A Multimodal Approach for Human Emotion Recognition from Bangla Text and Speech

Md Jabed Hosen and Iqbal Ahmed*

***Abstract*—Emotion recognition is crucial in Human-Computer Interaction (HCI). This research presents a multimodal emotion recognition approach for low-resource languages like Bangla, integrating text and speech modalities. We have applied several machine learning, deep learning, and BERT-based models for emotion classification and recognition. Among the models, the BanglaBERT (for text) and XGBoost (for speech) models achieved the highest accuracies of 64% and 72% respectively. Subsequently, we conducted the Multimodal Late (Decision level) fusion using the BanglaBERT and XGBoost models, resulting in a 70% accuracy, which surpasses several existing multimodal emotion recognition studies.**

***Index Terms*—Multimodal Emotion Recognition, Decision Level Fusion, Deep Learning, Bangla Text and Speech**

I. Introduction

IN today's digital world, it is important to understand the sense of any language by computers as Human-Computer Interaction (HCI) becomes more essential. Numerous studies have explored emotion classification[1], [2], [3] and recognition [4], [5]. Most studies consider a single modality (such as text, image, speech, or video) for emotion classification or recognition tasks. These studies have also found significant results for Bangla text-based emotion classification [6], [7], [8], [9] and for Bangla speech-based emotion recognition [10], [11], [12]. However, there are still some loopholes in these studies. Just depending on a single modality, we could not able to extract the exact emotions in various conflicting scenarios. Moreover, textual features[13], [14], [15] and patterns would not be sufficient to classify emotion as speech tone. Intensity[16] of tone also plays a crucial role in these cases. In addition, finding accurate speech data is also a challenging task, as it can be affected by background noise, unclear pronunciation, variations in pronunciation, and gender-based speaker differences [17]. To address these challenges, we propose a multimodal emotion recognition approach for the Bangla language, considering the text and speech modalities. This research makes the following key contributions:

1. Propose a novel multimodal emotion recognition approach for the Bangla Language by combining the text and speech data as modalities using the multimodal Late (decision-level) fusion.
2. Address the challenge of limited multimodal emotion datasets for Bangla text and speech modality by introducing a practical solution to align and balance emotion classes across two separate datasets (text and speech) for effective multimodal decision-level fusion.
3. Utilize an improved late (decision) fusion strategy using weighted averages for each modality (text and speech models) alongside confidence scores for each emotion prediction, to calculate the final emotion score and determine the predicted emotion class.

Manuscript received February 23, 2025; revised May 4, 2025.

The research work is supported and funded by The Research and Publication Cell, University of Chittagong, Bangladesh.

Md Jabed Hosen is a postgraduate student of the Department of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh (email: jabed.hosen.cu@gmail.com).

Iqbal Ahmed* is a Professor at the Department of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh (Corresponding author, Phone: +8801711481086, email: iqbal.ahmed@cu.ac.bd).

II. Related Works

This section discusses many related studies of the multimodal approach. Some key parameters, such as modality, the number of data, language, fusion approach (feature/ decision level), applied models, and research findings based on evaluation metrics, are considered for discussion.

This study [18] focused on applying a multimodal fusion approach to recognize emotion with 4600 sample images containing Bangla captions. They found that the ResNet50 (for image) and BiLSTM (for Bangla text) models achieved a weighted F1-score of 0.7750 in the feature fusion case. In the decision fusion scenario, the InceptionV3 (for image) and BiLSTM (for Bangla text) models performed best with a weighted F1-score of 0.7587.

To add diversity in multimodal emotion recognition, this study [19] considered the three root modalities (Text, Audio, Video) with 1002 data samples consisting of four emotion classes. They found that their multimodal feature fusion-based approach worked well, with an F1 score of 0.64, using the YamNet model for audio, the BanglaBERT model for text, and the DeepFace model for video.

For binary classification (Hate vs. Non-Hate) of Bangla speech, Karim et al. [20] worked with 4500 samples of images containing Bangla captions. Their research shows that the multimodal fusion of DenseNet-161 (For Image) and XLM-RoBERTa (For Text) performed well with an F1-score of 0.80.

For sentimental analysis, Elahi et al. [21] used image and text modalities using multimodal fusion with 4372 images (including captions). This research introduces diversity in the multimodal approach by working with Bangla and English captions. They found that the combination of ResNet50 (for image) and BanglishBERT (for text) performed well, achieving an F1-score of 0.71.

Hossain et al.[22] conducted a multimodal fusion approach to classify Bangla memes with 4368 image samples containing Bangla text captions. They found that the combination of ResNet50 and BiLSTM performed well, achieving an F1-score of 0.62 for the feature fusion case and 0.64 for the decision fusion case using the ResNet50 and the text-based CNN models.

For multimodal sentiment analysis, this study [23] used the image and text modalities of the English language. They considered the two MVSA datasets [24], which contain 4,347 and 17,024 image-text samples, respectively. They applied SVM, BiLSTM, and CNN models for unimodal and fusion-based approaches. Their research found that the feature-based fusion performed well with an F1-score of 0.6629.

In this study, Srivastava et al. [25] focused on multimodal decision fusion of text and audio for sentiment analysis.They used two datasets: one with English audio and another with English text. For audio, they considered 11700 samples of audio data. The audio data was transcribed to text using Python scripting. They also processed a text dataset containing 8,000 samples labeled as positive and negative. The combination of CNN and BERT models was used for the decision fusion which achieved a good F1-score of 0.88 for sentiment analysis.

To recognize emotions from text and speech, this study [26] applied a cross-model fusion. It used the MELD [27] dataset that contains a total of 13708 samples of audio and text. This study found a moderately weighted F1-score of 0.6628 using deep learning models for multimodal fusion.

This research [28] focused on multimodal emotion recognition through audio and video modalities. Salas et al. considered three publicly available multimodal datasets, such as RAVDESS[29], SAVEEE[30], which [31] is described in detail, and CREMA-D [32] for this experiment. They considered feature fusion and found that their multimodal approach outperforms existing research in most cases.

This multimodal approach [33] based research worked on multimodal emotion recognition using the MELD [27] dataset and IEMOCAP [34] dataset. Meng et al. applied some deep learning models. They found that their applied Class Boundary Enhanced Representation Learning (CBERL) model strategy performed well with an F1-score of 0.6927 for the IEMOCAP dataset and 0.6689 for the MELD dataset.

III. Methodology

This section covers the research approach for multimodal emotion recognition using Bangla text and speech. The overall approach is divided into some key tasks such as data collection, data preprocessing, and feature extraction. Next, we applied machine and deep learning models on the Bangla text and speech dataset for training. Finally, the trained models are tested on unseen data to evaluate their performance. The overall research process is depicted in Fig. 1.

