A Machine Learning Approach for Emotion Classification in Bengali Speech

Md. Rakibul Islam[1], Amatul Bushra Akhi[2], Farzana Akter[3], Md Wasiul Rashid[4], Ambia Islam Rumu[5], Munira Akter Lata[6], Md. Ashrafuzzaman[7]

Dept. of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh[1, 2, 4]
Dept. of IoT and Robotics Engineering, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh[3]
Dept. of English, Daffodil International University, Dhaka, Bangladesh[5]
Dept. of Educational Technology, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh[6, 7]

***Abstract*—In this research work, we have presented a machine learning strategy for Bengali speech emotion categorization with a focus on Mel-frequency cepstral coefficients (MFCC) as features. The commonly utilized method of MFCC in speech processing has proved effective in obtaining crucial phoneme-specific data. This paper analyzes the efficacy of four machine learning algorithms: Random Forest, XGBoost, CatBoost, and Gradient Boosting, and tackles the paucity of research on emotion categorization in non-English languages, particularly Bengali. With CatBoost obtaining the greatest accuracy of 82.85%, Gradient Boosting coming in second with 81.19%, XGBoost coming in third with 80.03%, and Random Forest coming in fourth with 80.01%, experimental evaluation shows encouraging outcomes. MFCC features improve classification precision and offer insightful information on the distinctive qualities of emotions expressed in Bengali speech. By demonstrating how well MFCC characteristics can identify emotions in Bengali speech, this study advances the field of emotion classification. Future research can investigate more sophisticated feature extraction methods, look into how temporal dynamics are incorporated into emotion classification models, and investigate practical uses for emotion detection systems in Bengali speech. This study advances our knowledge of emotion classification and paves the way for more effective emotion identification systems in Bengali speech by utilizing MFCC and machine learning techniques. Our work addresses the need for thorough and efficient techniques to recognize and classify emotions in speech signals in the context of emotion categorization. Understanding emotions is essential for many applications, as they are a basic component of human communication. By investigating cutting-edge strategies that show promise for enhancing the precision and effectiveness of emotion recognition, this study advances the field of emotion classification.**

Keywords—XgBoost*; *gradient boosting; CatBoost; random forest*; *MFCC

I. Introduction

Emotions significantly impact people's attitudes and decision-making processes in various everyday activities, making them a crucial aspect of communication among individuals. Emotional understanding promotes mutual understanding and contributes to a fulfilling social life [1]. In today's technologically advanced world, where voice commands and manual instructions are exchanged between humans and machines, the ability of artificial intelligence or machine learning to identify and assess someone's emotional state from their voice has emerged as a pressing challenge [2][3]. There is little study on emotion classification in non-English languages like Bengali, which presents particular difficulties for speech analysis. Systems currently in use frequently struggle with linguistic diversity and cultural subtleties. In order to solve these problems, our research advances the area by introducing a machine learning technique that improves the accuracy of emotion recognition. Our objective is to enhance emotional intelligence and human-machine communication, especially in underrepresented languages.

The contributions of this paper are twofold. Firstly, we provide a comprehensive investigation into the application of machine learning algorithms for emotion classification in Bengali speech, which is a relatively unexplored area of research. This study addresses the need for emotion recognition systems tailored to Bengali, which can enhance communication and human-machine interactions within Bengali-speaking populations. Secondly, we offer insights into the performance of Random Forest, XGBoost, CatBoost, and Gradient Boosting models in the context of emotion classification in Bengali speech. By comparing and evaluating the results of these models, we aim to identify the most effective approach for accurate emotion recognition.

Overall, our research aims to contribute to the advancement of speech emotion recognition, specifically in the Bengali language, and pave the way for improved human-machine interactions, affective computing applications, and speech therapy within Bengali-speaking communities.

This paper is organized logically and cogently. It starts with an introduction that sets the scene by supplying the backdrop and rationale for the research, outlining the goals of the study in detail, and highlighting its importance. The literature review section that follows conducts a detailed examination of earlier work on speech emotion identification, discusses several methods for categorizing emotions, and identifies the research gap in Bengali speech emotion categorization. The basis for the following sections is laid by this. Details regarding the Bengali speech dataset used in the study are provided in the dataset and data collecting section, along with information on the methods used for data collection and participant selection, as well as the preprocessing procedures performed on the speech data. The feature extraction section emphasizes the importance of Mel-frequency cepstral coefficients (MFCC) features in the categorization of emotions by emphasizing their extraction from the preprocessed speech data. In the section on machine learning algorithms, the algorithms Random Forest, XGBoost, CatBoost, and Gradient Boosting are introduced and discussed, along with details on their implementation and parameter settings. The experimental setup, evaluation measures, and results of the implemented algorithms are examined in the experimental evaluation section. The discussion part concludes by interpreting the results, outlining any ramifications, and addressing any study limitations.

