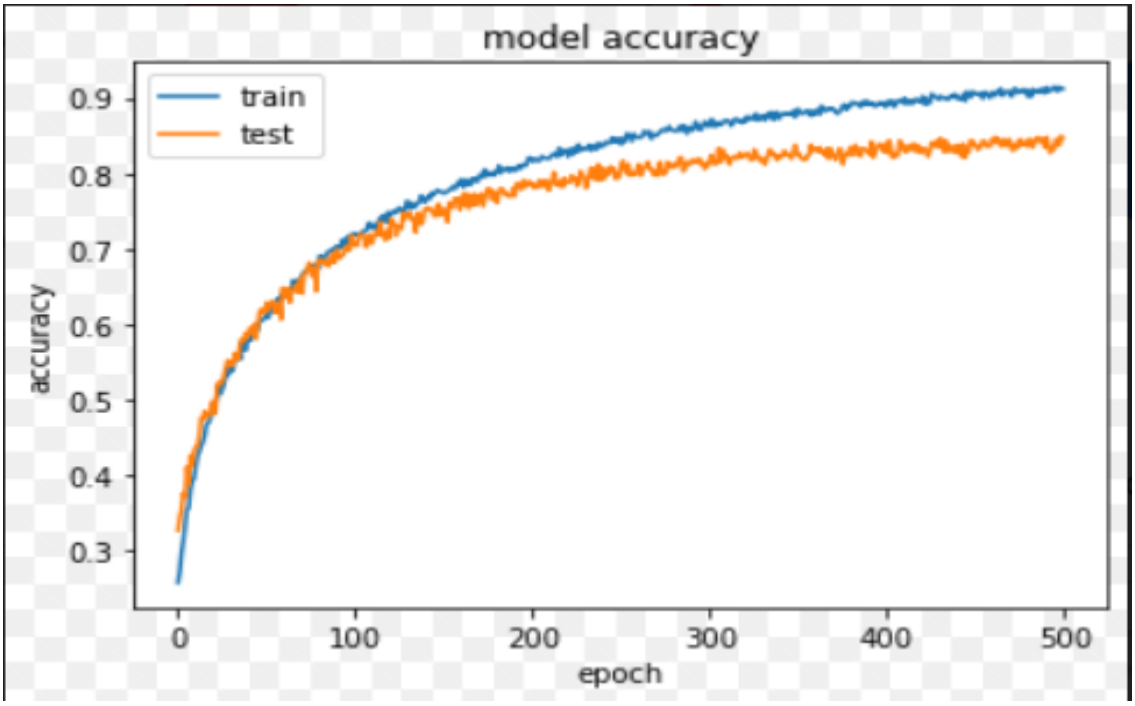


Figure 14: Training and validation accuracy curve (CNN on augmented data set)

Key results include:

Table 3: CNN performance on augmented dataset (6636 samples)

Metric	Value
Training Accuracy (Final)	90–95%
Validation Accuracy	83%
Test Accuracy	83%
Roc-Auc (Average)	0.95–0.99
Overfitting Status	Minimal

The accuracy of the test increased by a factor of nearly 2 (i.e., 83 percent) compared to before (44 percent). Moreover, most emotion classes were close to 1, which means that the ROC-AUC values are very reliable when it comes to classifying data by category.

4.3.3 ROC Curves and Class-wise Performance

Plots the ROC curves of the augmented dataset, and it can be seen that the separations between the different emotions were close to perfect. The augmented model displayed much better class-wise precision and recall, including the previously challenging classes (panic and threat), than the original dataset.

