ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 70-76 (2022)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Bangla Speech Emotion Detection using Machine Learning Ensemble Methods

Roy D Gregori Ayon, Md. Sanaullah Rabbi, Umme Habiba, Maoyejatun Hasana*

Department of Computer Science and Engineering, Asian University of Bangladesh, Dhaka, 1341, Bangladesh

ARTICLE INFO

Article history:
Received: 15 May, 2022
Accepted: 08 October, 2022
Online: 13 November, 2022

Keywords:
Bangla
Emotion Detection
Machine Learning

ABSTRACT

Emotion is the most important component of being human, and very essential for everyday activities, such as the interaction between people, decision making, and learning. In order to adapt to the COVID-19 pandemic situation, most of the academic institutions relied on online video conferencing platforms to continue educational activities. Due to low bandwidth in many developing countries, educational activities are being mostly carried out through audio interaction. Recognizing an emotion from audio interaction is important when video interaction is limited or unavailable. The literature has documented several studies on detection of emotion in Bangla text and audio speech data. In this paper, ensemble machine learning methods are used to improve the performance of emotion detection from speech data extracted from audio data. The ensemble learning system consists of several base classifiers, each of which is trained with both spontaneous emotional speech and acted emotional speech data. Several trials with different ensemble learning methods are compared to show how these methods can yield an improvement over traditional machine learning method. The experimental results show the accuracy of ensemble learning methods; 84.37% accuracy was achieved using the ensemble learning with bootstrap aggregation and voting method.

1. Introduction

Emotions play an important role in understanding human behaviors, thoughts, and actions. There is a plethora of applications such as human-to-human communication [1], human computer interaction [2], affective computing [3], remote patient monitoring system [4], etc. where emotion detection is a vital part for decision making, problem solving or understanding the mental state of a subject. Human emotions can be divided into primary and compound emotions. Primary emotions consist of eight types of emotions such as anger, fear, sadness, disgust, surprise, anticipation, acceptance, and joy. Compound emotions can be derived by conjugating two or more primary emotions. On the other hand, emotions may vary not only from person to person but also in different contexts, communities, cultures, and languages. This work detects emotions from Bangla speech data extracted from audio.

Working with Bangla speech data to detect emotions is quite difficult and different in terms of accent, pitch, rhythm, intonation, pronunciation, and voice modulation. Selecting the right set of features is necessary to correctly classify emotions from Bangla speech data. There are several feature extractions approaches such as perceptual linear prediction (PLP), linear prediction coding (LPC) and Mel-frequency Cepstrum Coefficients (MFCC), which have been used for speech recognition from speech data. In this study, MFCC is used to extract features from Bangla audio speech data collected from Bangla speaking participants.

Literature has documented numerous traditional machine learning approaches to classify an emotion from different types of data such as text, audio, video, image, brainwaves, etc. There are few notable works that detect emotions from Bangla speech data [5]-[9]. In [5], the authors have investigated the optimum number of MFCCs to recognize an emotion from speech data and suggested that MFCCs should be 25. In [6], the authors have developed a Gated Recurrent Unit (GRU) based deep neural network model to classify users' comments on Facebook pages. The authors have collected 5,126 Bangla comments and classified them into six classes: hate speech, communal attack, inciteful, religious hatred, political comments, and religious

*Corresponding Author: Maoyejatun Hasana, Asian University of Bangladesh, Dhaka-1341, Bangladesh, mjhasana@aub.edu.bd

comments. The accuracy of GRU based model is 70.10%. In [7], the Recurrent Neural Network (RNN) is used to classify six emotions: joy, sadness, anger, surprise, fear, and disgust from Bangla speech and achieved 51.33% accuracy. In [8], Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) and Deep Neural Network-Hidden Markov Model (DNN-HMM) are used to search emotions from 49 different speakers of a vocabulary of 500 unique words. The performance criterion of the models is considered Word Error Rate (WER) and achieved 3.96% WER for GMM-HMM, whereas 5.30% WER for DNN-HMM. In [9], an ensemble method of several supervised classifiers has been used to classify emotions from speech data and achieved 70% accuracy. We can see that the existing works have not achieved significant accuracy in detecting, and/or recognizing emotions from audio data. This work uses ensemble machine learning methods to detect four types of emotion such as happy, sad, angry, and neutral. Different trials of ensemble machine learning methods have been conducted to achieve better accuracy. The specific contributions of this work are as follows:

- Bangla speech data collection with a careful avoidance in data biases.
- Implementation of a noise reduction module which has been used during pre-processing.
- Apply a different set of ensemble machine learning methods to achieve better accuracy in emotion detection.