The motivation section underlines the importance of emotions in human communication and draws attention to the dearth of research on emotion classification in non-English languages, particularly Bengali. The Speech Emotion Recognition (SER) methodology—which includes data collection, preprocessing, feature extraction, model selection, training, evaluation, optimization, and deployment—is the primary topic of discussion in Section IV. With the goal of accurately detecting emotions, enhancing human-machine interactions, and enhancing emotional comprehension, the project focuses on creating a machine learning technique for Bengali speech emotion categorization.

In Section V, we go over the steps involved in classifying emotions in Bengali speech. These include gathering data, analyzing dataset properties, extracting features using MFCC, and applying machine learning algorithms (Random Forest, XGBoost, CatBoost, and Gradient Boosting) to the classification process. The diversity of the dataset, the value of MFCC in capturing emotional information, and the prowess of each classification method in handling complex patterns are all emphasized in this section.

II. Motivation

The need to address the dearth of studies concentrating on non-English languages, notably Bengali, in the field of emotion classification led to the research on "A Machine Learning Approach for Emotion Classification in Bengali Speech". Despite the importance of emotions in human communication, the majority of study so far has focused on English or other commonly used languages. This investigation into the categorization of emotions in Bengali speech tries to close this gap.

The major goal is to create a machine learning model that can recognize emotions in Bengali speech that is accurate and efficient. A useful tool for deciphering and comprehending emotions communicated in Bengali would be provided by such a model. It has applications in areas including sentiment analysis for Bengali-based platforms and services, human-computer interaction, and virtual assistants.

The study also recognizes the diversity of languages and cultures found among Bengali speakers. For effective emotion classification, Bengali's limited linguistic resources, dialectal variances, and pronunciation quirks present special difficulties. In order to create emotion recognition algorithms that are both culturally diverse and contextually pertinent, the project intends to solve these issues.

Studies on the identification of emotions from Bangla speech data are scarce [4],[24]-[27]. 25 MFCCs were suggested by researchers who investigated the optimum number of MFCCs for emotion recognition in speech data in [4]. To categorize user comments on Facebook pages, a deep neural network model based on Gated Recurrent Units was created in [24]. 5126 comments in Bangla were gathered, and these were divided into six categories: incitement, political, religious, hate speech, communal attack, and incitement. The accuracy of the GRU-based model was 70.10%. With an accuracy of 51.33%, a Recurrent Neural Network (RNN) was used in [25] to categorize six emotions in Bangla speech: joy, sadness, anger, surprise, fear, and disgust. [26] used a vocabulary of 500 distinct words using the Gaussian Mixture Model-Hidden Markov Model and Deep Neural Network-Hidden Markov Model to identify emotions in 49 distinct speakers. Using Word Error Rate to gauge model performance, they discovered that the GMM-HMM had 3.96% WER while the DNN-HMM had 5.30% WER. In [27], speech data was subjected to emotion categorization using an ensemble method that used multiple supervised classifiers, with an accuracy rate of 70%. It is clear that the level of accuracy attained by previous studies in audio-based emotion detection and recognition has not been very high.

In conclusion, the research is driven by the need to enhance communication technology, enable applications that can better recognize, interpret, and react to human emotions within the Bengali-speaking community, and advance our understanding of emotions in Bengali speech.

III. Related Work

While there have been extensive research studies conducted in the field of Speech Emotion Recognition (SER) for various languages, particularly English, only limited efforts have been made to establish SER for Bengali (Bangla). In 2018, Rahman et al. introduced a method for Bangla emotion classification using Dynamic Time Warping (DTW) assisted Support Vector Machines (SVM) [5]. The features utilized for classification were the first and second derivatives of Mel-frequency Cepstral Coefficients (MFCC). The proposed system achieved an average accuracy of 86.08% on a small dataset consisting of only 200 words.

N. Kholodna et al. [6] constructed a machine-learning model to automate the detection of emotions from speech, to monitor public emotions. They selected a manually annotated dataset and transformed it into a textual representation using a vectorization technique. Deep learning approaches, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and perceptron models, were employed to identify emotions from the textual data. The classification accuracy of the resulting model was found to be relatively low, with random forest achieving 77%, regression achieving 74%, and the naive Bayesian classifier achieving 73.5%.

Some researchers work with Bengali speech. That study employs pitch and MFCC variables to identify emotions in

Bengali speech [7]. The suggested method produced an accuracy rate of 87.50% on a self-created dataset of Bengali emotional speech, with accuracies of 80.00% for joyful, 75.00% for sad, 85.00% for angry, and 75.00% for neutral emotions. For a thorough evaluation of the study, further information is required.