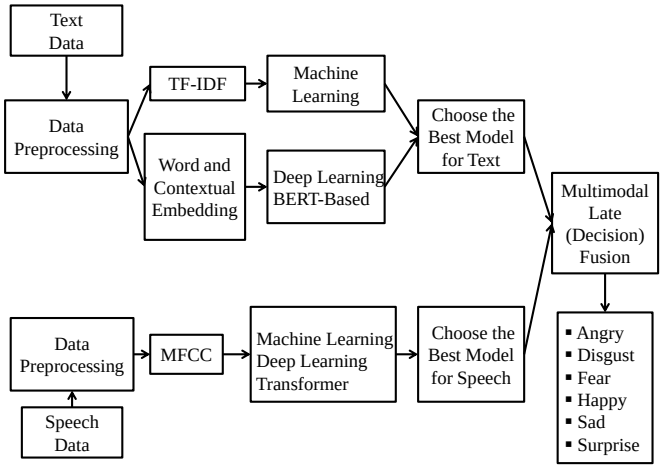

Fig. 1. Proposed Method for Multimodal Emotion Recognition

A. Data Collection

We have explored various publicly available datasets. To ensure optimal alignment for multimodal decision-level fusion of text and speech modalities, we have decided to choose the "Bengali Ekman's Six Basic Emotion" dataset [35] for Bangla text data and the "SUBESCO" dataset [36] for Bangla speech data. The text dataset has 36000 Bangla text samples with 6 emotions (Angry, Disgust, Fear, Happy, Sad, and Surprise) classes. The speech dataset has 7000 speech samples consisting of 7 emotions (Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise) classes. It is important to note that the dataset should be well aligned and balanced between modalities in multimodal fusion. As we have two separate datasets without matching context, it is not feasible to apply feature-level fusion. However, decision-level fusion could be applied using different datasets, as it is a technique by which we can combine the predictions from the different modalities and then aggregate their final output through majority voting or weighted average.

TABLE I
Emotion Dataset Details

Dataset	Original	Adjusted
Speech	7,000 samples (1,000 per class)	6,000 samples (Neutral class removed)
Text	36,000 samples (6,000 per class)	6,000 samples 5,940 (From Text), 60 transcribed (From Speech)

We have refined Ekman's Bangla text dataset and the SUBESCO speech dataset by matching emotion classes and sample sizes. The 'Neutral' emotion class was removed from the speech dataset. As a result, the speech dataset consists of in total of 6000 samples. The speech dataset was transcribed into text using Python scripting. We found 10

unique sentences that are used to express emotion in different ways. We planned to add them (A total of 60 texts, per class 10) to the text dataset for better multimodal decision-level fusion. In the text dataset, there are 36000 text data samples. We downsampled the dataset, reducing it to 5940 data samples. Then, we add the 60 text samples from the speech dataset. Thus, the text dataset contains a total of 6,000 samples, including 5,940 primary samples and 60 additional samples derived from the speech dataset. Finally, we have 6000 samples for both text and speech datasets, containing 1000 samples per emotion class. The summary of the dataset is represented in Table I.

B. Data Preprocessing

We have applied several data preprocessing techniques to the text and speech dataset. Regular expressions were used to remove non-Bangla characters, special characters, digits, and punctuation from the Bangla text dataset. The Bangla stopwords were removed using Spark NLP's pre-trained filter. An example of the data preprocessing for the Bangla text is depicted in Fig. 2. Butterworth low-pass filtering was

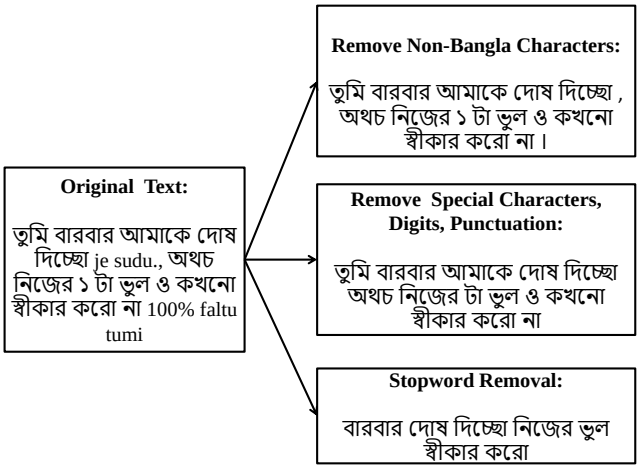

Fig. 2. Data Preprocessing Example for Bangla Text

used to process the speech data to reduce noise. Finally, the Bangla text and speech datasets are divided into an 80:10:10 ratio for training, validation, and testing sets.

C. Feature Extraction

Feature extraction is the approach to transform raw data into a meaningful attribute so that a machine could understand the pattern of the data. We have used the Term Frequency-Inverse Document Frequency (TF-IDF) [37] technique for machine learning models to extract the feature from the Bangla text data. The word embedding technique and contextual embedding were used for deep learning and BERT-based models, respectively. TF-IDF is a statistical approach to assigning the impact of a word within a text document relative to a larger collection of documents or datasets.

TF (Term Frequency): It calculates how frequently a word appears in a document.

$$TF(t,d) = \frac{f_t}{T_d}$$

where:
f_t = Number of occurrences of term t in document d
T_d = Total number of terms in document d

IDF (Inverse Document Frequency): It calculates how unique a word is across all documents in the corpus.

$$IDF(t,D) = \log \frac{N}{\#_t}$$

where:
N = Total number of documents
$\#_t$ = Number of documents that contain the term t

TF-IDF: Now, we combine the term frequency and inverse document frequency and multiply them to find the TF-IDF.

$$TF\text{-}IDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$$

Word Embeddings: The word (keras) embedding layer is a kind of embedding approach that is generated during training. The main principle of this word embedding approach is that Bangla words are converted into a numeric value using a tokenizer and then fed into Bi-LSTM and CNN models

Contextual Embedding: Like Keras, Bangla words are tokenized and passed into pre-trained models (BanglaBERT and mBERT) that generate dynamic, context-aware embeddings. These models consider surrounding words to create meaningful representations and are fine-tuned on emotion-labeled data to classify emotions in Bangla text.

The Mel Frequency Cepstral Coefficients (MFCC) technique was used to extract the features from the Bangla speech data. The description of the MFCC features of 13 coefficients is shown in Table II.

TABLE II
DESCRIPTION OF THE MFCC FEATURES

MFCC	Feature Name	Description
0	Energy/Intercept	Signal energy or loudness
1	Spectral Slope	Spectrum tilt
2	Spectral Curvature	Spectrum curvature
3	Spectral Sharpness	Sharpness of spectral peaks
4	High Frequency	High-frequency dominance
5	Spectral Flatness	Smoothness of spectrum
6	Spectral Roll-off	95% energy frequency
7	Spectral Centroid	Spectrum center of mass
8	Spectral Spread	Spectrum spread
9	Spectral Flux	Rate of spectral change
10	Spectral Entropy	Spectrum randomness
11	Spectral Contrast	Peaks vs valleys difference
12	Spectral Roughness	Spectrum irregularities

This technique is mainly applied to extract the features, analyze the sound, and generate coefficient values (13) for each speech clip.