The rest of this paper is organized as follows. Related research is given in the next section. Dataset information is given in section 3. After that, the detail of the proposed method is described in section 4. Following the methodology, results and analysis are drawn from the experiments. Then a conclusion is drawn.

2. Related Work

From the last few decades, enormous research works have been accomplished in the field of emotion detection, recognition, and/or classification. Emotional intelligence is widely used to develop an emotionally aware healthcare monitoring system or a safe driving system or during computer games. This section will focus on reviewing different studies on emotion detection as well as studies other than the English language.

2.1. Study of Speech Emotion Detection

In [10], the authors developed a machine learning model for automatic emotion detection from speech. The model is used to monitor public emotions. The authors chose a manually annotated dataset and represented it as text using a vectorization method. Deep learning methods, convolutional, recurrent neural networks, and perception are used to detect emotions in textual data. The accuracy of the obtained classification model is quite low, which is 77% for random forest, 74% for regression, and 73.5% for naive Bayesian classifier.

In [11], the authors presented an ANN approach to predict emotion in the field of Music Emotion Recognition. 167 voices were analyzed, and 76 features were extracted from International Affective Digital Sounds Dataset (IADS). This audio dataset was segmented into three parts for the purpose of training (70%), validation (15%) and testing (15%). In the prediction stage, the ANN model accounted for 64.4% in arousal and 65.4% in valence. The result showed that the shallow neural network performs better than the regression model.

In [12], the authors presented a method for detecting emotion using speech using IoT based deep learning. The authors implemented a real time system based on IoT, and then classified emotions. The authors proposed an integrated deep learning model named Speech Emotion Detection (SED) using 2D convolutional neural network. The accuracy rate achieved by SED is approximately 95%.

In [13], the authors presented a new set of acoustic features for automatic emotion recognition from audio. The author proposed a feature based perceptual quality metric which is based on the masked perceptual loudness. The features computed emotion based on emotional difference such as “happy/excited”, “angry/anxious”, “sad/bored”, and “relaxed/serene” in the reference set of data. The authors used the proposed set referred as a perceptual feature set that consists of a 9-dimensional feature vector with 7 low level and 2 statistical descriptors. GMM and SVM classifiers are used for computing emotions. A decision rule to be interpreted as an S-MV rule was proposed by the authors and it showed an improved recognition performance specially for valence which was valid in both acted and natural emotions.

In [14], the authors explained architecture for modeling conversation through language models encoding, and classification stages. The authors used transfer learning through the universal language modeling that is composed of Bi-LSTM units. The authors also list the hyperparameters which are used for building and training these models. The F1-score of this model is 0.7582.

The EmoDet2 system is presented in [15] to detect emotion using a deep learning approach. EmoDet2 takes English textual dialogue as an input, and from text it detects four types of emotion such as happy, sad, angry, and others. The authors combined neural network architecture and BiLSTM neural network to obtain substantial improvement over the baseline model. The performance of EmoDet2 is quite satisfactory, with an F1-score of 0.78.

In [16], the authors developed a system to detect emotion from the Roman Urdu text. The authors developed a comprehensive corpus of 18k sentences that converged from distinct domains, and annotated it with six different classes. The authors also applied different baseline algorithms like KNN, Decision tree, SVM, Random Forest on their corpus. The authors gained an accuracy rate of 69.4% and an F-measure of 0.69.

A method to recognize emotion collected from social media like Twitter is described in [17]. The authors classify English text into six different emotions which are happiness, sadness, fear, anger, surprise, and disgust. The authors used natural language processing and machine learning classification

algorithms. The authors also managed to create a large bag of emotional words along with their emotion-intensities. The authors achieved an accuracy of 91.7% for SMO and 85.4% for J48.