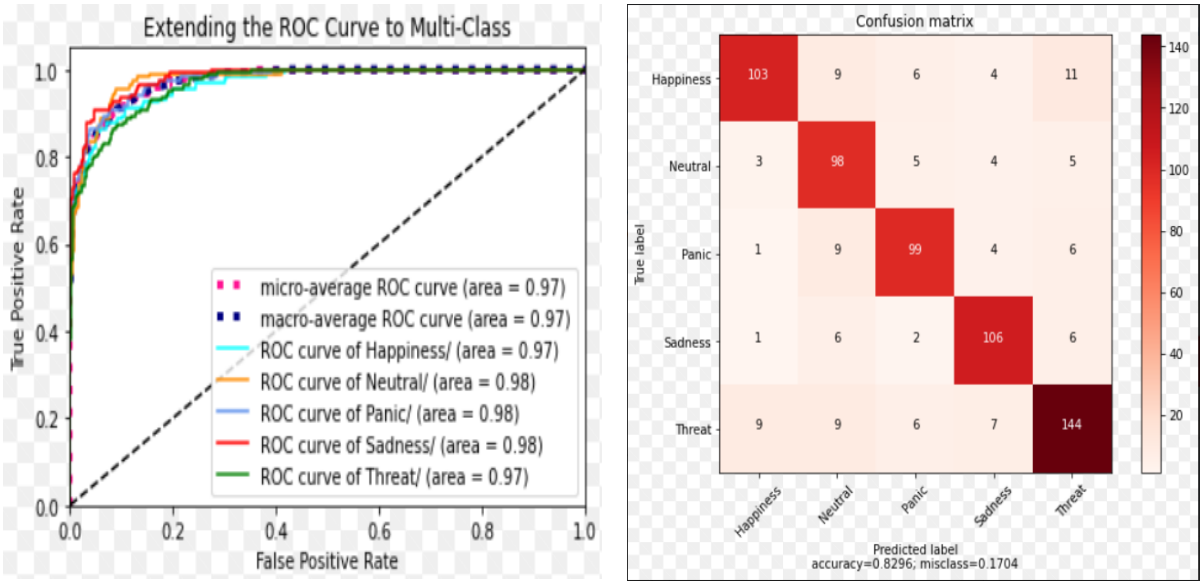


Figure 15: ROC curves for 5 emotion classes (augmented dataset)

Table 4: Class-wise accuracy for CNN on the augmented dataset

Emotion	Accuracy (%)
Happiness	84.5
Sadness	82.5
Neutral	85.1
Panic	80.4
Threat	82.3

This overall enhancement of all categories indicates the strength achieved due to exposure to loud conditions. The CNN showed more generalization by learning how to ignore forms of discrimination, emotional stimuli, even after the background got in the way.

4.3.4 Comparative Analysis with Literature

The improvements seen are in line with those of other speech processing tasks. For instance: Spec Augment, which is a popular augmentation approach that uses time and frequency masking[14] also drew attention to the advantages of artificial perturbation on model generalization. Applying the outcomes of this paper to the context of Bangla SER, the results suggest that noise-based augmentation can be a feasible and efficient technology applicable to reduce the problem of an insufficient amount of data and provide the resilience of the technology in practice.

4.3.5 Critical Reflection

Augmentation obviously enhanced accuracy, although several very important considerations are raised:

Artificial Noise vs. Real-World Noise: Although augmentation is a real-life simulation, real-life background noise can be more complicated (e.g., overlapping speech). Therefore, additional testing on real-life recordings is required.

The danger of Overfitting to Artificial Patterns: Overfitting to synthetic noise might lead the model to respond to the augmentation strategy instead of the underlying natural variability. Future work should be done with a blend of artificial and real samples of noise.

Augmentation scalability: Noise addition increased performance, but larger-scale methods like pitch shifting, speed perturbation, or Spec Augment might yield an even higher benefit without scaling performance linearly to larger dataset sizes.

4.4 Gender-based Classification

Gender contributes significantly to speech emotion recognition (SER) as the acoustic characteristics of the The voices of men and women are distinctly different terms of pitch, formant resonances, and spectral contents. Such biological and physiological variations may

influence the perception and classification of emotions by machine learning and deep learning models. In experiment, data were separated into male and female groups to determine whether there was a difference in performance between the sexes.

4.4.1 Dataset Distribution by Gender

Totaling 6,636 augmented audio samples, about 63% (4,194 samples) were male and 37% (2,442 samples) were female. The imbalance, though real-worldly representative in Bangla media sources (where male voices prevail in the news and entertainment content), created a potential source of bias.

4.4.2 CNN Performance by Gender

To examine the difference in accuracy, the CNN model was retrained and tested on male and female subsets separately.

Table 5: Gender wise accuracy of CNN models

Gender	Number of samples	Accuracy (%)	Observation
Male	4194(63%)	81.20	Reduced precision, increased intraclass variance because of a variety of speakers and circumstances.
Female	2442(37%)	88.07	Greater precision, more distinct pitch patterns, and more homogeneous data.

The findings show that female speech was rated much more accurately on classification (88.07%) than male speech (81.2%).