In Fig. 3, the MFCC coefficients of the 'Angry' class sample capture various speech characteristics of the speaker. The first few coefficients (MFCC0, MFCC1, MFCC2) represent overall energy and spectral slope, while higher values capture finer spectral details. Negative values suggest variations in sharpness (MFCC3), brightness(MFCC4), and irregularities

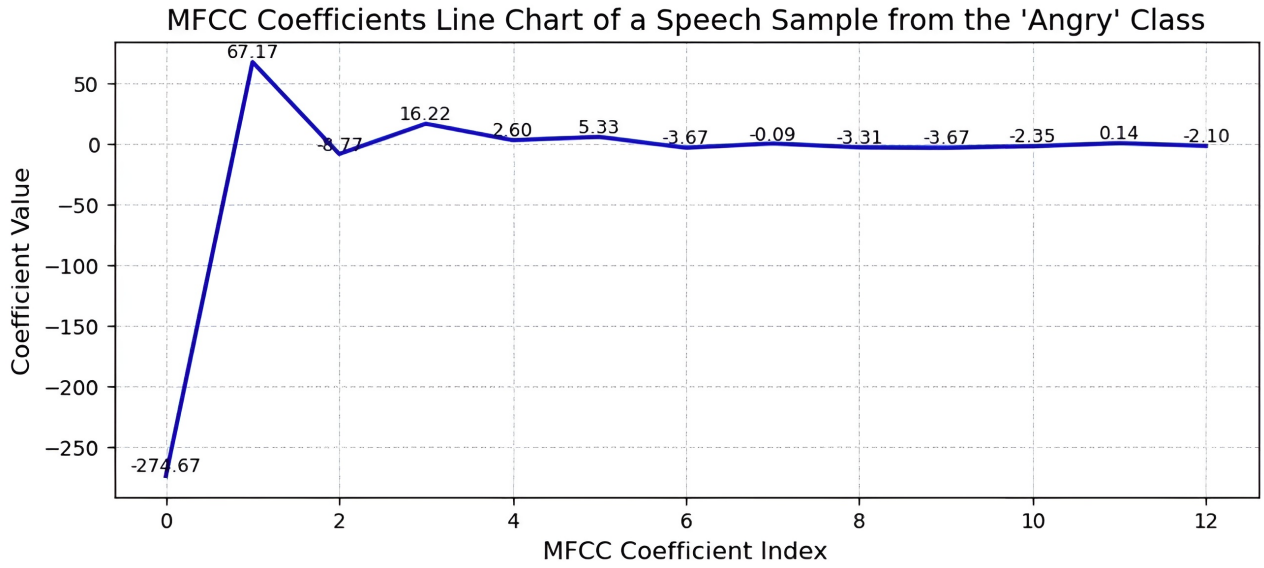

Fig. 3. Line Chart of MFCC Coefficients for a Speech Sample from the 'Angry' Emotion Class

(MFCC12) in the sound. The range of MFCC coefficients mainly highlights the dynamic changes in the 'Angry' class speech sample. This provides an effective representation of the tone and characteristics for emotion analysis.

D. Applied Models

For Bangla text-based emotion classification, several machine learning models have been applied, such as Support Vector Machine (SVM), Random Forest (RF), Multinomial Naive Bayes (MNB), deep learning models such as Bi-LSTM, CNN models, and BERT-based models such as BanglaBERT and mBERT. For Bangla speech-based emotion recognition, machine learning models such as Extreme Gradient Boosting (XGBoost), and Random Forest, deep learning models such as BiLSTM, CNN, a hybrid model CNN+BiLSTM, and a transformer-based model like Conformer have been applied.

E. Multimodal Late (Decision) Fusion

Multimodal fusion is the process of aggregating information or features from multiple modalities to handle several dimensions of data at a time.

There are two primary types of multimodal fusion: Early fusion and Late fusion. In early fusion, we combined the features from the different modalities at a time at the input level before supplying them to a model. As in our work, text and speech datasets are not properly aligned. Therefore, a proper feature-fusion result might not be achievable. In the late fusion, the multimodal model combines predictions from multiple models (from different modalities) after each model has independently classified the input. Finally, we chose the late fusion technique in our cases. The complete procedure is shown in Algorithm 1. Late fusion operates using confidence-based selection combined with a weighted average. The weights are assigned according to the performance of the models. The confidence score is generated according to emotion prediction by models using Softmax activation.

Algorithm 1 Multimodal Decision Level Fusion for Emotion Recognition

Input: text_input, speech_input

Output: Final_Emotion

Load Text_Model and Speech_Model
$T\_Pred \leftarrow$ Text_Model.predict(text_input)
$S\_Pred \leftarrow$ Speech_Model.predict(speech_input)
$w\_text, w\_speech \leftarrow$ Get_Model_Weights()
Initialize Final_Scores as an empty dictionary
for each emotion e in (T_Pred, S_Pred) **do**
    $Final\_Scores[e] \leftarrow (w\_text \times T\_Pred[e]) + (w\_speech \times S\_Pred[e])$
end for
$Final\_Emotion \leftarrow$ argmax$(Final_Scores)$
Return Final_Emotion

F. Model Evaluation

Several models have been used for emotion classification and recognition. Various evaluation metrics, such as accuracy, recall, precision, F1-score, and confusion matrix considered to observe and evaluate the performance of all these applied models in our study.

IV. Experimental Result and Analysis

Various models are applied to classify and recognize emotion from Bangla text and speech using multimodal fusion. This section covers the experimental findings of these models.

A. Performance of the Models used for the Bangla Text-based Emotion Classification

For text-based emotion classification, several machine learning models such as Support Vector Machine(SVM),

Multinomial Naive Bayes (MNB), and Random Forest(RF) models have been used. Additionally, deep learning models such as Bidirectional LSTM and CNN models (use 10 epochs), and BERT-based models such as BanglaBERT and mBERT models(use 10 epochs), have also been used. Table III compares the performance of all applied models that are used for the Bangla text-based emotion classification. The macro average is considered for the precision, recall, and F1-score.

TABLE III
MODEL'S PERFORMANCE FOR BANGLA TEXT-BASED EMOTION CLASSIFICATION

Model	Precision	Recall	F1-Score	Accuracy
SVM	0.45	0.46	0.46	0.46
MNB	0.42	0.42	0.42	0.42
RF	0.60	0.60	0.60	0.60
Bi-LSTM	0.57	0.56	0.56	0.56
CNN	0.61	0.60	0.60	0.59
BanglaBERT	0.65	0.64	0.64	**0.64**
mBERT	0.61	0.59	0.59	0.59

From Table III, it is evident that the BanglaBERT model performs fairly well, achieving an accuracy score of 64% in text-based emotion classification. In contrast, machine learning models, such as SVM, MNB, and deep learning models such as BiLSTM, CNN, and mBERT models, exhibit lower performance for the Bangla text-based emotion class. Next, the performance of the BanglaBERT model for text-based emotion classification is discussed in detail.

1) BanglaBERT Model's Performance for Text-based Emotion Classification: The BanglaBERT model performs best among the applied models for classifying emotions from the Bangla text. We have used the BanglaBERT model to classify emotions based on Bangla text. In the next Fig. 4, we have described the general procedure of the BanglaBERT model for text-based emotion classification.