A. Majeed et al. [8], a system was created by the authors to detect emotions from Roman Urdu text. They constructed a comprehensive corpus comprising 18k sentences, sourced from various domains, and annotated it with six different emotion classes. The authors further employed baseline algorithms such as K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Random Forest on their corpus. The achieved accuracy rate was 69.4%, with an F-measure of 0.69

S. Cunningham et al. [9], the authors introduced an Artificial Neural Network (ANN) approach for predicting emotions in the domain of Music Emotion Recognition. A total of 167 voices were analyzed, and 76 features were extracted from the International Affective Digital Sounds Dataset (IADS). To facilitate model training and evaluation, the audio dataset was partitioned into three segments: training (70%), validation (15%), and testing (15%). During the prediction phase, the ANN model achieved accuracy rates of 64.4% for arousal and 65.4% for valence. These results indicated that the shallow neural network outperformed the regression model in terms of performance.

In our study, we were successful in classifying Bengali speech emotions with a high degree of accuracy. We achieved outstanding results by utilizing cutting-edge machine learning methods including Random Forest, XGBoost, CatBoost, and Gradient Boosting. Our top accuracy score was 82.85%. This outstanding performance demonstrates how well our method works at correctly identifying and categorizing emotions in Bengali speech. These findings help Bengali language emotion recognition technology progress and present prospects for real-world applications in numerous fields.

IV. Speech Emotion Recognition

The goal of speech emotion recognition (SER), a crucial field of study, is to automatically identify and categorize the emotions indicated in speech signals. To process and analyze the voice data properly, the system includes a number of crucial steps. In the beginning, a dataset of labeled speech recordings is gathered, capturing emotions displayed in regulated or naturalistic settings. To improve their quality and retrieve pertinent information, the speech signals go through preprocessing procedures such as segmentation and noise removal. The preprocessed signals are used to extract acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs), energy-based features, and prosodic features. These characteristics act as clear illustrations of the emotional content. When deep learning or machine learning models like Support Vector Machines, Convolutional Neural Networks, or Long Short-Term Memory networks are used for classification, model selection is critical. The models are developed using the labeled data, tuned using hyperparameters, and their performance is assessed. In the end, the trained models can be used to predict emotions from speech data that hasn't been seen, which will help with applications like affective computing, human-computer interaction, and speech therapy. The processes of data collection, preprocessing, feature extraction, model selection, training, assessment, optimization, and deployment are all included in speech-emotion recognition systems. This study aims to investigate a machine learning method for Bengali speech emotion classification. The work seeks to accomplish accurate emotion detection using methods like MFCC-based feature extraction and models like Random Forest, XGBoost, CatBoost, and Gradient Boosting. Understanding and creating efficient speech emotion detection algorithms have the potential to significantly improve human-machine interactions, and emotional comprehension, and facilitate applications in many other fields.

V. The Materials and Method

To identify the proper emotional state, our suggested system goes through several significant phases, including generating a dataset for Bengali speech, feature extraction, feature classification, and decision-making based on extracted features and classifications. Fig. 1 is the process of emotion classification.

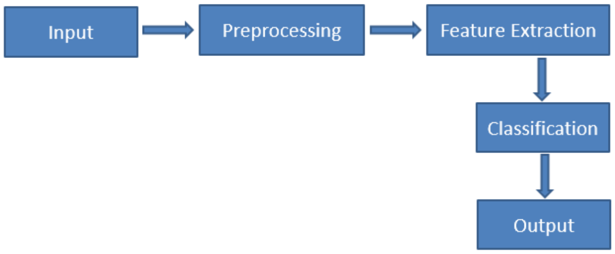


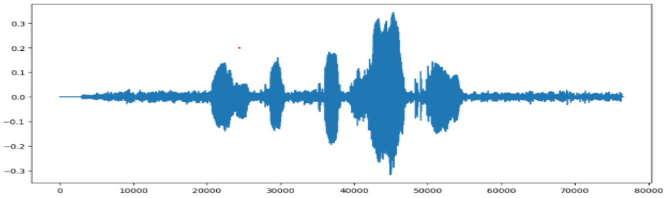
Fig. 1. Process of emotion classification.

A. Preparing Datasets

1) The technique of collecting the data for Bengali speech's classification of emotions requires several processes and meticulous planning. A total of 40 sentences were originally composed, and 25 of them were chosen for recording. Following this, these lines were uttered and recorded by people ranging in age from 9 to 48, creating a total of 25 distinct recordings. The clips ranged in length from 2 to 4 seconds, with an average of about 3 seconds, and each recording was converted to the mp3 format.

A Sample of data -
Bangla- তোমার নাম কি?
English – What is your name?
Phonemic- Tomar name ki?

Fig. 2 shows the audio frequency of each emotion.