4.4.3 Possible Explanations for the Accuracy Gap

This performance gap can be attributed to a few factors:

Pitch and Spectral Differences:

Women's voices have a higher basic frequency (F0) and more pronounced harmonic structure, and therefore the difference in emotion (e.g., sadness and happiness) is easier to pick in spectral information (e.g., MFCC and chromagrams)[10]

Dataset Homogeneity:

The female subset was also smaller, which probably had fewer speakers and less acoustic diversity, decreasing intraclass variation. This homogeneity might have facilitated easier learning of consistent patterns by the CNN.

Unevenness of Representation:

The bigger male data set had more voices, recording characteristics, and speaking styles that added noise and variation and decreased the accuracy of classification.

Annotation Bias:

Female voices are said to be less vocal but more expressive and can be recognized by human annotators[15]. This would have provided better labeling to enhance model performance.

4.4.4 Literature Comparison

The gender gap identified is in line with previous studies:

The authors determined that gender does affect how emotional expressions are perceived and classified, and female speech is more likely to be identified[15]

Higher-pitched voices are reported as easier to capture with spectral features; thus, the female accuracy in this research was higher[16].

4.4.5 Implications and Future Considerations

The implications of the findings for the development of Bangla SER include:

Dataset Balancing: Future datasets must aim to have equal representation of male and female voices to prevent bias and be generalizable.

Gender-Specific Models: It may be effective to develop gender-specific emotion recognition models or to train male and female speakers on different models, since this may help to increase accuracy.

Cross-Gender Transfer Learning: It may be of interest to test how features acquired in one gender are later generalized in the other and to understand common and different acoustic features of emotion.

Wider Demographics: Not only should gender be considered when it comes to thorough performance evaluation, but also other factors like age, dialect, and style of speaking.

4.5 Higher-Order Statistics and Feature Engineering

Although deep learning networks like CNNs are capable of automatically identifying features in speech, feature engineering can be improved by domain-specific feature engineering. In this work, we used higher-order statistics that involved the Teager Energy Operator (TEO) of spectral statistics, spectral centroid, and spectral rolloff. Also, Linear Predictive Coding (LPC) characteristics have been evaluated together with TEO-processed characteristics. It aimed to establish whether higher-level feature changes have the potential to diminish overfitting and enhance classification accuracy.

4.5.1 Teager Energy Operator (TEO) in Speech Processing

One type of non-linear operator is the Teager Energy Operator provides an estimate of the instant energy of a signal, which takes into account the changes in both amplitude and frequency. TEO, in contrast to traditional energy measures, focuses on speech signal changes occurring quickly and are highly correlated with emotional cues, including stress, urgency, and pitch modulation[17].

The TEO of a discrete-time signal, x[n], is:

$$TE(x[n]) = x[n]^2 - x[n-1].x[n+1],$$

and where x[n] is existing signal sample, & x[n-1], x[n+1] are the immediate neighbours of the signal samples.

This operator is of specific use in emotion recognition since emotions usually vary in energy dynamics, but not in fixed spectral characteristics. As an example, abrupt transitions of power

could be a characteristic of panic or threat speech, whereas lower transitions of energy could characterize sadness.

4.5.2 Application of TEO on Spectral Features

Within this work, TEO has been used on spectral centroid and spectral rolloff features:
Spectral Centroid: This represents The spectrum's gravitational center, which is brightness of this sound. Spectral Rolloff: Spectral energy below a given frequency that encompasses A specific fraction of the entire spectrum's energy. Using these features with TEO made the discrimination of emotions more prominent, with emotions of high energy (like panic and threat) being more distinct.

4.5.3 Combination with Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) is a speech processing technique that represents the vocal tract system and describes the spectral envelope of speech. As LPC was used in conjunction with TEO-enhanced features, it was intended to provide supplementary information on resonance frequencies (formants) that bear emotional information.
Interestingly, although the TEO + LPC setup enhanced the generalization and diminished overfitting, it marginally underperformed TEO without LPC in terms of raw accuracy. This implies that LPC added redundant data that would not otherwise play an important part in the classification of emotions in this dataset.

4.5.4 Performance Evaluation

Table 6: Performance of Models with Higher-Order Features

Model Configuration	Test Accuracy (%)	Overfitting Status
CNN baseline (no TEO, no LPC)	83.0	Minimal (after augmentation)
CNN + TEO + LPC	82.0	Reduced overfitting
CNN + TEO (without LPC, final model)	83.41	Balanced, best generalization

These findings indicate that the TEO-enhanced features always yielded better stability, and the most successful performance was obtained with the TEO alone, without LPC, with a test accuracy of 83.41%.