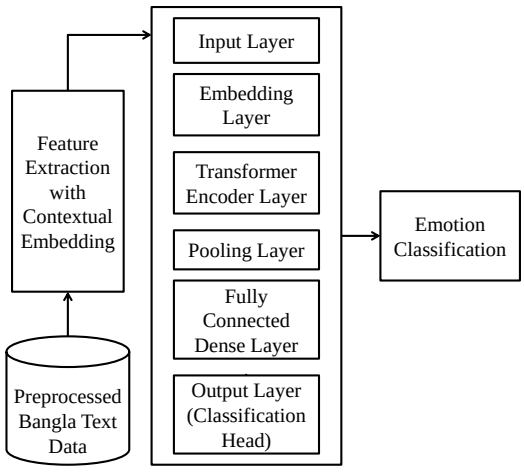

Fig. 4. BanglaBERT Model For Bangla Text Emotion Classification

1. **Preprocessing and Feature Extraction:** The entire process starts with pre-processed Bangla text. Then, contextual embedding is applied, which is a dynamic word representation format that allows capturing the meaning of Bangla words based on their surrounding context.
2. **Input Phase:** The input layer takes the tokenized:
   - Input ID, which represents the sequence of words, and
   - Attention mask specifies which Bangla text tokens should be considered and which should be avoided.
3. **Embedding Layer:** The embedding layer then takes the tokenized Bangla text and converts it into dense vectors. This layer also uses the pre-trained word embedding of the Bangla text collection.
4. **Transformer Encoder Layer:** In this step, the Multi-head self-attention mechanism allows the model to focus on several parts of the input sequence simultaneously, making the model more contextually understandable. There is also a feedforward layer that processes the output of the attention mechanism and a normalization layer for better learning and model convergence.
5. **Pooling Layer:** The pooling layer is used as the summarizer for the input sequence, which is then fed into the classification head.
6. **Dense Layer:** The transformer output is passed into a dense layer of fully connected neurons. This layer converts the feature vector of the input sequence into a well-structured form. A ReLU activation function is also used.
7. **Output Layer:** Finally, in the output layer, softmax activation is used to predicts the emotion class of the Bangla text by producing the probability score for each class.

Now, we first set the hyperparameter settings for the model configuration in Table IV:

TABLE IV
HYPERPARAMETERS SETTING FOR BANGLABERT MODEL

Hyperparameters	**Values**
Pre-trained Model	csebuetnlp/banglabert
Hidden Layer Size	768
Number of Transformer Layers	12
Max Sequence Length	128
Optimizer	Adam
Learning Rate	5×10^{-5}
Batch Size	32
Dropout Rate	0.1
Number of Epochs	10
Attention Heads	12
Weight Decay	0.01
Classification Task Output	6

The BanglaBERT model is now evaluated using the confusion matrix and classification report. Firstly, the confusion matrix of the BanglaBERT model is presented here for emotion classification.

In Fig. 5, the confusion matrix shows that the BanglaBERT model predicts 382 accurate predictions (considering all true positives) out of 600 test instances. It is observed that the

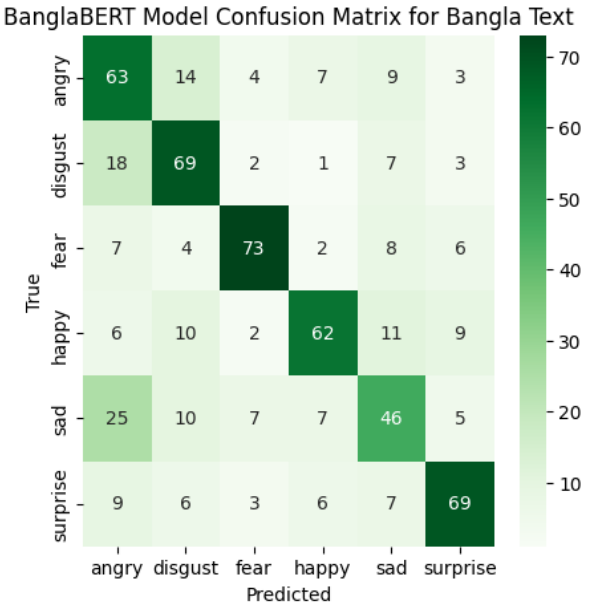

Fig. 5. Confusion Matrix for BanglaBERT Model

BanglaBERT model predicts the highest number of correct instances for the 'fear' class (73), followed by 69 each for both the 'disgust' and 'surprise' classes, and the lowest for the 'sad' class, with 46 correct predictions. Next, the classification report of the BanglaBERT model for text-based emotion classification is given in Table V.

TABLE V
CLASSIFICATION REPORT FOR BANGLABERT MODEL FOR TEXT-BASED EMOTION CLASSIFICATION

Class	Precision	Recall	F1-score	Support
Angry	0.49	0.63	0.55	100
Disgust	0.61	0.69	0.65	100
Fear	0.80	0.73	0.76	100
Happy	0.73	0.62	0.67	100
Sad	0.52	0.46	0.49	100
Surprise	0.73	0.69	0.71	100
Accuracy	-	-	**0.64**	600
Macro Avg	0.65	0.64	0.64	600
Weighted Avg	0.65	0.64	0.64	600

From the classification report, it is noticed that the BanglaBERT model's accuracy is better than any other applied models so far for Bangla text-based emotion classification. It is a strong transformer model that is trained on millions of Bangla words. Therefore, it represents a good performance for the Bangla text-based emotion classification. Upon analyzing the class-wise performance for the BanglaBERT model, it is found that the model demonstrates its best performance for the 'fear' class, with a high F1-score of 0.76, which reflects a good level of balance between precision and recall for that class. In contrast, the 'angry' and 'sad' classes have a lower F1-score of 0.55 and 0.49, respectively, which hurts the model's performance. The accuracy of the BanglaBERT model is 64% for the Bangla text-based emotion classification.

B. Performance of the Models used for the Bangla Speech-based Emotion Recognition

Several machine learning models, such as the Random Forest (RF) model and Extreme Gradient Boosting (XGBoost) model, have been used for speech-based emotion recognition. Moreover, deep learning models such as Bidirectional LSTM and CNN models (30 epochs), CNN+BiLSTM (30 epochs) model, and transformer-based models such as Conformer model (30 epochs) have also been used. Table V compares the performance of models that are used for speech emotion recognition. The macro average is considered for the precision, recall, and F1-score. Table VI shows that the XGBoost model performs best, achieving an accuracy of 72% for the recognition of emotions based on the Bangla speech. The hybrid model (CNN + Bi-LSTM) also shows a fair performance for recognition with an accuracy of 66%. The performance of the XGBoost model is now discussed for speech-based emotion recognition.

TABLE VI
MODEL'S PERFORMANCE FOR BANGLA SPEECH-BASED EMOTION RECOGNITION

Model	Precision	Recall	F1-Score	Accuracy
RF	0.56	0.55	0.54	0.55
XGBoost	0.72	0.72	0.72	**0.72**
Bi-LSTM	0.61	0.58	0.58	0.58
CNN	0.59	0.55	0.55	0.56
CNN + BiLSTM	0.67	0.66	0.66	0.66
Conformer	0.62	0.62	0.61	0.62

1) Extreme Gradient Boosting (XGBoost) Model's Performance for Speech-based Emotion Recognition: The XGBoost model performs best among the applied models for speech-based emotion recognition. The methodology of this model is shown below in Fig. 6.

