Emotion Recognition with Intensity Level from Bangla Speech using Feature Transformation and Cascaded Deep Learning Model

Md. Masum Billah[1], Md. Likhon Sarker[2], M. A. H. Akhand[3], Md Abdus Samad Kamal[4]
Department of Computer Science and Engineering,
Khulna University of Engineering and Technology, Khulna-9203, Bangladesh[1, 2, 3]
Graduate School of Science and Technology, Gunma University, Kiryu 376-8515, Japan[4]

***Abstract*—Speech Emotion Recognition (SER) identifies and categorizes emotional states by analyzing speech signals. The intensity of specific emotional expressions (e.g., anger) conveys critical directives and plays a crucial role in social behavior. SER is intrinsically language-specific; this study investigated a novel cascaded deep learning (DL) model to Bangla SER with intensity level. The proposed method employs the Mel-Frequency Cepstral Coefficient, Short-Time Fourier Transform (STFT), and Chroma STFT signal transformation techniques; the respective transformed features are blended into a 3D form and used as the input of the DL model. The cascaded model performs the task in two stages: classify emotion in Stage 1 and then measure the intensity in Stage 2. DL architecture used in both stages is the same, which consists of a 3D Convolutional Neural Network (CNN), a Time Distribution Flatten (TDF) layer, a Long Short-term Memory (LSTM), and a Bidirectional LSTM (Bi-LSTM). CNN first extracts features from 3D formed input; the features are passed through the TDF layer, Bi-LSTM, and LSTM; finally, the model classifies emotion along with its intensity level. The proposed model has been evaluated rigorously using developed KBES and other datasets. The proposed model revealed as the best-suited SER method compared to existing prominent methods achieving accuracy of 88.30% and 71.67% for RAVDESS and KBES datasets, respectively.**

***Keywords*—Bangla speech emotion recognition; speech signal transformation; convolutional neural network; bidirectional long short-term memory**

I. Introduction

Speech is the most commonly used method of interaction and emotional expression. In artificial intelligence and deep learning, the Speech Emotion Recognition (SER) task involves identifying and categorizing emotional elements of speech. The SER task has two basic steps: extracting features from the speech signal and classifying emotions. Existing SER methods employ different signal transformation and feature extraction methods on speech signals and then employ different machine learning (ML) and/or deep learning (DL) methods to the feature-transformed signal for emotion recognition [1] [2]. SER performance relies on the quality of the extracted emotional features from speech and methods used to classify the features.

In the ML or DL domain, SER is a language-specific research task, as natural languages have remarkable variations. Most of the existing SER tasks are performed in a few languages for which quality corpuses are available. On the other hand, SER studies are very limited, even for major languages, due to the legging of resources. As an example, Bangla is a major speaking language in the world, which is spoken by more than 210 million people [3]. However, Bangla is not resourceful for SER studies; hence, a few studies are available on Bangla SER. Therefore, the Bangla SER study is a promising research domain in ML and DL domains.

The study aims to develop a DL-based Bangla SER (BSER) system to identify the most suitable match for emotional speech. Existing BSER studies have only looked at classifying emotions from speech without considering the intensity level. However, the level of intensity (normal or strong) of an emotional expression such as sadness or anger is crucial. When someone is experiencing intense emotions, they may engage in harmful behaviors. The recent increase in online broadcasts of such behavior on social media has raised the alarm and the need for action to address these issues [4]. As a result, the automatic recognition of both emotions and their intensity level has become a highly relevant and essential issue that is the focus of this study.

