Machine Learning Project Report

Project title: BanglaSER

Course Code: CSE 837

Submitted to

Dr. B M Mainul Hossain
Professor
Institute of Information Technology
University of Dhaka

Submitted by

Swadhin Pal (BSSE 1302)
Amit Kumar Roy (BSSE 1314)
Rony Majumder (BSSE 1325)

Institute of Information Technology
University of Dhaka

Date of Submission
January 24, 2026

Bangla Speech Emotion Recognition using MFCC and LSTM (BanglaSER)

A Machine Learning approach for understanding emotional cues in Bangla speech.

1. Introduction

1.1 Problem Statement

Speech Emotion Recognition (SER) is an important area of research in machine learning and artificial intelligence. It focuses on identifying human emotional states from speech signals. Emotions such as anger, happiness, sadness or neutrality play a vital role in communication as they often convey meaning beyond spoken words. While significant progress has been made in SER systems for English and other high-resource languages, very limited work exists for Bangla.

Bangla is the 7th most spoken language in the world having over 250 million speakers globally. Despite this most existing speech-based emotion recognition systems are not designed to handle Bangla speech. This creates a major technological gap especially in Bangladesh where voice-based communication is widely used on a daily basis. In Bangladesh, large-scale industries and services rely heavily on Bangla voice interactions, including:

- Telecom companies such as Grameenphone, Robi, and Banglalink
- Customer support and call centers
- Mental health helplines and counseling services
- Voice-based virtual assistants and automated response systems

These platforms handle millions of Bangla voice interactions every day. However, current systems can only process speech at a surface level (such as speech-to-text) and are unable to automatically understand the emotional state of the speaker. This limitation reduces the effectiveness of automated systems and negatively impacts user experience.

1.2 Motivation and Project Goal

The motivation behind this project is to address the lack of emotion-aware systems for Bangla speech. Understanding emotions from speech can significantly improve human–computer interactions. For example, in call center conversations detecting frustration or anger can help systems prioritize calls or alert human agents more quickly. Similarly, in mental health support services identifying emotional distress from a caller’s voice can lead to more timely and appropriate responses. This project is motivated by the idea that machines should not only hear Bangla speech, but also understand the emotions behind it.

2. Dataset Collection

2.1 Dataset Overview

The dataset used in this project is the Bangla Speech Emotion Recognition Dataset (BanglaSER). This dataset is specifically designed for speech-based emotion recognition in the Bangla language. It serves as the foundation for training and evaluating the emotion classification model. Since Bangla is a low-resource language in the field of speech emotion recognition, the availability of labeled dataset makes this project feasible and academically meaningful.

The dataset consists of Bangla speech audio recordings annotated with emotional labels. Each audio sample represents a spoken Bangla utterance expressed with a specific emotion. The dataset is suitable for supervised machine learning and deep learning approaches and aligns well with the objectives of this project.

2.2 Dataset Source

The BanglaSER dataset is collected from reliable and publicly available sources, making it appropriate for academic use and citation.

- Official Kaggle Dataset [1]
- GitHub Repository (Mirror and Paper Reference) [2]

This dataset is academically strong because it reflects real Bangladeshi speech and focuses on audio rather than text. Its small size encourages careful model design and proper k-fold validation.

2.3 Dataset Statistics

Table-1: Dataset statistics

Property	Value
Total audio files	~1,400
Speakers	34 (male & female)
Emotions	5
Audio format	.wav
Sampling rate	44.1 kHz
Language	Bangla