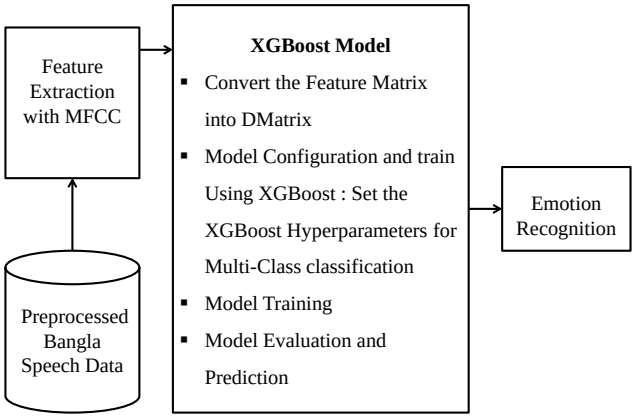

Fig. 6. XGBoost Method for Bangla Speech Emotion Recognition

To recognize the emotion from Bangla Speech using the XGBoost model, we have the following steps:

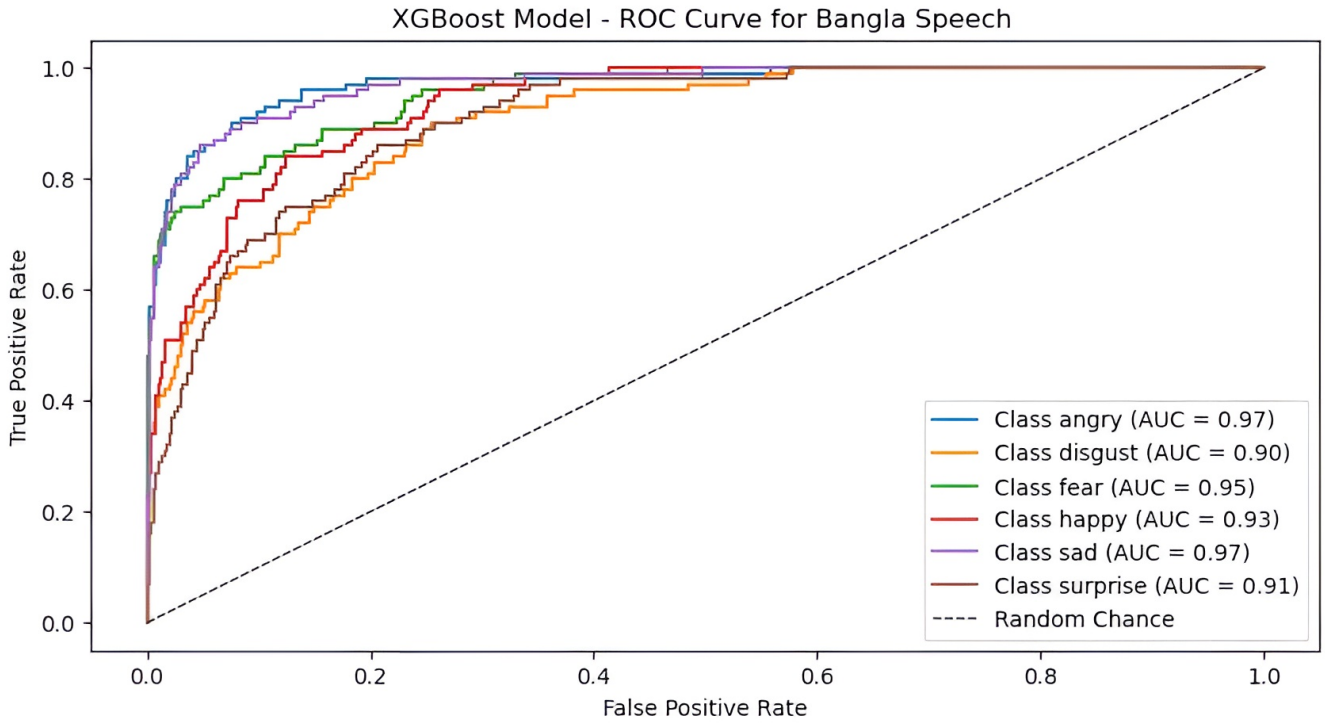

Fig. 7. XGBoost - ROC Curve for Bangla Speech

1. Firstly, we started with preprocessed Bangla speech data.
2. Then, we divided the dataset for training, testing, and validation.
3. After splitting, we convert the feature matrix to DMatrix and configure the model.
4. For configuring the model, we set the XGBoost parameters, such as the objective, num_class, max_depth, learning rate, etc.
5. We also set the evaluation metrics (e.g., mlogloss) and other parameters like the number of estimators, tree depth, and subsampling.
6. Then, we trained the model, tested the model on new speech data, and evaluated the performance using evaluation metrics.

The XGBoost model is now evaluated using the ROC curve, confusion matrix, and classification report.

Firstly, the ROC curve is presented in Fig. 7. This Receiver Operating Characteristic (ROC) curve represents the performance of an XGBoost model in classifying different emotions. The AUC score (ranging from 0 to 1) indicates the model's ability to distinguish between emotion classes. The higher AUC value means better classification performance. The XGBoost model performs exceptionally well in recognizing the 'angry', 'sad', and 'fear' emotions classes in a range of AUC (0.95- 0.97). The lowest performance is observed for the 'disgust' class (AUC = 0.90), which still indicates a strong classification ability. The curves are well above the diagonal line, which indicates that the XGBoost model significantly outperforms random guessing.

In Fig. 8, the confusion matrix shows that the XGBoost model predicts 431 correct predictions (considering all true positives) out of 600 test instances. This model shows a strong performance for speech recognition. Considering the

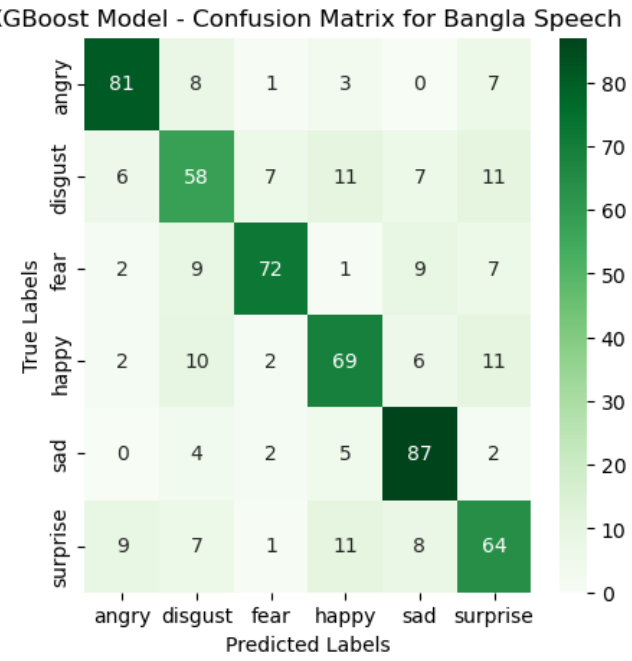

Fig. 8. XGBoost - Confusion Matrix for Bangla Speech

class-wise performance, it is noticed that the XGBoost model predicts the highest number of instances (87) correctly for the 'sad' class among all the classes, and the lowest 58 for the 'disgust' class. Finally, the classification report of the XGBoost model is given in Table VII for speech-based emotion recognition. Analyzing the class-wise performance, it is observed that this model demonstrates its best performance for the 'angry', 'fear', and 'sad' classes, with F1-scores of 0.81,0.78, and 0.80, respectively. This indicates a good level of balance between precision and recall for those classes.