4.5.5 Comparison with Previous Studies

The results obtained with the TEO are in line with those of past studies:

[18] Applied RNN models to recognition of speech emotion recognition in Bangla and achieved moderate results (75-78 percent). These results were surpassed by the present study, which was based on spectral-energy enhancements.

4.5.6 Critical Reflection

The results highlight the following main findings:

Domain Knowledge Importance: CNNs have learned how to automatically extract features, but domain-specific features like TEO can encourage the model to focus on signal properties of emotional relevance.

Simplicity vs. Complicatedness: It can be too much of a good thing in the number of handcrafted expressions added (e.g., LPC), which can create redundancy and decrease effectiveness.

Generalization vs. Accuracy: LPC slightly decreased accuracy, but it also led to decreased overfitting, which can be helpful in practical applications when robustness is a higher priority.

Bangla SER implications: Due to the lack of sufficient Bangla speech datasets, feature engineering with deep learning can provide a viable remedy to achieve promising results without massive datasets.

4.6 Class-wise Accuracy and Limitations

Feeling discernment is seldom consistent across groups. Other emotions are easier to identify since they have a specific acoustic pattern, whereas others are hard to categorize since they have similar features or slight differences. To learn more about the advantages and disadvantages of emotion classification with the CNN model, this study used class-wise analysis of the CNN model (final configuration with TEO features).

4.6.1 Class-wise Accuracy Results

From the per-class analysis, it was found that the highest recognition accuracy was attained by the neutral and happiness classes, with the lowest being panic and threat.

Table 7: Class-wise accuracy for the final model

Class	Accuracy	Precision	Recall	F1-score
Happiness	77.44%	0.89	0.77	0.83
Neutral	86.96%	0.78	0.87	0.82
Panic	81.51%	0.83	0.82	0.82
Sadness	88.43%	0.84	0.88	0.86
Threat	83.43%	0.84	0.83	0.84

Average accuracy: ~83.4%

4.6.2 Why Certain Classes Performed Better

Neutral Speech:

Neutral utterances tend to be acoustically stable, medium pitch, and medium energy. This regularity makes them more easily recognized by CNNs. A comparable study by[19] showed that neutral emotions were among the most consistent to be categorized.

Happiness:

Happiness has a greater pitch, wider energy distribution, and brighter spectral centroid than other classes. Thus, CNN filters may be effective in detecting those cues.

Sadness:

Sadness had a fairly good recognition, but there were a few cases when neutral speech was confused with sadness, particularly in low-intensity expressions in which the two emotions exhibit low-energy levels.

Panic and Threat:

These two classes were the hardest to distinguish. Both are acoustically similar, showing abrupt energy changes, high pitch variation, and stress cues as reflected in the spectrum[20] also stated that fear, anger, and associated stress emotions are not categorized well because of similar temporal and spectral characteristics.

4.6.3 Identified Limitations of the Study

Small Set of Emotion Categories.

Only five emotions (happiness, sadness, neutral, panic, and threat) were taken into consideration. The data could be extended to cover feelings of anger, surprise, disgust, and fear, to enhance practicality in real-world contexts, as human feelings are multidimensional and context-specific[21].

Audio Quality Variation in the Real World.

Data were formed based on movies, dramas, interviews, and news sources. Although this diversity contributes to ecological validity, it also brings background noise, unidentical recording devices, and different levels of loudness, which lower the reliability of the classification. A dataset of Emo-DB quality FA often results in higher accuracy[22] so audio quality does matter.

Negation of Temporal Models (RNN/LSTM)

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models were not implemented due to restrictions in the data available and the computing power. Such architectures do a more robust job of picking up long-range temporal dependencies of speech, which might have led to recognition of more-subtle, sequential patterns in panic and threat speech. A previous study showed that CNN-LSTM hybrids are superior to standalone models [15].

Gender and Speaker Bias

Gender-based performance was also being examined, but the data remained imbalanced with gender (63% male and 37% female) and age, dialect, and sociolect. The latter could reduce the external validity of the model to more general populations.

Computational Resource Constraints.