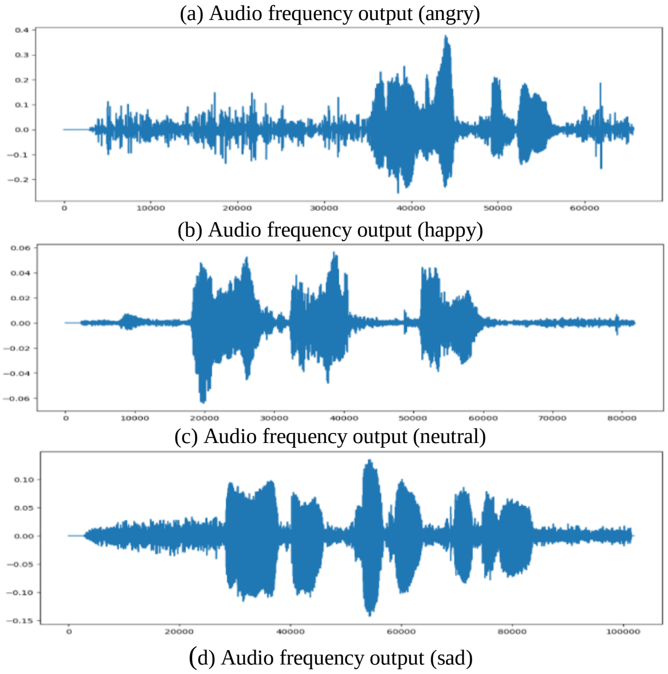


Fig. 2. Audio frequency output (angry, happy, neutral, sad).

Our Bengali speech dataset's main sources include unrestrained or natural emotions, performed or driven emotions, and elicited or created emotions [10]. Happy, Sad, Natural, and Angry were the four main emotions that were the focus of the data gathering. 14 speakers, including both male and female speakers, were enlisted to take part in the recording process to capture a wide spectrum of emotions. When compared to comparable datasets like the BanglaSER database, our dataset has a considerable volume [11]. This research produced a sizable dataset of 1400 audio recordings, 350 of which were assigned to each emotion group. Fig. 3 represents the total dataset and accuracy of each algorithm.

SL	Name of Algorithm	Number of data	Accuracy (%)
1	Random Forest	350	80.01
2	XGBoost	350	80.03
3	Gradient Boosting	350	81.19
4	CatBoost	350	82.85

Fig. 3. Total Number of dataset and accuracy of each algorithm.

The data collection method required a lot of time and effort because the participants were not actors but rather members of the general public. In order to capture a realistic depiction of genuine emotional expression, it was essential to guarantee the speaker’s independence. The dataset is important because it is diverse in terms of speakers and emotions. The dataset offers a wide range of vocal qualities and expressions with four female voices and ten male sounds. An 80-20 split was used to ensure the dataset's usefulness, with 80% of the data (1120 clips) going toward training the models and the other 20% (280 clips) going toward testing and assessment.

B. Extraction of Features

Feature extraction is the process of taking a little amount of information from a speech signal in order to address each speaker separately [12]. Mel-frequency cepstral coefficients (MFCC) are a frequently employed method for feature extraction, which is a vital stage in the analysis and classification of speech data. In order to extract pertinent acoustic features that capture the unique qualities of the audio signals, MFCC is used in the context of emotion categorization in Bengali speech. Parts of human speech production and perception are represented by MFCC. The human auditory system's logarithmic perception of loudness and pitch is depicted by MFCC [13].

MFCC features are based on human hearing perception [14]. The first step in the MFCC feature extraction procedure is to split the recorded speech data into brief frames, which typically last 20–40 milliseconds and have a minimal amount of overlap. The next step is to perform a series of operations on each frame to determine the MFCC coefficients. A full diagram of MFCC is given in Fig. 4.

The average frequency perception capability of humans is over the range of 1KHz [15]. To lessen spectral leakage, the speech signal is first divided into manageable chunks, and a windowing function, such as the Hamming window, is then applied. The signal is then transformed from the time domain to the frequency domain on each segment using the Fast Fourier Transform (FFT).

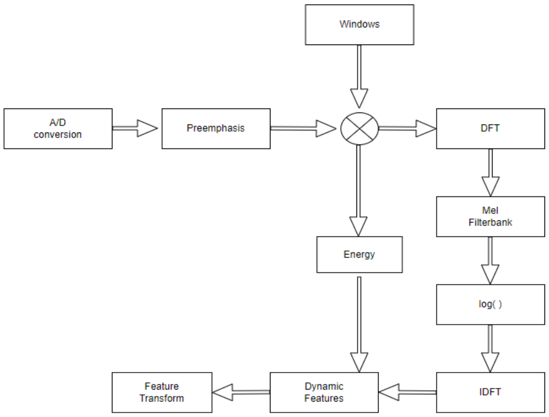


Fig. 4. Block diagram of MFCC.

The resulting frequency spectrum is then altered using a Mel filter bank to simulate the non-linear perception of sound by the human auditory system. The purpose of this filter bank is to simulate how sensitive different frequencies are to the human ear. The magnitude of the spectral components is represented by the logarithm of the energy contained within each Mel filterbank.