An integrated speech signal transformation and a cascaded model with hybrid DL architecture are considered in this study to achieve better SER performance. The proposed method employs the Mel-Frequency Cepstral Coefficient (MFCC), Short-Time Fourier Transform (STFT), and Chroma STFT (CSTFT) feature transformation on the speech signals individually. The transformed two-dimensional (2D) MFCC, STFT, and CSTFT features are combined into a 3D form and inputted into the cascaded DL model for emotional classification with intensity. The cascaded DL model performs the task in two stages: classify emotion first in Stage 1 and then measure the intensity of the classified emotion in Stage 2. DL architectures used in both stages are the same, which consist of 3D Convolutional Neural Network (CNN), a Time Distribution Flatten (TDF) layer, a Long Short-term Memory (LSTM), and a Bidirectional LSTM (Bi-LSTM). The CNN extracts features from the 3D transformed speech signal using four feature blocks in series, and the TDF layer then flattens the features. The flattened features are inputted into the Bi-LSTM, the outcomes of Bi-LSTM are inputted into the LSTM, and finally, a fully connected layer is used for emotion classification or intensity measure. The proposed model has

been evaluated rigorously using our developed KBES dataset and other BSER datasets. The proposed cascaded revealed as the best-suited SER method while compared to existing prominent methods. The major contributions of this study are briefly summarized as follows:

1) Integrated 3D speech feature formation using MFCC, STFT, and CSTFT;

2) Emotion classification with intensity using cascaded DL model with hybrid DL architecture;

3) Rigorous experimental studies with different feature transformations and different DL models on our developed KBES dataset and other BSER datasets; and

4) Performance comparison of the proposed BSER model with existing prominent methods.

The structure of the remaining paper is as follows. Section II is the literature review, briefly explaining several related SER studies. Section III demonstrates the proposed method illustrating individual components. Experimental studies, including outcomes of the proposed cascaded model and performance comparison with the existing methods, are placed in Section IV. Finally, Section V concludes the paper with a few remarks on the present study and several future research directions.

II. Literature Review

Several DL-based SER studies are available with different corpuses for English and Germany, such as RAVDESS and IEMOCAP. Badshah et al. [5] investigated a CNN architecture composed of three convolutional layers and three fully connected (FC) layers. The model was trained on the Berlin emotional corpus to differentiate seven emotions using spectrograms collected from the stimuli. Satt et al. [6] introduced an SER model extracting log-spectrogram feature vectors from the IEMOCAP speech dataset. They tested two architectural models: convolution-only and convolution-LSTM deep neural networks. Etienne et al. [7] also used a CNN-LSTM architecture to classify emotions using spectrogram information. Chen et al. [8] used a portion of the IEMOCAP dataset for three different models: a shallow CNN combined with a Bi-LSTM, a deep CNN with a shallow Bi-LSTM, and a deep CNN with a deep Bi-LSTM. Manohar and Logashanmugam [9] integrated different meta-heuristic and deep-learning methods for SER with selected features. Wen et al. [10] Introduced self-labeling feature frames in their DL-based SER study.

Zhao et al. [11] investigated different DL-based models with CNN and LSTM for SER using the Emo-DB corpus and part of the IEMOCAP corpus. Among various combinations of convolutional layers and LSTM, the combination of 6 convolutional layers and LSTM is found to be the best-performing one. Zhang et al. [12] developed an attention-based, fully convolutional network on the IEMOCAP corpus and claimed to outperform state-of-the-art models. Ghosal et al. [13] proposed a graph neural network-based technique called the Dialogue Graph Convolutional Network (DialogueGCN) for recognizing emotions in conversations. A 2D CNN architecture extracted features from audio recordings, and a Support Vector Machine (SVM) was utilized for emotion classification. They evaluated the performance of their architecture on three datasets: IEMOCAP, AVEC, and MELD. Zhao et al. [14] employed a Connectionist Temporal Classification (CTC) with attention-based Bi-LSTM using the IEMOCAP dataset. Zhao et al. [15] combined Bi-LSTM with a CNN using the IEMOCAP dataset. Recently, Is-lam et al. [4] developed a 3D transformed feature and CNN Bi-LSTM-based cascaded DL model for the RAVDESS dataset.

Only a few attempts have been made to develop SER for Bangla, even though it is a significant natural language worldwide. Sultana et al. [16] employed a DL model combining CNN with TDF and Bi-LSTM layers to analyze the SUBESCO [17] dataset. The model consists of four CNN feature blocks (FB), each with a convolution layer, as well as layers for batch normalization (BN), exponential linear units (ELU), and max-pooling. The first two FBs used 128 kernels, while the last two FBs used 64 kernels. The proposed model's performance of Bangla SER was compared to three other models [18] [19] [20]. Sultana and Rahman [21] recently investigated essential speech features for ML-based BSER using the SUBESCO dataset.