TABLE VII
CLASSIFICATION REPORT FOR XGBOOST MODEL FOR SPEECH-BASED EMOTION RECOGNITION

Class	Precision	Recall	F1-Score	Support
angry	0.81	0.81	0.81	100
disgust	0.60	0.58	0.59	100
fear	0.85	0.72	0.78	100
happy	0.69	0.69	0.69	100
sad	0.74	0.87	0.80	100
surprise	0.63	0.64	0.63	100
Accuracy			**0.72**	600
Macro Avg	0.72	0.72	0.72	600
Weighted Avg	0.72	0.72	0.72	600

C. Performance of the Models Used for the Multimodal Decision Level Fusion

We have planned to apply decision-level fusion, which means we will combine the outputs of multiple models by aggregating their predictions. In decision-level fusion, we have used the weighted-based fusion with the confidence score of the emotion prediction for each class. The overall process is shown by an example in Fig. 9.

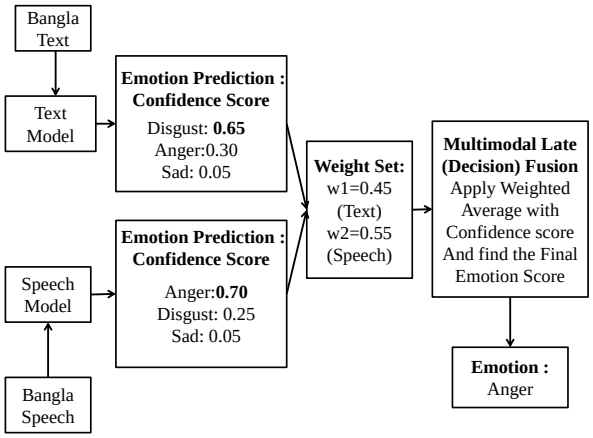

Fig. 9. A Scenario of the Multimodal Emotion Recognition Task.

We decided to apply two multimodal decision-level fusions using the combination of the best two models of text and speech. These combinations are the BanglaBERT + XGBoost model and the BanglaBERT + (CNN+BiLSTM) model. We do not consider the Random Forest model here as its accuracy score is less than the BanglaBERT, XGBoost, and CNN+BiLSTM models. Table VIII shows the performance comparison of the multimodal approaches for emotion recognition. Weights were set as $w_1 = 0.4$ and $w_2 = 0.6$ for the text and speech models, respectively, since the XGBoost model for speech achieved higher accuracy (72%) compared to the BanglaBERT model for text (64%). Multimodal decision-level fusion has also been applied using the combination of the BanglaBERT and CNN+BiLSTM model, where the confidence scores of the predicted emotion classes were combined through weighted averaging for each model (text and speech model). For this, weights are set for

TABLE VIII
PERFORMANCE COMPARISON OF THE MULTIMODAL APPROACHES FOR EMOTION RECOGNITION

Model	Precision	Recall	F1-Score	Accuracy
BanglaBERT + XGBoost	0.73	0.70	0.69	**0.70**
BanglaBERT + (CNN + Bi-LSTM)	0.72	0.67	0.66	0.67

both the text and speech model 0.5 ($w_1 = w_2 = 0.5$) as the accuracy of the text model (64%) and speech model (66%) is close. From Table VIII, it is observed that the BanglaBERT +XGBoost model achieves an accuracy of 70% for the Bangla emotion recognition. This multimodal approach outperforms the accuracy of the BanglaBERT model (64%) in text-based emotion classification, but it falls short of the accuracy of the XGBoost model (72%) in speech-based emotion recognition. If we observe the performance of the BanglaBERT+ (CNN+BiLSTM) multimodal, we see that this multimodal accuracy is 67% which is less than the BanglaBERT +XGBoost model performance.

Fig. 10 shows a comparison of the unimodal (Best three models) and multimodal performance for emotion recognition.

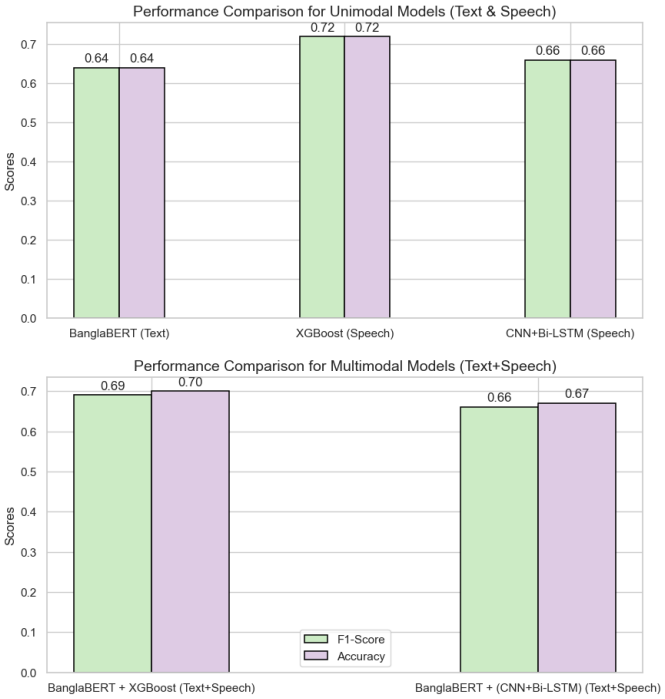

Fig. 10. Performance Comparison of the Unimodal (Best 3) and Multimodal for Emotion Recognition

We also observe that the multimodal BanglaBERT+ (CNN+BiLSTM) outperforms both the accuracy of the BanglaBERT model (64%) and the CNN+BiLSTM model(66%).

The confusion matrix and classification report of the best multimodal (BanglaBERT + XGBoost) are given below in Fig.11 and Table IX, respectively. In Fig. 11, the confusion matrix reveals that this multimodal correctly predicts 420

out of 600 instances(considering all true positives) in the test dataset, which indicates a satisfactory performance level for Bangla emotion recognition. The multimodal model predicts the highest number of correct instances for the 'angry', 'fear', and 'disgust' classes, with 94, 82, and 79 instances correctly classified, respectively. In contrast, it struggles in capturing the pattern of the 'sad' and 'surprise' classes, which adversely hurts the model's performance.

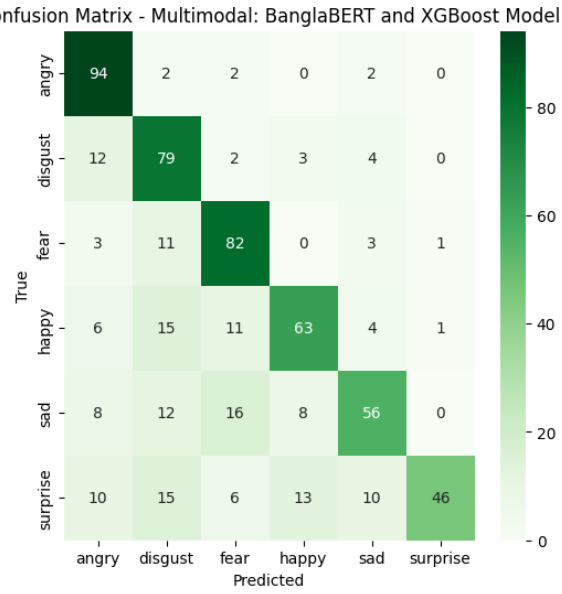

Fig. 11. Confusion Matrix of the Multimodal: BanglaBERT + XGBoost

In Table IX, the multimodal classification report indicates effective performance in the six emotion classes.