In [18], the authors presented a model to detect multiclass emotion detection from Bangla text. The authors used a Multinomial Naïve Bayes (MNB) classifier with various features. The model can classify three emotions: happy, sad, and angry from text with an accuracy of 78.6%.

In [19], the authors explained a machine learning method to recognize four major types of human emotions which are anger, sadness, joy, and pleasure. The authors incorporated electrocardiogram (ECG) signals to recognize emotions. The authors combined four ECG signal-based techniques which are heart rate variability, empirical mode for decomposition within beat analysis, and frequency spectrum analysis. The frequency spectrum analysis used in this work is proposed in this work. By comparing it with the best biosensor-based model, this ensemble model attained an accuracy rate of 10.77%.

In [20], the authors presented a framework of Long Short-Term Memory (LSTM) and 2D Convolutional Neural Network (CNN) to detect emotion from physiological signals acquired using wearable, low-cost sensors. In [21], the authors used two ensemble classification schemes: stacked generalization and unweighted voting for spoken emotion recognition. Stacked generalization is an approach to combining predictions from multiple classifiers. In an unweighted voting method, the class predictions of the base-level classifiers are abridged and the class which gets majority votes is selected as the final class. Numerous deep learning architectures have been used in [22] for emotion detection from both speech and text data.

Most of these works have been conducted to recognize emotions in English language, which could not be useful for detecting emotions in languages other than English.

2.2. Study of speech emotion detection in languages other than English

Beyond the English language-based speech emotion detection studies, researchers have worked on many other languages such as Persian [23], Urdu [24], Arabic [25], Hindi [26], etc. Since each language has different kind of expressions to show emotional states, generalizing emotions for all languages would be a difficult task. Hence, speech emotion detection systems are generally developed language-dependently. As native speakers of Bangla language, we choose to work on emotion detection from spoken Bangla using our own dataset.

3. Dataset

To detect different emotions from speech audio data, we needed to develop an emotion speech dataset. In this work, we have collected data from 20 participants, of which 12 are males, and 8 are females. All participants are native speakers of Bangla language. Table 1 and Table 2 show the detailed information of participants' age, gender, and occupation. We have collected data from different age groups ranging from 18 to 32 years, and with different occupations such as job holder, student, businessman, and self-employed. We have collected 452 samples and the duration of each sample is from 3 to 5 seconds. We have labeled these samples in four emotional categories based on the type of speech data. The emotion categories are angry, happy, sad, and neutral. We have recorded the speech audio data on smartphones. Since the participants had different models of smartphone, we had to convert the recorded data into one common audio format. In this work, we have converted the data into wav format. The volume of our dataset is not large compared to other datasets such as RAVDESS multimodal database [27]. However, we could achieve a good accuracy with fewer samples of data that is discussed in the result section.

Table 1: Age and gender of the participants

Gender	Age			Total
	18-22	23-27	28-32	
Male	04	05	03	12
Female	03	04	01	08

Table 2: Occupation of the participants

Gender	Occupation			
	Job Holder	Student	Business	Self-Employed
Male	02	06	01	03
Female	01	05	00	02

4. Methodology

Voice is the prominent medium to communicate. Our objective is to detect the emotion from audio data using several machine learning models including traditional, and ensemble models. We have gone through a number of pre-processing steps before training our dataset. The pre-processing is done to remove any unwanted noises from the audio data. We will discuss each phase of pre-preprocessing in the next subsection.