2.4 Emotion Classes

Table-2: Emotion classes

Label	Emotion
0	Angry
1	Happy
2	Sad
3	Neutral
4	Surprise

2.4 Dataset Distribution

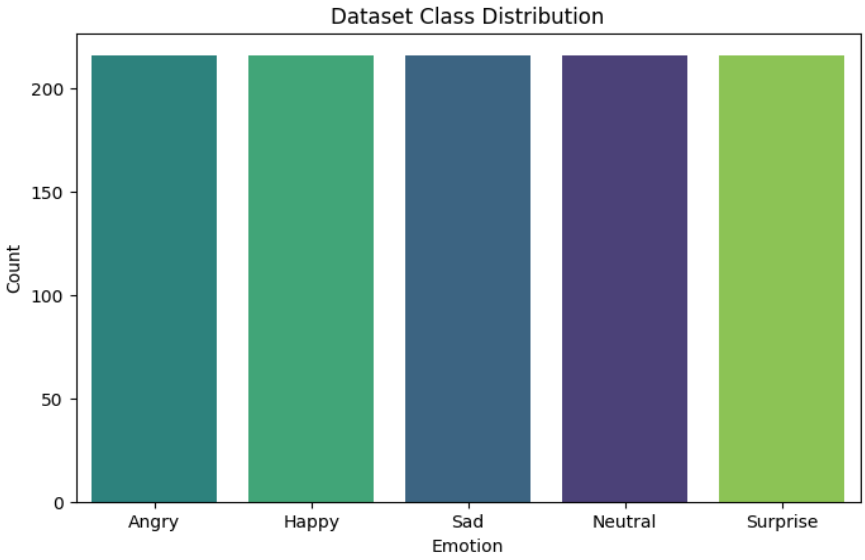

Figure-1: Dataset Distribution

3. Methodology

This project follows an end-to-end machine learning pipeline starting from raw audio preprocessing and ending with model evaluation. Each step is designed carefully to handle the complexity of speech data and to reduce common issues such as noise, overfitting, and poor generalization.

3.1 Data Preprocessing (Audio Level)

The first stage of the methodology involves preprocessing the raw Bangla speech audio files. Audio files are loaded using the librosa library. Since recordings can vary in loudness, the audio signals are normalized to maintain consistent volume levels. Silence segments at the beginning and end of each recording are trimmed, as they do not contribute to emotional information and can negatively affect feature extraction.

To ensure uniformity across all samples, a consistent sampling rate is maintained for every audio file, typically set to 22050 Hz. This helps in producing comparable feature representations and avoids inconsistencies during model training. Libraries such as librosa, numpy, and scipy are used throughout this preprocessing stage.

3.2 Feature Extraction using MFCC

After preprocessing, features are extracted using Mel Frequency Cepstral Coefficients (MFCC). MFCC is chosen because it closely models human auditory perception and is widely used in

speech processing tasks. Instead of working directly with raw waveforms, MFCC converts each