The Discrete Cosine Transform (DCT) is used to decorrelate the Mel-frequency coefficients as the last stage in the MFCC feature extraction process. This treatment decreases dimensionality, compresses the data, and keeps the most important information.

The signal produced by our speech often comprises a large number of criteria that indicate the signs of different emotions. Pitch, energy, format, as well as spectral features including Mel-frequency cepstrum coefficients (MFCC), are recovered in this study. The approach known as MFCC (Mel-frequency cepstral coefficients) is commonly used in speech recognition and voice emotion recognition. It is predicated on the knowledge that speech impulses are not linearly processed by the human auditory system. A non-linear Mel scale filter bank is used, which amplifies the lower frequencies while attenuating the higher ones, to capture the phoneme-specific information included in the lower frequency components of speech.

The short-term power spectrum of a speech frame is represented by the Mel frequency cepstrum in speech processing. It can be found by taking the logarithm of the power spectrum on a Mel frequency scale and performing a linear cosine transform. The formula m = 2595 * log10(f/700+1) converts the normal frequency (f) to the Mel frequency (m). We are able to convert any frequency to its Mel equivalent using the presented equation. [16].

$$m = 2595log10(\frac{f}{700} + 1)$$

C. Classification Method

Four well-known algorithms were used to categorize emotions in Bengali speech using machine learning: Random Forest, XGBoost, CatBoost, and Gradient Boosting. These algorithms were selected because they handled classification jobs well and could identify intricate patterns in the data. Many of the researchers use many algorithms but small work with those algorithms.

1) Random forest: An ensemble learning system called Random Forest mixes various decision trees to produce predictions. Each tree in the forest is trained using a portion of the data, and the combined forecasts of all the individual trees are used to make the final prediction. Fig. 5 shows the Random Forest architecture. Double randomness is the random forest's primary attribute [17]. Random Forest is renowned for its resistance to overfitting and capacity for handling large-scale, multidimensional data. In Fig. 7, it shows the confusion matrix of Random Forest.

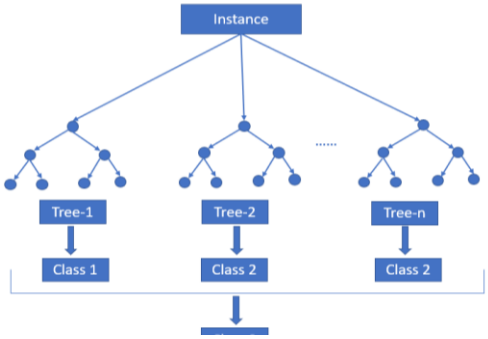


Fig. 5. RF classifier architecture.

2) XGBoost: It is an ensemble machine-learning technique based on decision trees that makes use of an improved gradient-boosting framework [18]. Due to its outstanding results in numerous machine learning contests, the gradient boosting technique known as XGBoost (Extreme Gradient Boosting) has become increasingly popular. In Fig. 6, Gradient descent is used to incrementally improve an ensemble of poor prediction models, such as decision trees. The capacity of XGBoost to handle intricate relationships and identify non-linear patterns in the data is well known. The major benefit of this algorithm is its regularization technique [19]. Fig. 8 shows the confusion matrix of XGBoost.

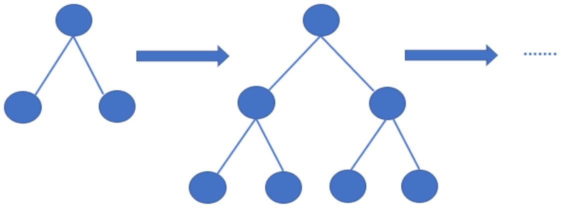


Fig. 6. XGBoost architecture.

3) CatBoost: Another gradient boosting method created specifically to handle category features effectively is CatBoost. It combines ordered boosting and gradient-based optimization techniques to obtain high accuracy in classification applications. When working with datasets that combine numerical and categorical variables, CatBoost is especially helpful. By applying the process of encoding, this method is able to handle the many types of data in the form as they emerge. Such a category can be handled without difficulty by Catboost [20]. Fig. 9 shows the confusion matrix of CatBoost.

4) Gradient boosting: A broad ensemble learning approach called gradient boosting combines several weak learners to produce a powerful predictive model. It is among the effective models created for making predictions. Three components make up the technique [21]. A decision tree to strengthen weak learners, a differentiable loss function, and an additive model to help choose the best decision tree model [22]. Each decision tree's nodes use a unique subset of characteristics to determine the appropriate split. In this method, each tree is distinct and capable of extracting a distinctive signal from the data points. Additionally, every new tree is built on top of the mistakes made by the prior tree, and all of these procedures are carried out sequentially [23]. In Fig. 10, it shows the confusion matrix of Gradient Boosting.