III. Cascaded DL Model for Bangla SER With Intensity

There is a novel aim to develop Bangla SER with intensity. Speech is the primary method of communication in various live media platforms like Facebook and YouTube, which often express emotions. In the case of emotional expression, the level of intensity (normal or strong) has a crucial impact. For example, extreme sadness or anger may lead a person towards harmful behaviors, even suicidal effects. Identifying harmful emotions, such as extreme sadness or anger, is a challenging computational intelligence task to prevent individuals from taking harmful actions that could impact society.

Fig. 1 illustrates the general architecture of the proposed method of identifying appropriate emotions from speech signals. The main components of the pro-posed approach are the feature transformation of the input speech signal and classifying emotions and their intensity level. The method developed an integrated 3D feature form from three different individual features (i.e., MFCC, STFT, and Chroma STFT) on the input speech signal. Then, an appropriate DL model is used to analyze 3D features through multiple layers, eventually leading to the classification of emotions. At a glance, the proposed system takes in speech signals as input and generates output that classifies emotions as Neutral, Happy, Sad, Angry, and Disgust and further categorizes each emotion as Low or High based on the intensity level. The following subsections explain the individual components briefly.

A. Speech Signal Transformation

The speech signal transformation in the proposed system considers three popular signal transformation methods: Mel-Frequency Cepstral Coefficients (MFCC), Short-Time Fourier Transform (STFT), and Chroma STFT. Each MFCC, STFT, and Chroma STFT produces three different same-sized two-dimensional (2D) transformed features. The three 2D features

are integrated into a 3D transformed feature, which is inputted into the cascaded DL model for emotion recognition. The basic elaborated descriptions of the feature transformation are available in the literature [4].

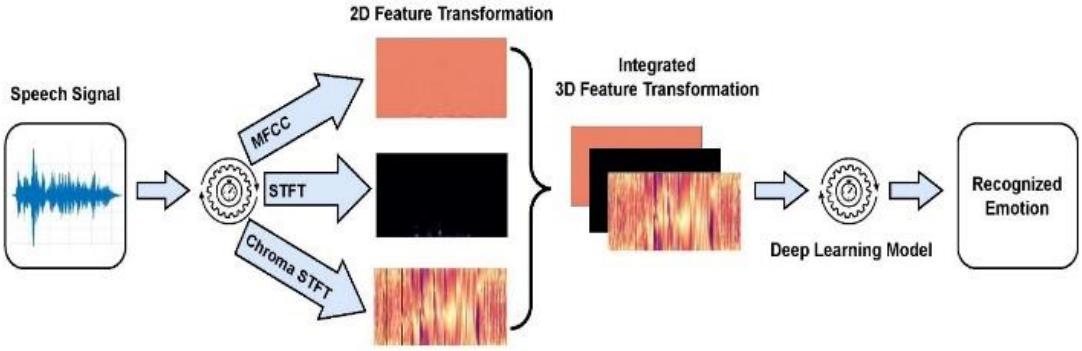
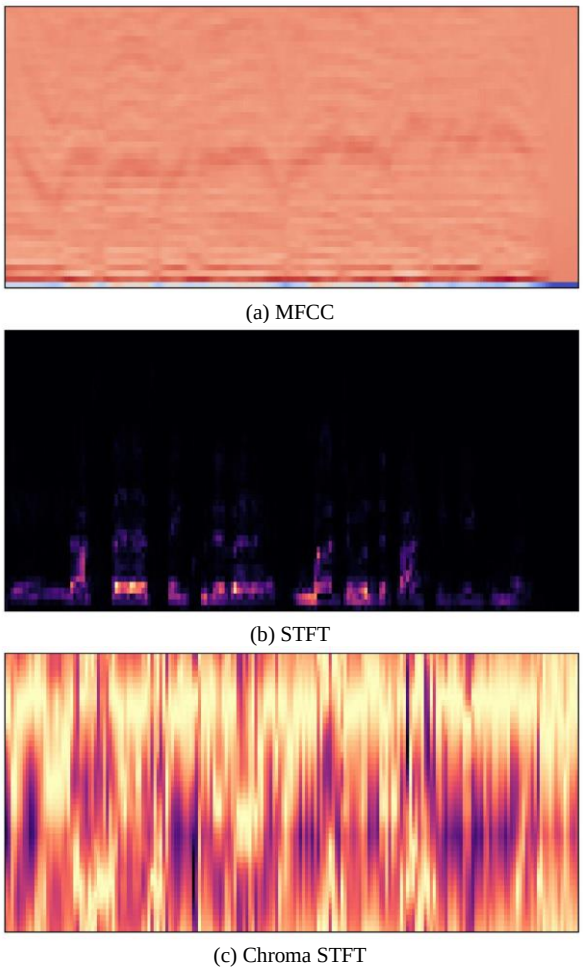