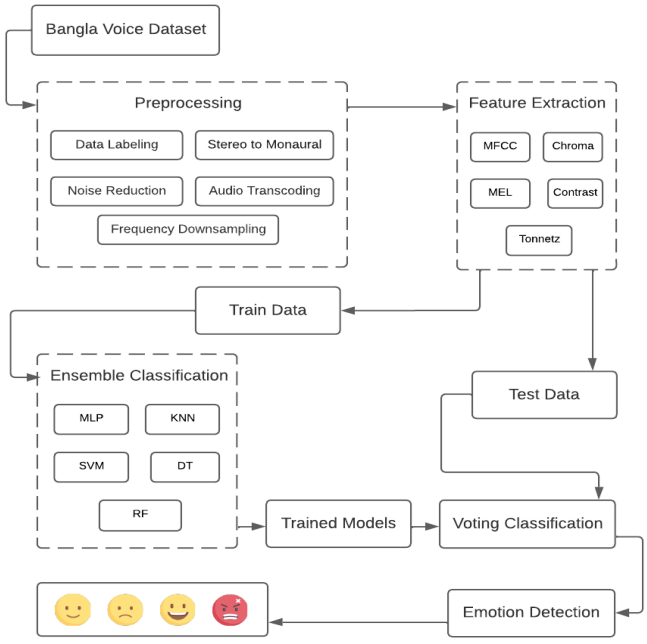

Figure 1: Architecture of the proposed methodology

A block diagram of emotion detection methodology is shown in Figure 1 where the pre-processed data go through the feature extraction phase. Here, we extract several voice features such as MFCC, MEL, contrast, tonnetz, etc. After the feature extraction phase, the audio data are sent to ensemble machine learning models for training. After training the models, our next step is to detect different types of emotions from test data.

4.1. Pre-processing

Pre-Processing is a technique that transforms raw data into understandable format. It is not suitable for feature extraction modules with raw data directly because the data are collected from different platforms and environments. Our dataset was collected in various formats such as MP3, MP4 AAC, M4A, WMA, FLAC, OGG, etc. Also the dataset was not properly labeled, and was contaminated with a lot of extra unwanted noises. In order to simplify the further steps, we processed the dataset and cleaned them as follows:

- Data Labeling: Our dataset is formatted based on RAVDESS multimodal database [27] of emotional speeches and songs. Each of the 452 samples has been labeled based on the speech emotion data. The filename consists of a 7-part numerical identifier (e.g., 03-01-05-01-02-01-12.wav). These identifiers define stimulus characteristics of the speech data. Third identifier of the filename is defined as Emotion (e.g., 01 = neutral, 03 = happy, 04 = sad, 05 = angry). For example in the filename: 03-01-05-01-02-01-12.wav, the 3rd identifier is 05, and is referred to emotion angry. The other identifiers are not necessary for this study.
- Audio Transcoding: We have used the *librosa* module to convert the audio data. This module returns WAV audio format from raw data such as MP3, MP4 AAC, M4A, WMA, FLAC, OGG etc. Audio transcoding is done to convert different audio formats to one common audio format.
- Noise Reduction: Noise reduction is the process of removing noise from a signal. Noise reduction techniques exist for audio and images. In this work, a python module *pydub* is used for audio segmentation to remove extra noise from audio data.
- Stereo to Monaural: Monaural or monophonic sound (mono) reproduction is sound intended to be heard as if it is emanating from one position. Mono channel is used when there is only one source of audio, and the stereo channel is used when there are multiple sources of audio. Since we are only taking speech data without contamination of any music or instruments, we have converted stereo to mono channel to reduce the usage of bandwidth and storage space. We have used *FFmpeg* multimedia framework for converting our audio data from stereo to mono.
- Frequency Downsampling: We downsampled the audio data to adjust frequency to 16kHz using *FFmpeg* multimedia framework.

4.2. Feature extraction

Choosing a suitable set of features is an important step in detecting emotions from speech data. Speech features can be divided into spectral, excitation, acoustic features [25]. We have selected several features such as MFCC, Chroma feature, LFCC, LPC, RC, Contrast, Tonnetz, etc. We use a minimal set of features to reduce the complexity in emotion detection. Since there is no general agreement on the right number of features for detecting emotions from speech data, we have chosen features that are effective and computationally efficient.

4.3. Train model

After the feature extraction phase, the extracted features are used to train machine learning models. In this stage, different machine learning models are trained using the dataset. For this, the dataset is split into 85% for training and 15% for testing. Since the volume of our dataset is small, we have split the dataset based on empirical findings. After training a model, the dataset is gone through the testing phase to evaluate model's accuracy.

4.4. Detection of emotion

Machine learning classifiers have been used in predicting, recognizing, and detecting the desired target of a given dataset. In this work, we have trained and tested traditional and ensemble machine learning models to detect emotions from Bangla speech audio data.