VI. RESULT AND DISCUSSION

An experimental study was carried out to gauge how well the machine learning approach performed in classifying emotions in Bengali speech. The dataset used for this assessment was made up of recorded Bengali speech samples from different people, covering a variety of emotions including happy, sad, natural, and angry.

For the emotion categorization challenge, four well-known machine learning algorithms—Random Forest, XGBoost, CatBoost, and Gradient Boosting—were used. Using features taken from the Bengali speech samples, such as Mel-frequency cepstral coefficients (MFCC), each algorithm was trained on the training subset. Following the training phase, the classification models were used to forecast the emotions present in the Bengali speech samples using the testing subset. The accuracy of each algorithm was determined by contrasting the anticipated labels of the test samples with their actual labels.

A. Accuracy

A model's overall accuracy in foretelling both positive and negative events is measured by accuracy. It is determined by dividing the total number of forecasts by the number of correct guesses. Although accuracy offers a broad evaluation of model performance, it may not be appropriate when classes are unbalanced.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

B. Precision

Precision measures a model's ability to categorize positive cases among all instances that it properly predicted as positive. It is determined by dividing the total number of true positives (positives that were successfully predicted) by the sum of true positives and false positives (positives that were mistakenly forecasted). The model's precision indicates its capacity to prevent false positives.

$$Precision = \frac{TP}{TP + FP}$$

C. Recall

The capacity of a model to properly identify every positive case is measured by the recall. It is determined by dividing the total number of true positives by the total number of false negatives, which are positives that were mistakenly labeled as negatives. Recall shows how well the model can prevent false negatives.

$$Recall = \frac{TP}{TP + FN}$$

D. F-1 score

The F1 score is a statistic that combines precision and recall. By using the harmonic mean of precision and recall, it offers a fair evaluation of the model's performance. The F1 score is calculated as 2 * ((precision * recall) / (precision + recall)). It has a scale of 0 to 1, with 1 representing the highest attainable score.

$$F - 1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

These findings show how well Bengali speech can be classified according to emotions using machine learning. The most accurate results came from CatBoost, which demonstrated a great capacity to identify underlying patterns and relationships in the data. The ability of Gradient Boosting, XGBoost, and Random Forest to correctly predict emotions in Bengali speech was also demonstrated by their respectable performance.

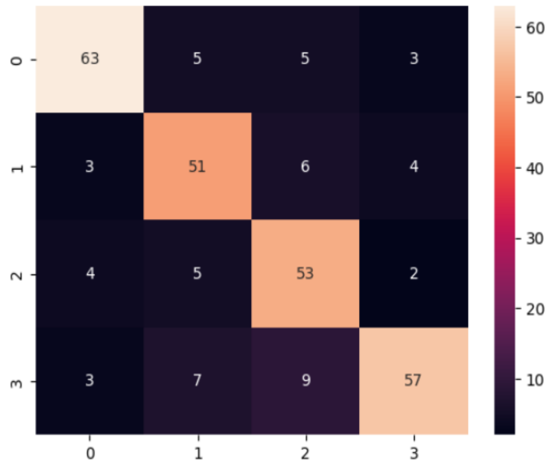


Fig. 7. Confusion matrix of random forest.

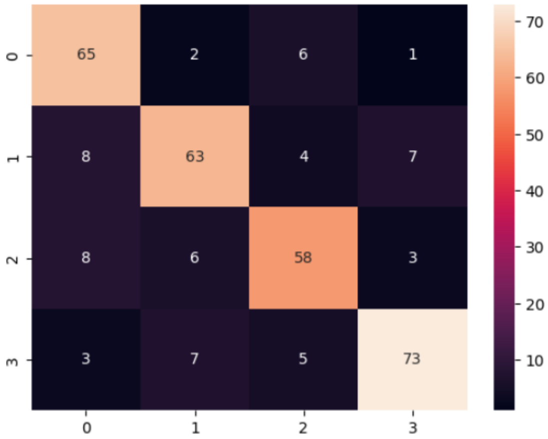


Fig. 8. Confusion matrix of XGBoost.

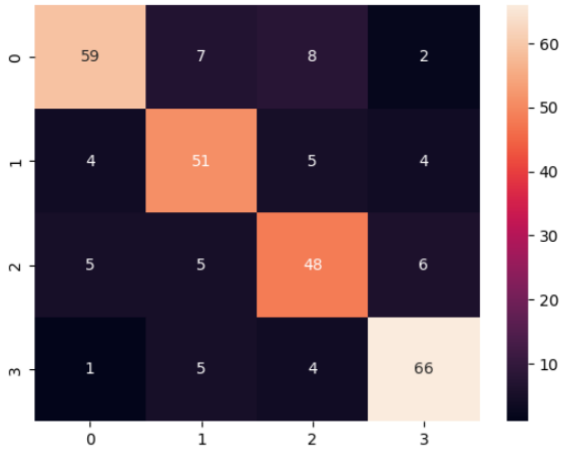


Fig. 9. Confusion matrix of CatBoost.