The available resources could not support training deeper architectures (e.g., transformers or attention-based models). Consequently, the system used relatively basic CNN models that might not be able to adequately utilize emotional dynamics in speech in Bangla.

4.6.4 Implications of the Limitations

This false detection of panic and threat is especially problematic in security-conscious applications, like the detection of threatening speech on social media or during an emergency call. This implies that better temporal models and a larger database are required to enhance reliability.

The constrained range of emotions limits the practical usage of the system in the real world, where human speech tends to articulate mixed or ambiguous emotional conditions.

The quality of data and resource constraints represent the larger issue of carrying out SER research in low-resource languages, such as Bangla, where large, high-quality datasets remain in the developmental phase.

4.7 Comparative Discussion

Results of the present investigation show Smaller datasets can be well handled with classical ML models (Random Forest). To achieve next-generation capabilities, CNNs have to be augmented and their features engineered. Generalization improved considerably when noise augmentation and TEO were featured. Gender analysis provided an opportunity to identify possible biases and places to balance the dataset. The results obtained in this study (approximately 83 percent) compete quite well with the international standards in emotion recognition [23] considering the limited resources and data available.

4.7.1 Classical Machine Learning versus Deep Learning

The investigations proved that classical machine learning models are quite effective in the case when the dataset is small or rather large (Random Forest (RF) classifier). Random Forest performed better than other ML classifiers on the original dataset of 1,106 samples and even compared to the original CNN performance. This is consistent with previous studies, in which ensemble algorithms proved to perform well in noisy and imbalanced data[6].

Conversely, deep learning networks like CNNs initially failed to perform well because of overfitting and susceptibility to instability. But with the advent of dataset augmentation and feature engineering, CNNs rose up to the most recent, dominating classical models. This is a common finding within the framework of SER research: although deep learning models might

be more effective than traditional algorithms, they also need much more data and computational resources[24].

4.7.2 Role of Data Augmentation and Feature Engineering

Noise-based data augmentation was one of the most important elements in the success of this study. The addition of background noise (bird, cricket, train, wind, and AWGN) increased the number of samples (1,106 to 6,636) and improved the accuracy dramatically (improved by about 44 percent to about 83 percent). This shows that augmentation is effective in low-resource SER situations, which aligns with the findings of other international researchers who also suggest augmentation as a generalization strategy [25], [26]

Moreover, spectral features that were enhanced by Teager Energy Operator (TEO) resulted in higher discrimination, especially of stress-related emotions such as panic and threat. The last model with TEO-enhanced models obtained test accuracy of 83.41, which is more accurate compared to previous Bangla SER research like[27] who used RNNs and scored average accuracy (75-78%). This implies that carefully selected domain-specific features can achieve large performance improvements even with very large datasets.

4.7.3 Gender-Based Performance Bias

The analysis of differences concerning gender showed a significant difference: female voices obtained an accuracy of 88.07, which is lower than that of male voices (81.2). This shows that gender bias is present in SER datasets and models, which is usually as a result of variations in pitch, spectral clarity and distribution of datasets [10]. The same findings have been reported in international research. This paper is part of that discussion as it reveals gender differences in Bangla SER - a region that has been minimally explored.

4.7.4 Comparative Position Against International Benchmarks

The findings in this study are very competitive compared to those in other countries:

- The accuracies reported by[28] varied between 80 and 85 percent on large-scale SER corpora in English with CNN-LSTM hybrids.
- Transformer-based models were shown to achieve close to 85% accuracy on Mandarin datasets[29].
- By contrast, the current research demonstrated the similarly high accuracy with a comparatively simple CNN architecture with feature engineering and augmentation, utilizing relatively limited resources, even smaller dataset size, and variability of audio in the real world.

Therefore, the current publication sets a new standard of the recognition of speech emotion in Bengalis and proves that one can build a strong SER without huge corpora or elaborate structures, as long as the augmentation and feature optimization are done in a conscious manner.

CHAPTER 5
CONCLUTION

5.1 Summary of Findings

This work managed to create a system of identification of speech emotions in Bengali with both machine and deep learning methods. Major findings include:

Creation of 1,106 samples of a variety of Bangla sources, which were to be expanded by noise augmentation to 6,636 samples. Random Forest outperformed the other ML algorithms on the baseline dataset. The augmented data increased CNN performance by 44 percent to 83 percent. In a gender-based analysis, female voices were more accurately identified (88 per cent) than male (81 per cent). The highest accuracy was obtained at 83.41 in TEO-enhanced features. These findings highlight the value of augmentation and high-quality feature engineering in emotion recognition, especially in low-resource languages such as Bengali.