4.4.1. Traditional Machine Learning Models

Five traditional machine learning models are used in this work. They are Multi-Layer Perceptron (MLP), K-Nearest Neighbors (KNN), Decision tree (DT), Random Forest (RF), and Support vector machines (SVM). These machine learning approaches are used in detecting different emotional states such as happy, sad, angry, and neutral from the extracted speech data.

4.4.2. Ensemble Machine Learning Model

Ensemble machine learning model is a combination of different sets of models. It provides a final output by combining several outputs of different ML models. Hence, ensemble machine learning model gives more accurate performance. The final decision can be taken by the ensemble model by using different methods such as hard voting, bootstrap aggregations, boosting, etc. It provides more accurate results by relying on a decision tree rather than one model as shown in Figure 2.

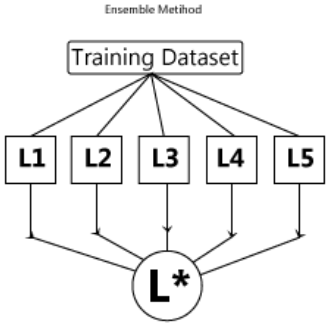

Figure 2: An example ensemble method

In Figure 2, there are 5 weak learners that are L_1, L_2, L_3, L_4 and L_5. They are going through training models to become a vigorous learner as L*.

In this work, we used a bagging ensemble model. Bagging (Bootstrap Aggregation) is generally used to reduce the contradiction. In bootstrap aggregation, multiple similar algorithms are trained separately, and then merge them all to determine the model's average. We used five different types of algorithms for the bootstrap aggregation method. In Figure 3 shows ensemble method with bootstrap aggregation.

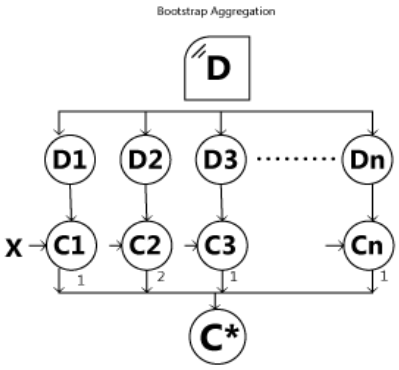

Figure 3: An example of Bootstrap Aggregation

In Figure 3, we have provided an original dataset. The bootstrap aggregation process generates multiple subsets which are D_1, D_2, D_3, … D_n from the given dataset. On each subset, a machine learning algorithm is fitted. The fitting algorithm is then trained using multiple subsets to produce various models which are C_1, C_2, C_3, … C_n. The produced models are called weak learners or base models. Now, we have our multiple base models which are trained in parallel at this stage. In the last stage, the multiple predictions made by the base models are combined to produce a single final model which is C*. The final model C* will have low variance and a high accuracy score.

5. Result and Discussion

In order to evaluate the performance of our work, we have considered precision, recall, and F1-score for each emotion class and then we have calculated the average accuracy. We have experimented with different ML models to find out which model performs better in terms of detecting emotions with higher accuracy.

We have tested multiple machine learning models for our work. Since these machine learning models failed to give our desired result, we go with ensemble methods with bootstrap aggregation that not only give the better accuracy but also improve the stability of machine learning models, prevent model overfitting, and reduce variance.

Table 3 gives the precision, recall and the weighted-average F1-scores for the multiclass classification using ensemble model with SVM, MLP, KNN, DT and RF classifiers for training dataset. In Table 3, DT with ensemble model performs better than other classifiers with an average accuracy of 99% approximately.

Table 4 shows the precision, recall and the weighted-average F1-scores for test dataset. The RF model gives slightly better accuracy than other models, which is 78%. Here, we also see that, DT has the accuracy score of 77%, while KNN has 74%, and MLP has 71% accuracy. On the other hand, precision, recall, f1-score and accuracy percentage are quite low for the SVM classifier compared to other classifiers. SVM gives only 65% accuracy.