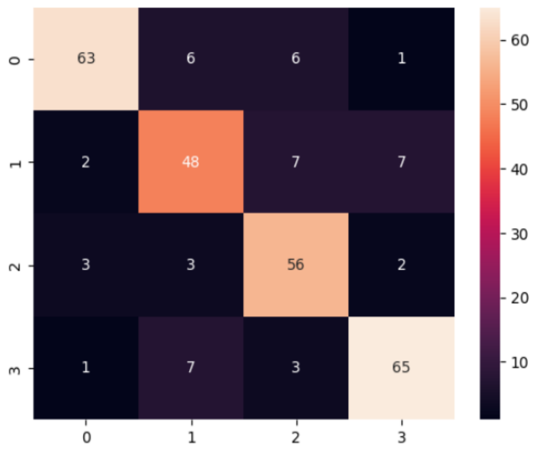


Fig. 10. Confusion matrix of gradient boosting.

TABLE I. CONFUSION MATRIX

ML Model	Emotion	Precision	Recall	F1-score	Average accuracy
Random Forest	angry	0.86	0.83	0.85	0.80
	happy	0.75	0.80	0.77	
	neutral	0.73	0.83	0.77	
	sad	0.86	0.75	0.80	
XGBoost	angry	0.86	0.78	0.81	0.80
	happy	0.75	0.80	0.77	
	neutral	0.74	0.75	0.74	
	sad	0.85	0.87	0.86	
Gradient	angry	0.77	0.88	0.82	0.81
	happy	0.81	0.77	0.79	
	neutral	0.79	0.77	0.78	
	sad	0.87	0.83	0.85	
CatBoost	angry	0.77	0.88	0.82	0.82
	happy	0.81	0.77	0.79	
	neutral	0.79	0.77	0.78	
	sad	0.87	0.83	0.85	

The experimental evaluation sheds important light on the effectiveness of various algorithms and emphasizes the significance of choosing the right machine-learning methods for Bengali speech emotion categorization. To acquire a deeper knowledge of the models' performance, additional research and assessment could be conducted to study different evaluation measures, such as precision, recall, and F1 score.

We shall start by outlining the judicial system of our suggested model. We have taken into account the Table I accuracy, precision, recall, and F-1 score.

The experimental results showed how each algorithm performed in terms of accuracy. Gradient Boosting, CatBoost, Random Forest, and XGBoost all produced results with accuracy higher than 81.19%, 82.85%, 80.01%, and 80.03% respectively.

Overall, the experimental findings support the effectiveness of the machine learning method for correctly detecting emotions in Bengali speech, advancing emotion identification systems designed for the Bengali language.

VII. CONCLUSION AND FUTURE WORK

In conclusion, this research focused on developing a machine-learning approach for emotion classification in Bengali speech. The experimental evaluation demonstrated the effectiveness of the Random Forest, XGBoost, CatBoost, and Gradient Boosting algorithms in accurately predicting emotions in Bengali speech samples. The achieved accuracy rates ranged from 80% to 83%, with CatBoost exhibiting the highest accuracy. This research contributes to addressing the scarcity of studies on emotion classification in non-English languages, specifically Bengali, and provides valuable insights into the potential of machine learning algorithms for capturing emotions in Bengali speech.

Incorporating sequence modeling techniques, such as recurrent neural networks or transformers, as well as taking into account the temporal dynamics of emotions in a speech could capture the temporal dependencies and boost the precision of emotion classification.

We realize that our study has some limitations. There are a total of seven different emotional categories; initially, we concentrated on the classification of four distinct emotions. Second, while MFCC was the main methodology we used for feature extraction, there are other feature extraction methods like LPC and PLP that can provide different insights and advantages.

In future work, we plan to expand our approach by incorporating a broader range of algorithms for emotion classification. Additionally, we intend to explore additional feature extraction techniques, including LPC and PLP, to enhance the robustness and accuracy of our emotion classification system.

REFERENCES

[1] M. Sidorov, S. Ultes, and A. Schmitt, "Emotions Are A Personal Thing: Towards Speaker-Adaptive Emotion Recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014, pp. 4836-4840. doi: 10.1109/ICASSP.2014.6854514.

[2] D. Czerwinski and P. Powroźnik, "Human Emotions Recognition with the Use of Speech Signal of Polish Language," in Proceedings of the IEEE Conference on Electrical Power and Energy Conference (EPMCCS), 2018, pp. 1-6. doi 10.1109/EPMCCS.2018.8596404.

[3] A. Agarwal and Dr. A. Dev, "Emotion recognition and conversion based on segmentation of speech in the Hindi language."

[4] M.R. Hasan, M.M Hasan, M.Z. Hossain, "How many Mel-frequency cepstral coefficients to be utilized in speech recognition?," The Journal of Engineering, 12, 817-827, 2021,doi:10.1049/tje2.12082..