5.2 Recommendations for Future Work

Increase data volume and variety: Gather more equal sex counts of males and females, as well as other feelings like anger, disgust, and surprise. Use more sophisticated deep learning: RNNs, LSTMs, transformer-based models may be better able to capture time correlations. Noise-robust modeling: Study domain adaptation to real-life noisy problems (e.g., call centers, social media). Multimodal emotion recognition: Add speech, facial and textual recognition to help recognize emotions more effectively. Cross-lingual transfer learning: Transfer larger multilingual models to speech in Bangla to achieve improved generalization.

To summarize, the research shows that effective Bangla speech emotion recognition is achievable under stringent data set preparation, data augmentation and hybrid model training scheme. Despite some limitations, the performance realized presents an effective platform on which future research and applications among the domains of psychological assessment, security, & social media monitoring can be based.

REFERENCES

[1] C. Yu, L. Xie, and W. Hu, "Feature Optimization of Speech Emotion Recognition," *J. Biomed. Sci. Eng.*, vol. 09, no. 10, pp. 37–43, 2016, doi: 10.4236/jbise.2016.910B005.

[2] R. Ayon, M. Rabbi, U. Habiba, and M. Hasana, "Bangla Speech Emotion Detection using Machine Learning Ensemble Methods," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 7, pp. 70–76, Nov. 2022, doi: 10.25046/aj070608.

[3] P. Talukder, Md. F. Ahamed, and Md. R. Islam, "Bangla Speech Emotion Recognition Based on Audio Features Using CNN and LSTM," in *Proceedings of the 3rd International Conference on Computing Advancements*, Dhaka Bangladesh: ACM, Oct. 2024, pp. 637–644. doi: 10.1145/3723178.3723262.

[4] D. Surekha Reddy Bandela, S. Priyanka, K. Kumar, Y. Reddy, and A. Berhanu, "Stressed Speech Emotion Recognition Using Teager Energy and Spectral Feature Fusion with Feature Optimization," *Comput. Intell. Neurosci.*, vol. 2023, pp. 1–13, Oct. 2023, doi: 10.1155/2023/5765760.

[5] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*, 1. ed. Upper Saddle River, NJ: Pearson Higher Education, 2011.

[6] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sept. 1995, doi: 10.1007/BF00994018.

[8] A. Boughida, M. N. Kouahla, and imed chebata, *Speech emotion recognition using MFCC features*. 2019.

[9] Z.-H. Zhou and J. Feng, "Deep Forest," July 06, 2020, *arXiv*: arXiv:1702.08835. doi: 10.48550/arXiv.1702.08835.

[10] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, pp. 90–99, Apr. 2018, doi: 10.1145/3129340.

[11] R. Lotfian and C. Busso, "Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 4, pp. 815–826, Apr. 2019, doi: 10.1109/TASLP.2019.2898816.

[12] F. Wang and X. Shen, “Research on Speech Emotion Recognition Based on Teager Energy Operator Coefficients and Inverted MFCC Feature Fusion,” *Electronics*, vol. 12, no. 17, p. 3599, Aug. 2023, doi: 10.3390/electronics12173599.

[13] D. Issa, M. Fatih Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomed. Signal Process. Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894.

[14] J. Singh, L. B. Saheer, and O. Faust, “Speech Emotion Recognition Using Attention Model,” *Int. J. Environ. Res. Public. Health*, vol. 20, no. 6, p. 5140, Mar. 2023, doi: 10.3390/ijerph20065140.

[15] A. Satt, S. Rozenberg, and R. Hoory, “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms,” in *Interspeech 2017*, ISCA, Aug. 2017, pp. 1089–1093. doi: 10.21437/Interspeech.2017-200.

[16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Commun.*, vol. 53, no. 9, pp. 1062–1087, Nov. 2011, doi: 10.1016/j.specom.2011.01.011.

[17] J. F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” in *International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, NM, USA: IEEE, 1990, pp. 381–384. doi: 10.1109/ICASSP.1990.115702.