TABLE IX
CLASSIFICATION REPORT FOR MULTIMODAL: BANGLABERT AND XGBOOST MODELS

Class	Precision	Recall	F1-score	Support
Angry	0.71	0.94	0.81	100
Disgust	0.59	0.79	0.68	100
Fear	0.69	0.82	0.75	100
Happy	0.72	0.63	0.67	100
Sad	0.71	0.56	0.63	100
Surprise	0.96	0.46	0.62	100
Accuracy	-	-	**0.70**	600
Macro Avg	0.73	0.70	0.69	600
Weighted Avg	0.73	0.70	0.69	600

The model demonstrates its best performance for the 'angry', 'fear' classes, achieving F1 scores of 0.81 and 0.75, respectively, indicating a fair level of precision and recall for those classes. In contrast, lower F1-scores of the 'sad' and 'surprise' classes hurt the performance of the model. The overall accuracy of this multimodal approach for the recognition of Bangla emotions is 70%.

D. Performance Comparison with Existing Research for Multimodal Fusion

In this subsection, we compare the performance of our proposed approach with existing research. While comparing our study we consider some parameters.

In Table X, 'Modalities' refers to the types of data used for multimodal fusion, 'Tasks' indicates the goal of these studies, 'Data' represents the number of samples used for these studies, 'Applied Model' parameter indicates what the models they have used for finding the best result, 'Language' specifies the language of the text or speech have used for these studies and finally, the 'F1-score' is considered for evaluating the performance of these studies.

1) Multimodal Fusion Using Text and Image Modalities: For emotion recognition and sentiment analysis through image-text modalities, these studies [18], [22], [21] provide good weighted F1 scores of 0.75,0.64,0.71, respectively. For binary detection of hate speech, [20] provides strong performance with a weighted F1 score of 0.81. While comparing this information with our study, we could understand that image-text modalities have a greater tendency compared to speech-text modalities for providing accurate results with less number of data samples of the Bangla language for decision fusion. In our research, we have used 12000 samples to recognize emotions through multimodal decision-level fusion. These studies [18], [20], [22], [21] used 4660, 4500, 4372, and 4368 data samples, respectively.

2) Multimodal Fusion Using Text and Image Modalities: This research approach[26] is different from ours, as we have used decision (late) level fusion for emotion recognition, which is a forward and independent step from their approach (mid-level fusion). We concluded that with 12000 samples, we have found a more accurate result from them (0.66) with an F1 score of 0.69. However, this research [25] provides a better result for the decision-level fusion technique than our research, since we have found an F1 score of 0.69 for emotion recognition using text and speech modalities. Note that the study worked on more data(19700) than our study(12000), as the English language already has huge available resources, whereas Bangla has very limited resources. These reasons might affect our multimodal performance if we compare it with their research.

3) Multimodal Fusion Using Audio/Speech, Video, and Text Modalities: This research approach[19] is different from our fusion technique, since we have applied the fusion of the decision level (late) for emotion recognition, where we combine the predictions from separate models and use the weighted average of the final prediction. However, their approach (feature-level fusion) combines the features of the model after the feature extraction but before feeding those features into the model. We found a better result with (F1 score =0.69), which is 0.64 for theirs. It is worth to mention that they produce good results using the Early(Feature) fusion technique with a very limited dataset (1002) whereas our study had more samples (12000 samples)

V. CONCLUSION

In this section, we provide a concise summary of the multimodal emotion recognition tasks conducted in

TABLE X
COMPARISON OF RESEARCH STUDIES IN MULTIMODAL APPROACH

Research	Modalities	Tasks	Data	Applied Model	Language	F1-score
[18]	Image, Text	Emotion Recognition	4660	ResNet50 (Image), BiLSTM (Text)	Bangla	0.75
[19]	Audio, Video, Text	Emotion Recognition	1002	YamNet (Audio), DeepFace (Video), BanglaBERT (Text)	Bangla	0.64
[20]	Image, Text	Hate Speech Detection	4500	XLM-RoBERTa (Text), DenseNet-161 (Image)	Bangla	0.83
[21]	Image, Text	Sentiment Analysis	4372	ResNet50 (Image), BanglishBERT (Text)	Bangla, English	0.71
[22]	Image, Text	Sentiment Analysis	4368	ResNet50 (Image), BiLSTM (Text)	Bangla	0.64
[26]	Text, Speech	Emotion Recognition	13708	RoBERTa (Text), OpenL3 (Audio)	English	0.66
[25]	Text, Speech	Sentiment Analysis	19700	BERT (Text), CNN (Audio)	English	0.88
Proposed Method	Text, Speech	Emotion Recognition	12000	BanglaBERT (Text), XGBoost (Speech)	Bangla	0.69

our research. Our study proposed a multimodal emotion recognition approach using text and speech modalities for limited-resource languages like Bangla. We have applied several machine learning (SVM, MNB, Random Forest, XGBoost), Deep Learning (Bi-LSTM, CNN, Conformer), Hybrid Model (CNN+BiLSTM), BERT-based models (BanglaBERT, mBERT) for emotion classification and recognition. Among the text models, the BanglaBERT model has performed the best with an accuracy of 64% and the Random Forest model also performed well with an accuracy of 60%. For the speech model, XGBoost and the hybrid model (CNN+BiLSTM) have achieved an accuracy of 72% and 66% respectively. Finally, we have chosen the BanglaBERT, XGBoost, and CNN+BiLSTM models for multimodal emotion recognition. We have found that the BanglaBERT + XGBoost model has performed best with an accuracy of 70%, while the combination of the BanglaBERT and (CNN+BiLSTM) models has performed with an accuracy of 67%. In this research, we have only used two modalities: text and speech. In future, our plan is to add more modalities(e.g., image, video) for the multimodal emotion recognition fusion tasks, which may add more diversity to this field. In addition, our goal is to integrate our proposed multimodal framework into mobile and web-based platforms for real-world applications. Moreover, as there is a lack of a multimodal emotion dataset for Bangla text and speech modalities, We are planning to make a balanced dataset for the scientific community of this field.

REFERENCES

[1] A. Das, M. M. Hoque, O. Sharif, M. A. A. Dewan, and N. Siddique, "Temox: Classification of textual emotion using ensemble of transformers," *IEEE Access*, vol. 11, pp. 109 803–109 818, 2023.

[2] R. Haque, M. B. Islam, P. B.D, K. G. Khushbu, S. Rahman, A. U. Rahman, M. H. Hossen, and T. I. Rohan, "Bengali emotion classification using hybrid deep neural network," in *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, 2023, pp. 1–7.