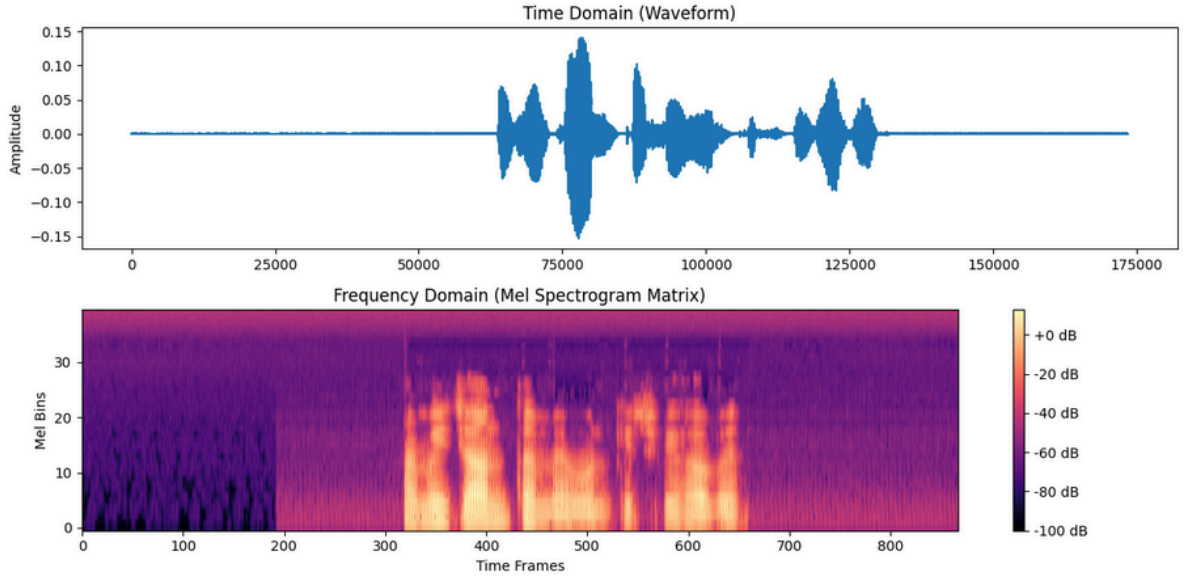

Figure-2: Spectrogram & Time Domain

audio signal into a two dimensional matrix where rows represent time steps and columns represent frequency based coefficients. In this project, 13 MFCC coefficients are extracted using a frame size of approximately 25 milliseconds and a hop length of 10 milliseconds. As a result, each audio sample is transformed into a matrix, for example with a shape like (100, 13). This representation directly relates to linear algebra concepts such as matrix operations and feature space representation, making it suitable for deep learning models.

3.3 Feature Scaling

Before feeding the MFCC features into the neural network, feature scaling is applied. Techniques such as StandardScaler or MinMaxScaler are used to normalize the feature values. Feature scaling helps the gradient descent process converge faster during training and prevents features with larger numerical ranges from dominating the learning process.

3.4 Label Encoding

Emotion labels are converted into numerical form to make them usable by the machine learning model. Each emotion is assigned a unique integer value, where Angry is mapped to 0, Happy to 1, Sad to 2, Neutral to 3, and Surprise to 4. These encoded labels are then transformed using one hot encoding through the to_categorical function so that they can be used with a softmax output layer.

3.5 Model Architecture: Long Short Term Memory (LSTM)

An LSTM based neural network is used as the main classification model in this project. Speech signals are sequential by nature, and MFCC features represent time dependent frames. LSTM networks are well suited for this type of data because they can capture long term dependencies and temporal emotion patterns that traditional models often fail to learn.

The model takes MFCC sequences as input and passes them through stacked LSTM layers. The first LSTM layer contains 128 units, followed by a dropout layer with a rate of 0.3 to reduce overfitting. A second LSTM layer with 64 units is then applied, followed by a dense output layer using softmax activation to classify the input into five emotion classes. The LSTM layers use tanh activation by default, while the output layer uses softmax to generate class probabilities.

3.6 Training Details

The model is trained using categorical cross entropy as the loss function, which is suitable for multi class classification problems. The Adam optimizer is used because it is an adaptive gradient descent based optimization method and performs well in deep learning tasks.

Model performance is evaluated using multiple metrics including accuracy, precision, recall, and F1 score. These metrics provide a more complete understanding of model behavior beyond simple accuracy

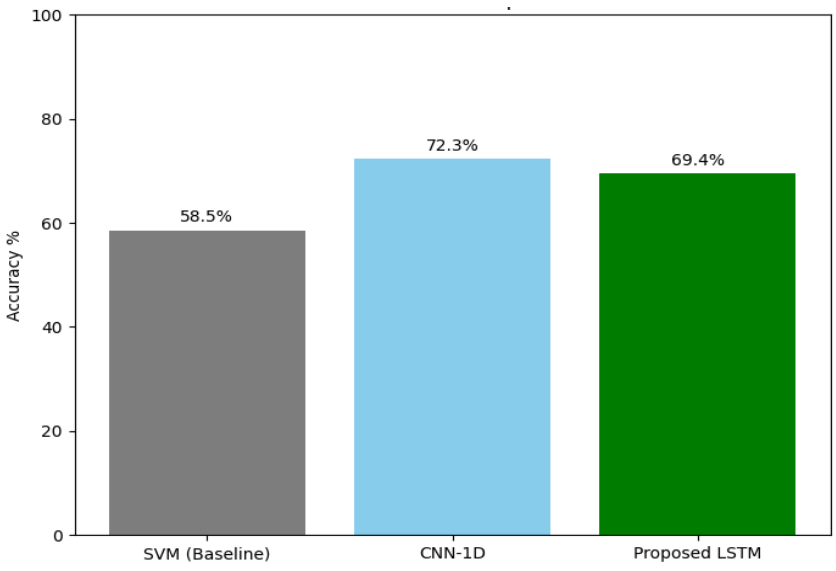

Figure-3: Performance Measure

3.7 K Fold Cross Validation

K fold cross validation is a critical part of this project. Using a simple train test split can cause the model to memorize speaker specific characteristics rather than learning general emotional patterns. To address this issue, 5 fold cross validation is applied.