[5] M.M. Rahman, D.R. Dipta, and M.M. Hasan, "Dynamic time warping assisted SVM classifier for Bangla speech recognition," in Proceedings of the International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018, pp. 1-6.

[6] N. Kholodna, V. Vysotska, and S. Albota, "A Machine Learning Model for Automatic Emotion Detection from Speech," in CEUR Workshop

Proceedings, vol. 2917, pp. 699-713, 2021.

[7] J.Y.O.T.I.R.M.A.Y. Devnath, S. Hossain, M.O.S.H.I.U.R. Rahman, H.A.S.I. Saha, M.A. Habib, and N. Sultan, "Emotion recognition from isolated Bengali speech."

[8] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," Personal and Ubiquitous Computing, vol. 25, no. 4, pp. 637-650, 2021. doi 10.1007/s00779-020-01389-0.

[9] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," Personal and Ubiquitous Computing, vol. 25, no. 4, pp. 637-650, 2021. doi 10.1007/s00779-020-01389-0.

[10] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features, and classifiers for speech emotion recognition: a review," International Journal of Speech Technology, vol. 21, 2018.

[11] R.K. Das et al., "BanglaSER: A speech emotion recognition dataset for the Bangla language," Data in Brief, vol. 42, p. 108091, 2022.

[12] A. Bala, A. Kumar, and N. Birla, "Voice command recognition System Based on MFCC and DTW," International Journal of Engineering Science and Technology, vol. 2, no. 12, pp. 7335-7342, 2010.

[13] M. Ghai et al., "Emotion recognition on speech signals using machine learning," in Proceedings of the International Conference on Big Data Analytics, Cloud Computing and Science (ICBDACI), 2017, pp. 34-39. doi 10.1109/ICBDACI.2017.8070805.

[14] R. Choudhary, G. Meena, and K. Mohbey, "Speech emotion based sentiment recognition using deep neural networks," Journal of Physics: Conference Series, vol. 2236, p. 012003, Mar 2022.

[15] M.A. Imtiaz and G. Raja, "Isolated Word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN," in Proceedings of the Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast), 2016.

[16] N. Yang et al., "Recognize basic emotional states in a speech by machine learning techniques using mel-frequency cepstral coefficient features," Journal of Intelligent & Fuzzy Systems, vol. 39, no. 2, pp. 1925-1936, 2020.

[17] R.R. Choudhary, G. Meena, and K.K. Mohbey, "Speech emotion based sentiment recognition using deep neural networks," Journal of Physics: Conference Series, vol. 2236, p. 012003, Mar 2022.

[18] S. Yan et al., "Speech Interactive Emotion Recognition System Based on Random Forest," in Proceedings of the International Wireless Communications and Mobile Computing (IWCMC), 2020. doi 10.1109/iwcmc48107.2020.9148117.

[19] "Data Analysis and Classification using XGBoost," [Online]. Available: https://www.kaggle.com/code/lucidlenn/data-analysis-and-classification-using-xgboost/notebook. [Accessed: March 7, 2023].

[20] M. Mohan, P. Dhanalakshmi, and R. Satheesh Kumar, "Speech Emotion Classification using Ensemble Models with MFCC," Procedia Computer Science, vol. 218, pp. 1857-1868, 2023. doi: 10.1016/j.procs.2023.01.163.

[21] V. Pujari et al., "Speech Emotion Recognition," International Research Journal of Engineering and Technology (IRJET), vol. 09, no. 3, pp. 3288-3294, 2022.

[22] P.T. Krishnan et al., "Emotion classification from speech signal based on empirical mode decomposition and non-linear features," Complex Intelligent Systems, vol. 7, pp. 1919-1934, 2021. doi: 10.1007/s40747-021-00295-z.

[23] J.H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. doi: 10.1214/aos/1013203451.

[24] A.M. Ishmam, S. Sharmin, "Hateful Speech Detection in Public Facebook Pages for the Bengali Language," in 18th IEEE international conference on machine learning and applications (ICMLA), 555-560, 2019, doi:10.1109/ICMLA.2019.00104.

[25] H.M. Hasan, M.A. Islam, "Emotion recognition from bengali speech using rnn modulation-based categorization," In 2020 third international conference on smart systems and inventive technology (ICSSIT), 1131-1136, 2020, doi:10.1109/ICSSIT48917.2020.9214196.

[26] J.R. Saurav, S. Amin, S. Kibria, M.S. Rahman, "Bangla speech recognition for voice search," in 2018 international conference on Bangla speech and language processing (ICBSLP), 1-4, 2018, doi:10.1109/ICBSLP.2018.8554944.

[27] N.T Ira, M.O. Rahman, "An efficient speech emotion recognition using ensemble method of supervised classifiers," in 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), IEEE, 1-5, 2020, doi:10.1109/ETCCE51779.2020.935091