[3] W. A. Aripin and S. Huda, "Multichannel convolutional neural network model to improve compound emotional text classification performance," *IAENG International Journal of Computer Science*, vol. 50, no. 3, pp. 866–874, 2023.

[4] 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," *IEEE Access*, vol. 10, pp. 564–578, 2022.

[5] M. J. Hasan, M. S. Hossain, S. N. Hassan, M. Al-Amin, M. N. Rahaman, and M. A. Pranjol, "Bengali speech emotion recognition: A hybrid approach using b-lstm," in *2022 4th International Conference on Electrical, Computer Telecommunication Engineering (ICECTE)*, 2022, pp. 1–7.

[6] M. S. U. Sourav, H. Wang, M. S. Mahmud, and H. Zheng, "Transformer-based text classification on unified bangla multi-class emotion corpus," 2023. [Online]. Available: https://arxiv.org/abs/2210.06405

[7] M. R. Faisal, A. M. Shifa, M. H. Rahman, M. A. Uddin, and R. M. Rahman, "Bengali banglish: A monolingual dataset for emotion detection in linguistically diverse contexts," *Data in Brief*, vol. 55, p. 110760, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340924007261

[8] A. Kabir, A. Roy, and Z. Taheri, "BEmoLexBERT: A hybrid model for multilabel textual emotion classification in Bangla by combining transformers with lexicon features," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, F. Alam, S. Kar, S. A. Chowdhury, F. Sadeque, and R. Amin, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 56–61. [Online]. Available: https://aclanthology.org/2023.banglalp-1.7/

[9] T. Ghosh, M. H. A. Banna, M. J. A. Nahian, M. N. Uddin, M. S. Kaiser, and M. Mahmud, "An attention-based hybrid architecture with explainability for depressive social media text detection in bangla," *Expert Systems with Applications*, vol. 213, p. 119007, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422020255

[10] A. Anika Namey, K. Akter, M. A. Hossain, and M. Ali Akber Dewan, "Cochleaspecnet: An attention-based dual branch hybrid cnn-gru network for speech emotion recognition using cochleagram and spectrogram," *IEEE Access*, vol. 12, pp. 190 760–190 774, 2024.

[11] A. C. Shruti, R. H. Rifat, M. Kamal, and M. G. R. Alam, "A comparative study on bengali speech sentiment analysis based on audio data," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2023, pp. 219–226.

[12] S. Aziz, N. H. Arif, S. Ahbab, S. Ahmed, T. Ahmed, and M. H. Kabir, "Improved speech emotion recognition in bengali language using deep learning," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023, pp. 1–6.

[13] S. M. Maheen, M. R. Faisal, M. R. Rahman, and M. S. Karim, "Alternative non-bert model choices for the textual classification in low-resource languages and environments," *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250390775

[14] S. K. Banshal, S. Das, S. A. Shammi, and N. R. Chakraborty, "Monovab : An annotated corpus for bangla multi-label emotion detection," *ArXiv*, vol. abs/2309.15670, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263152270

[15] T. Parvin and M. M. Hoque, "An ensemble technique

to classify multi-class textual emotion," *Procedia Computer Science*, vol. 193, pp. 72–81, 2021, 10th International Young Scientists Conference in Computational Science, YSC2021, 28 June – 2 July, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921020494

[16] A. C. Shruti, R. H. Rifat, M. Kamal, and M. G. R. Alam, "A comparative study on bengali speech sentiment analysis based on audio data," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2023, pp. 219–226.

[17] M. M. Billah, M. L. Sarker, M. A. H. Akhand, and M. A. S. Kamal, "Emotion recognition with intensity level from bangla speech using feature transformation and cascaded deep learning model," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, 2024. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.0150460

[18] Z. S. Taheri, A. C. Roy, and A. Kabir, "Bemofusionnet: A deep learning approach for multimodal emotion classification in bangla social media posts," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023, pp. 1–6.

[19] A. Das, M. S. Sarma, M. M. Hoque, N. Siddique, and M. A. A. Dewan, "Avater: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition," *Sensors*, vol. 24, no. 18, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/18/5862

[20] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi, "Multimodal hate speech detection from bengali memes and texts," in *Speech and Language Technologies for Low-Resource Languages*, A. K. M, B. R. Chakravarthi, B. B, C. O'Riordan, H. Murthy, T. Durairaj, and T. Mandl, Eds. Cham: Springer International Publishing, 2023, pp. 293–308.

[21] K. T. Elahi, T. Binte Rahman, S. Shahriar, S. Sarker, S. K. Saha Joy, and F. Muhammad Shah, "Explainable multimodal sentiment analysis on bengali memes," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023, pp. 1–6.

[22] E. Hossain, O. Sharif, and M. M. Hoque, "MemoSen: A multimodal dataset for sentiment analysis of memes," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1542–1554. [Online]. Available: https://aclanthology.org/2022.lrec-1.165/

[23] S. Biswas, K. Young, and J. Griffith, "Exploring multimodal features for sentiment classification of social media data," in *Proceedings of International Conference on Information Technology and Applications*, A. Ullah, S. Anwar, D. Calandra, and R. Di Fuccio, Eds. Singapore: Springer Nature Singapore, 2024, pp. 527–537.

[24] T. Niu, S. Zhu, L. Pang, and A. El-Saddik, "Sentiment analysis on multi-view social data," in *MultiMedia Modeling*, 2016, p. 15–27.

[25] H. Srivastava, S. Sunil, K. Shantha Kumari, and P. Kanmani, "Multimodal sentiment analysis using text and audio for customer support centers," in *Proceedings of ICACTCE'23 — The International Conference on Advances in Communication Technology and Computer Engineering*, C. Iwendi, Z. Boulouard, and N. Kryvinska, Eds. Cham: Springer Nature Switzerland, 2023, pp. 491–506.

[26] J. Luo, H. Phan, and J. Reiss, "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[27] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050/

[28] J. Salas-Cáceres, J. Lorenzo-Navarro, D. Freire-Obregón, and M. Castrillón-Santana, "Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics," *Multimedia Tools and Applications*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272877641

[29] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, vol. 13, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:21704094

[30] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," [Online] Available: http://kahlan.eps.surrey.ac.uk/savee/, 2015.

[31] S. Haq and P. Jackson, "Multimodal emotion recognition," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. IGI Global, 2010, pp. 398–423.

[32] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[33] T. Meng, Y. Shou, W. Ai, N. Yin, and K. Li, "Deep imbalanced learning for multimodal emotion recognition in conversations," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6472–6487, 2024.

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6

[35] S. Mustavi Maheen, M. Rahman Faisal, M. Rafakat Rahman, and M. S. Karim, "Alternative non-BERT model choices for the textual classification in low-resource languages and environments," in *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, C. Cherry, A. Fan, G. Foster, G. R. Haffari, S. Khadivi, N. V. Peng, X. Ren, E. Shareghi, and S. Swayamdipta, Eds. Hybrid: Association for Computational Linguistics, Jul. 2022, pp. 192–202. [Online]. Available: https://aclanthology.org/2022.deeplo-1.20/

[36] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *PLoS ONE*, vol. 16, no. 4, p. e0250173, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0250173

[37] T. Tokunaga and M. Iwayama, "Text categorization based on weighted inverse document frequency," 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:18257943