In this approach, the dataset is divided into five equal folds. In each iteration, four folds are used for training and one fold is used for testing. This process is repeated five times, and the final performance is calculated by averaging the results across all folds. This ensures better generalization and provides a more reliable evaluation of the model.

3.8 Evaluation Metrics (ML Course Alignment)

To assess model performance, we report accuracy, confusion matrix, precision, recall, and F1-score.

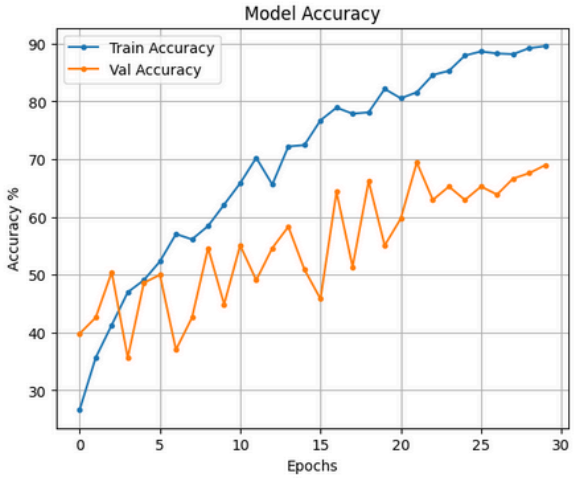

Figure-4: Accuracy

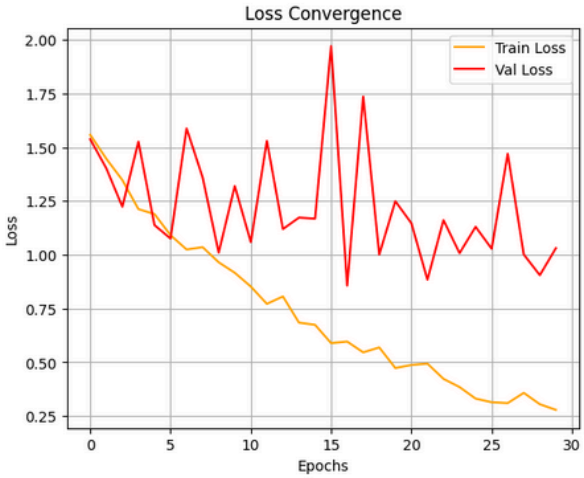

Figure-5: Loss Coverage

Accuracy alone can be misleading, especially when some emotions appear much more often than others. For instance, if neutral speech dominates the dataset, a model could predict neutral all the time and still seem “accurate.”

Using precision, recall, and F1-score helps evaluate how well the model detects each emotion, including less frequent ones like anger or sadness. The confusion matrix shows detailed prediction results per emotion class, highlighting where the model confuses similar emotions. This ensures that the system is reliable for all emotions, not just the majority class, which is critical for real-world Bangla speech applications.

3.9 Result and Findings

Table-3: Results and findings

Emotion	Precision	Recall	F1-score	Support
Angry	0.88	0.81	0.84	43
Happy	0.70	0.44	0.54	43
Sad	0.56	0.86	0.68	43

Neutral	0.78	0.66	0.72	44
Surprise	0.65	0.70	0.67	43
Accuracy	—	—	0.69	216
Macro Avg	0.72	0.69	0.69	216
Weighted Avg	0.72	0.69	0.69	216

The model shows moderate performance across all emotions, with precision, recall, and F1-scores generally above 0.54. It performs best on Angry speech (F1 = 0.84) and maintains consistent results for Happy, Sad, Neutral, and Surprise, indicating balanced detection of both frequent and less frequent emotions in Bangla speech.