Table 3: Comparison of different ensemble ML models for training dataset

ML Model	Emotion	Precision	Recall	F1-score	Average accuracy
Decision Tree	angry	1.00	0.98	0.99	0.99
	happy	0.94	0.98	0.96	
	neutral	0.99	0.97	0.98	
	sad	0.98	1.00	0.99	
Random Forest	angry	1.00	0.98	0.99	0.96
	happy	0.89	0.97	0.93	
	neutral	0.92	0.96	0.94	
	sad	0.97	0.99	0.98	
KNN	angry	1.00	0.98	0.99	0.98
	happy	0.94	0.98	0.96	
	neutral	0.99	0.97	0.98	
	sad	0.98	1.00	0.99	
MLP	angry	0.96	0.98	0.97	0.93
	happy	0.90	0.95	0.93	
	neutral	0.91	0.93	0.92	
	sad	0.95	0.86	0.90	
SVM	angry	0.98	1.00	0.99	0.97
	happy	0.93	0.97	0.95	
	neutral	0.97	0.99	0.98	
	sad	0.95	0.97	0.96	

Table 4: Comparison of ensemble ML models for test dataset

ML Model	Emotion	Precision	Recall	F1-score	Average accuracy
Decision Tree	angry	0.85	0.73	0.79	0.77
	happy	0.79	0.75	0.77	
	neutral	0.71	0.79	0.75	
	sad	0.75	0.81	0.78	
Random Forest	angry	0.64	0.94	0.79	0.78
	happy	0.95	0.57	0.76	
	neutral	0.69	0.71	0.70	
	sad	0.79	0.94	0.87	
KNN	angry	0.76	0.80	0.78	0.74
	happy	0.83	0.57	0.70	
	neutral	0.68	0.78	0.73	
	sad	0.67	0.81	0.74	
MLP	angry	0.63	0.73	0.68	0.71
	happy	0.79	0.71	0.75	
	neutral	0.77	0.69	0.73	
	sad	0.65	0.77	0.71	
SVM	angry	0.55	0.69	0.62	0.65
	happy	0.61	0.71	0.66	
	neutral	0.65	0.57	0.61	
	sad	0.76	0.69	0.71	

As we progressed, we see that we are not getting our desired result. So, we have applied an ensemble voting classifier for better accuracy. The intuition behind using hard voting is that to label the emotion that has been chosen most frequently by the classification models.

Hard voting is the simplest case of majority voting. Here, we label a class ŷ via majority (plurality) voting of each classifier C_j:

$$\hat{y} = mode\{C1(x), C2(x), \ldots, Cm(x)\}$$

Here C represents the models. We combine our five models in the voting classifier to get better performance in detecting emotions.

Table 5: Accuracy of ensemble methods with voting

Ensemble Method	Emotion	Precision	Recall	F1-score	Average accuracy
Voting classifier	angry	0.76	0.82	0.79	**0.84**
	happy	0.83	0.89	0.86	
	neutral	0.82	0.86	0.84	
	sad	0.83	0.91	0.87	

In Table 5, we have calculated the F1-score of four types of emotions: a79.5% for angry, 86.25% for happy, 84.75% for neutral and 87% for sad emotions. The average accuracy is 84.37% using the voting classifier. So, we can say that if we test any sample of Bangla speech data, we can get the success rate of emotion detection of 84.37%. During evaluation of ensemble model in our work, we have found that the size of the dataset, the number of features, and the classifiers affect the detection of emotions to some extent.

In Figure 4 and Figure 5, we have showed the ensemble bootstrap aggregation model accuracy for training and test dataset respectively. In our emotion recognition journey, we try to find out which classifier detects emotion accurately, so we have compared the ensemble classifiers and traditional classifiers. In Figure 5, we clearly see that our ensemble classifier gives better performance compared to traditional classifiers.