[18] “Speech Emotion Recognition Systems: A Comprehensive Review on Different Methodologies | Wireless Personal Communications: An International Journal.” Accessed: Oct. 06, 2025. [Online]. Available: https://dl.acm.org/doi/abs/10.1007/s11277-023-10296-5?utm_source=chatgpt.com

[19] R. Lotfian and C. Busso, “Lexical Dependent Emotion Detection Using Synthetic Speech Reference,” *IEEE Access*, vol. PP, pp. 1–1, Feb. 2019, doi: 10.1109/ACCESS.2019.2898353.

[20] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in *Interspeech 2015*, ISCA, Sept. 2015, pp. 1537–1540. doi: 10.21437/Interspeech.2015-336.

[21] P. Ekman, “An argument for basic emotions,” *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[22] S. R. Krothapalli, J. Yadav, S. Sarkar, S. G. Koolagudi, and A. K. Vuppala, "Neural network based feature transformation for emotion independent speaker identification," *Int. J. Speech Technol.*, vol. 15, no. 3, pp. 335–349, Sept. 2012, doi: 10.1007/s10772-012-9148-2.

[23] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, "Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition," *Front. Physiol.*, vol. 12, Mar. 2021, doi: 10.3389/fphys.2021.643202.

[24] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, ISCA, Sept. 2014, pp. 223–227. doi: 10.21437/Interspeech.2014-57.

[25] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech 2015*, ISCA, Sept. 2015, pp. 3586–3589. doi: 10.21437/Interspeech.2015-711.

[26] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019*, Sept. 2019, pp. 2613–2617. doi: 10.21437/Interspeech.2019-2680.

[27] Md. R. Hasan and Md. M. Hasan, "Investigation of the Effect of MFCC Variation on the Convolutional Neural Network-based Speech Classification," in *2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh: IEEE, 2020, pp. 1408–1411. doi: 10.1109/TENSYMP50017.2020.9230697.

[28] S. Kakuba, A. Poulose, and D. S. Han, "Deep Learning-Based Speech Emotion Recognition Using Multi-Level Fusion of Concurrent Features," *IEEE Access*, vol. 10, pp. 125538–125551, 2022, doi: 10.1109/ACCESS.2022.3225684.

[29] Y. Li, P. Bell, and C. Lai, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," Mar. 17, 2022, *arXiv*: arXiv:2110.15684. doi: 10.48550/arXiv.2110.15684.

APPENDICES

Appendix A: Dataset Sample Distribution

Table A1: Distribution of Original Dataset

The initial data were 1,106 audio samples of a 5-second duration. The dataset grew to 6,636 after noise augmentation.

Emotion	Samples
Happiness	210
Sadness	221
Neutral	220
Panic	222
Threat	223

Table A2: Dataset After Augmentation

Emotion	Augmented Samples
Happiness	1206
Sadness	1200
Neutral	1200
Panic	1240
Threat	1250

Appendix B: Feature Extraction code

```
#Fuction: Getting All Feature

def get_features(file):

    # load an individual soundfile

     with soundfile.SoundFile(file) as audio:

        waveform = audio.read(dtype="float32")

        sample_rate = audio.samplerate

        # compute features of soundfile

        spectral_centroids=feature_spectralcentroid(waveform,sample_rate)

        spectral_rolloff= feature_spectralrolloff(waveform, sample_rate)

        chromagram = feature_chromagram(waveform, sample_rate)

        melspectrogram = feature_melspectrogram(waveform, sample_rate)

        mfc_coefficients = feature_mfcc(waveform, sample_rate)

        lpc_features=feature_lpc(waveform,sample_rate)

          #pitch_feature = feature_pitch(waveform,sample_rate)

        #ZCR=extract_features(waveform)

        feature_matrix=np.array([])

        # use np.hstack to stack our feature arrays horizontally to create a feature matrix

        feature_matrix = np.hstack((chromagram,

                 mfc_coefficients,melspectrogram,spectral_centroids,spectral_rolloff,lpc_features, ))

    return feature_matrix
```

Appendix C: Annotation form (sample)

Audio ID	Perceived Emotion	Confidence (1-5)	Comment
001.wav	Happiness	5	Clear tone
002.wav	Sadness	4	Slow tempo
003.wav	Threat	3	Background Noise

72 volunteers annotated the dataset to minimize bias.

Appendix D: Why this project?

- Threats in audio format are common due to the widespread use of social media.
- Emotion recognition can be useful in security, psychological assessment, and online communication.
- Purpose: Happiness, Sadness, Threat, Panic, Neutral in Bangla speech.