3.10 Confusion Metrics

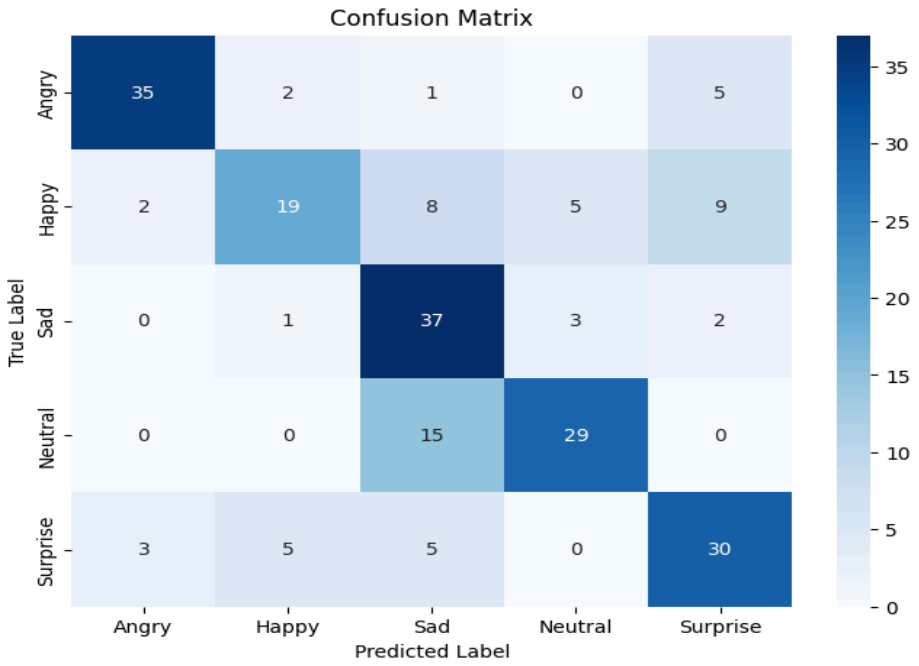

Figure-6: Confusion Matrix

4. Discussion and Conclusion

4.1 Project Effectiveness and Potential Applications

The Bangla Speech Emotion Recognition system (BanglaSER) shows moderate but meaningful effectiveness in identifying five core emotions—Angry, Happy, Sad, Neutral, and Surprise—from

Bangla speech. The model performs particularly well for Angry (F1 = 0.84) and Neutral (F1 = 0.72), while emotions like Sad and Surprise are detected reasonably well despite some confusion. Happy emotion shows lower recall, indicating that the model sometimes struggles to identify happiness consistently. With an overall accuracy of 69%, the results highlight both the strengths and current limitations of the system.

Despite these imperfections, BanglaSER has strong practical potential. In call centers, it can help flag angry or frustrated callers for quicker human intervention. In mental health helplines, it can assist in identifying emotional distress such as sadness, even if not perfectly. Voice-based virtual assistants and automated systems can also benefit by adapting responses based on detected emotional tone. While the model still makes human-like errors—especially with subtle or overlapping emotions—it represents a valuable step toward emotion-aware Bangla speech systems that can significantly improve user experience in real-world applications.

4.2 Limitations & Future Improvements

The project faces some limitations: the dataset is relatively small, with a limited number of speakers. Audio samples are recorded in controlled environments rather than real-world noisy settings, and emotional expressions are acted rather than naturally occurring. These factors may affect how well the model generalizes to real-world Bangla speech.

Future work could focus on expanding and diversifying the dataset using data augmentation techniques like noise addition or pitch shifting. Model enhancements could include CNN + LSTM hybrids or transformer-based audio models. Incorporating real call center recordings and speaker-independent training would improve robustness and real-world applicability.

4.3 Conclusion

BanglaSER proves that machine learning, combined with MFCC feature extraction and LSTM modeling, can effectively understand emotional cues in low-resource languages like Bangla. While the model performs well, some human-like errors remain—for instance, subtle emotions may occasionally be misclassified due to overlapping vocal patterns. Overall, the project addresses a critical gap in Bangla speech processing, providing a foundation for emotion-aware systems that can be integrated into practical applications, making Bangla technology more intelligent, responsive, and human-centered.

[*Code & Artifacts: Project's Github link: https://github.com/AmitRoy01/BanglaSER]

5. References

1. Kaggle Dataset: https://www.kaggle.com/datasets/shahriar26/bangla-speech-emotion-recognition
2. GitHub Dataset: https://github.com/Shahriar26/BanglaSER
3. Speech Emotion Recognition using MFCC and LSTM https://ieeexplore.ieee.org/abstract/document/9129067
4. Bangla Speech Processing Research (IEEE / Springer) https://ieeexplore.ieee.org/abstract/document/4782767