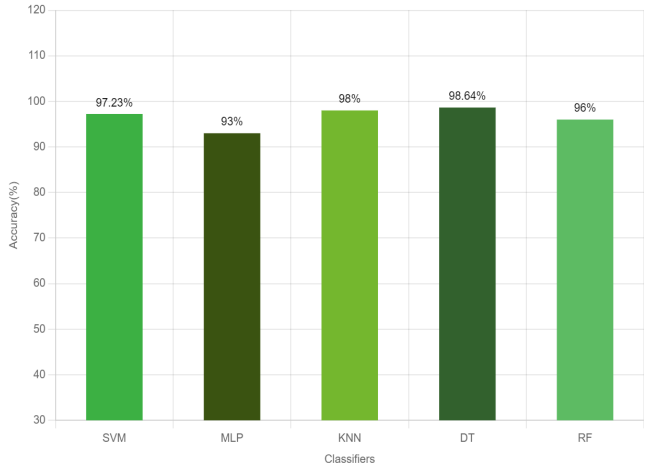

Figure 4: Ensemble bootstrap aggregation training model accuracy

In Figure 6, we have shown the emotion detection success rate using ensemble model with voting. From our observation, we can say that by using ensemble voting method, we can get better accuracy in compared to traditional methods, which is 84.37%.

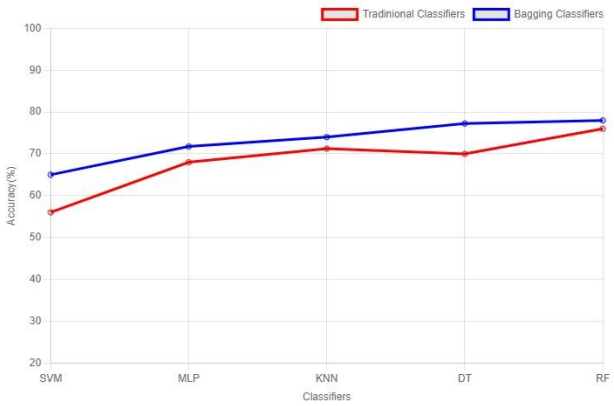

Figure 5: Comparison between begging classifier and traditional classifier

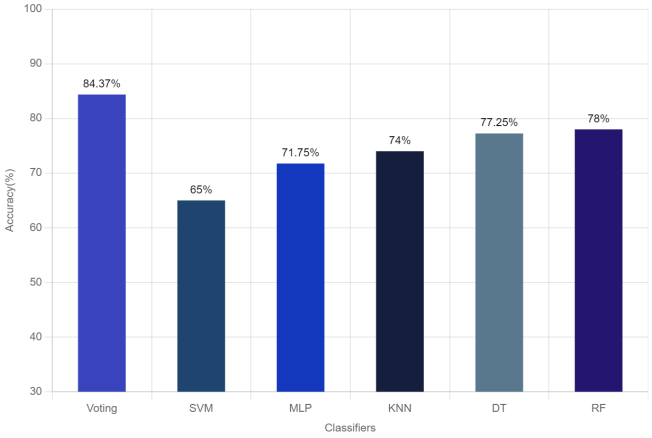

Figure 6: Accuracy of ensemble method with voting

6. Conclusion

In this work, we have explored emotion detection in Bangla language. We have presented an ensemble machine learning model with voting to detect emotion from speech data. The ensemble learning model with voting outperformed traditional machine learning models with better accuracy. However, our dataset has fewer samples of speech data. The amount is not sufficient to achieve a better result. Therefore, in future, more data samples would be considered for detecting broad ranges of emotions from Bangla speech data.

References

[1] S. C. Hauser, S. McIntyre, A. Israr, H. Olausson, G. J. Gerling, "Uncovering human-to-human physical interactions that underlie emotional and affective touch communication," in 2019 IEEE World Haptics Conference, 407–412, 2019, doi: 10.1109/WHC.2019.8816169.

[2] A. Al-Nafjan, K. Alharthi, H. Kurdi, "Lightweight building of an electroencephalogram-based emotion detection system," Brain Sciences, **10**(11), 781, 2020, doi:10.3390/brainsci10110781.

[3] S. Pal, S. Mukhopadhyay, and N. Suryadevara, "Development and progress in sensors and technologies for human emotion recognition," Sensors, **21**(16), 2021, doi:10.3390/s21165554

[4] C. Athavipach, S. Pan-Ngum, and P. Israsena, "A wearable in-ear EEG device for emotion monitoring," Sensors, **19**(18), 4014, 2019, doi:10.3390/s19184014.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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/ETCCE51779.2020.9350913.

[10] N. Kholodna, V. Vysotska, S. Albota, "A Machine Learning Model for Automatic Emotion Detection from Speech," in CEUR Workshop Proceedings, 2917, 699-713, 2021.

[11] S. Cunningham, H. Ridley, J. Weinel, R. Picking, "Supervised machine learning for audio emotion recognition," Personal and Ubiquitous Computing, **25**(4), 637-650, 2021, doi: 10.1007/s00779-020-01389-0.

[12] Z. Tariq, S.K. Shah, Y. Lee, "Speech emotion detection using iot based deep learning for health care," in 2019 IEEE International Conference on Big Data, 4191-4196, 2019, doi: 10.1109/BigData47090.2019.9005638.

[13] M.C. Sezgin, B. Gunsel, G.K. Kurt, "Perceptual audio features for emotion detection," EURASIP Journal on Audio, Speech, and Music Processing, **2012**(1), 1-21, 2012, doi:10.1186/1687-4722-2012-16.

[14] W. Ragheb, J. Azé, S. Bringay, M. Servajean, "Attention-based modeling for emotion detection and classification in textual conversations," arXiv preprint arXiv:1906.07020, 2019, doi:10.48550/arXiv.1906.07020.

[15] H. Al-Omari, M.A. Abdullah, S. Shaikh, "Emodet2: Emotion detection in english textual dialogue using bert and bilstm models," in 2020 11th International Conference on Information and Communication Systems (ICICS), 226-232, 2020, doi: 10.1109/ICICS49469.2020.239539.

[16] A. Majeed, H. Mujtaba, M.O Beg, "Emotion detection in roman urdu text using machine learning," in Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering Workshops, 125-130, 2020, doi: 10.1145/3417113.3423375.

[17] B. Gaind, V. Syal, S. Padgalwar, "Emotion detection and analysis on social media," arXiv preprint arXiv:1901.08458, 2019, doi:10.48550/arXiv.1901.08458.

[18] S. Azmin, K. Dhar, "Emotion detection from Bangla text corpus using Naïve Bayes classifier," in 2019 4th International Conference on Electrical Information and Communication Technology (EICT), 1-5, 2019, doi: 10.1109/EICT48899.2019.9068797.

[19] T. Dissanayake, Y. Rajapaksha, R. Ragel, I. Nawinne, "An ensemble learning approach for electrocardiogram sensor based human emotion recognition," Sensors, **19**(20), 4495, 2019, doi:10.3390/s19204495.

[20] M. N. Dar, M.U. Akram, S.G. Khawaja A.N. Pujari, "CNN and LSTM-Based Emotion Charting Using Physiological Signals," Sensors, **20**(16), 4551, 2020, doi:10.3390/s20164551.

[21] D. Morrison, R Wang, L.C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," Speech communication, **49**(2), 98-112, 2007, doi:10.1016/j.specom.2006.11.004.

[22] M. de Velasco, R. Justo, J. Antón, M. Carrilero, M.I. Torres, "Emotion Detection from Speech and Text," in IberSPEECH, 68-71, 2018, doi:10.21437/IberSPEECH.2018.

[23] O.M. Nezami, P.J. Lou, M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," Language Resources and Evaluation, **53**(1), 1-16, 2019, doi:10.1007/s10579-018-9427-x.

[24] A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, K. Fatima, "An Urdu speech corpus for emotion recognition," PeerJ Computer Science, 8, p.e954, 2022, doi:10.7717/peerj-cs.954.

[25] S. Klaylat, Z. Osman, L. Hamandi, R. Zantout, "Emotion recognition in Arabic speech," Analog Integrated Circuits and Signal Processing, **96**(2), 337-351, 2018, doi:10.1007/s10470-018-1142-4.

[26] A. Agrawal, A. Jain, "Speech emotion recognition of Hindi speech using statistical and machine learning techniques," Journal of Interdisciplinary Mathematics, **23**(1), 311-319, 2020, doi:10.1080/09720502.2020.1721926.

[27] S.R. Livingstone, 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,**13**(5), p.e0196391, 2018, doi:10.1371/journal.pone.0196391.