Fig. 1. Architecture of the proposed method identifying emotions from speech signals.



(a) MFCC

(b) STFT

(c) Chroma STFT

Fig. 2. Sample outcomes of speech signal feature trans-formation employed in the proposed method.

Fig. 2 shows 2D features from a sample speech signal using MFCC, STFT, and Chroma STFT. 250 hop lengths and 128 sequences were considered for MFCC; thus, the MFCC feature (see Fig. 2(a)) is with 128 sequences and 513 frames. Having a hop length of 250 and the FFT window length of 254, the STFT feature (see Fig. 2(b)) is also 128 sequences and 513 frames. The Chroma STFT feature is the same size as STFT, as shown in Fig. 2(c). Afterward, the three 2D features were integrated into the 3D form with size 128*513*3, which fed into the DL model.

B. Emotion Classification using Cascaded DL Model

Recognizing emotions using DL involves extracting features through multiple layers before finally recognizing them into emotion categories with intensity levels. A cascaded model is considered in this study to perform the task in two different stages: emotion classification and then the intensity of the classified category. Such a cascaded model is investigated in the recent study [4], but we have considered different DL architectures to achieve better performance for our intended application of Bangla SER.

Fig. 3 depicts the proposed cascaded DL model, which consists of two different stages. The DL architecture in Stage 1 classifies 3D-transformed features into five emotion categories: Neutral, Happy, Sad, Angry, and Disgust. Stage 2 uses another DL model to identify the intensity (i.e., Normal or Strong) of the emotion recognized in Stage 1. A single DL model is trained with the whole training set for Stage 1. For intensity classification in Stage 2, the training set was divided into five individual classes (i.e., Normal and four emotion classes), and then four distinct DL models were trained with individual emotion case samples for intensity classification. In the figure, DLEmotion in Stage 1 is the classifier for the emotion classification, and DLHappy, DLSad, DLAngry, and DLDisgust in Stage 2 are the classifiers for the intensity classification of Happy, Sad, Angry, and Disgust emotions, respectively. In other words, the DLEmotion extracts features from the input 3D transformed feature and then categorizes them into emotion cases. Similarly, a DL in Stage 2 extracts features from the same input used in Stage 1 but for the task of intensity measure (Low or High) of the classified emotion.

The main distinction between the DL architectures used in Stage 1 and Stage 2 is the number of output neurons: five in Stage 1 and two in Stage 2. Although the number of output classes and tasks differ, the basic structure is the same in all the DL architectures depicted in Fig. 4. Fig. 4(a) shows the building blocks of the DL architecture, which has four 3D CNN feature blocks (FB), a Time Distribution Flatten (TDF) layer, a Bi-LSTM layer, an LSTM layer, and a fully connected (FC) dense layer. The structure of a single CNN feature block

is shown in Fig. 4(b), which consists of a 3D CNN layer, batch normalization layer, ReLU activation layer, and 3D max-pooling layer. The use of 3D trans-formed features makes the 3D CNN ideal for feature extraction, while the LSTMs are selected for their ability to handle sequential data. Output from the CNN is passed through a TDF layer that serves as the input for the Bi-LSTM and then passed to the LSTM layer. The LSTM layer results are given to an FC layer to identify emotions or intensity levels.

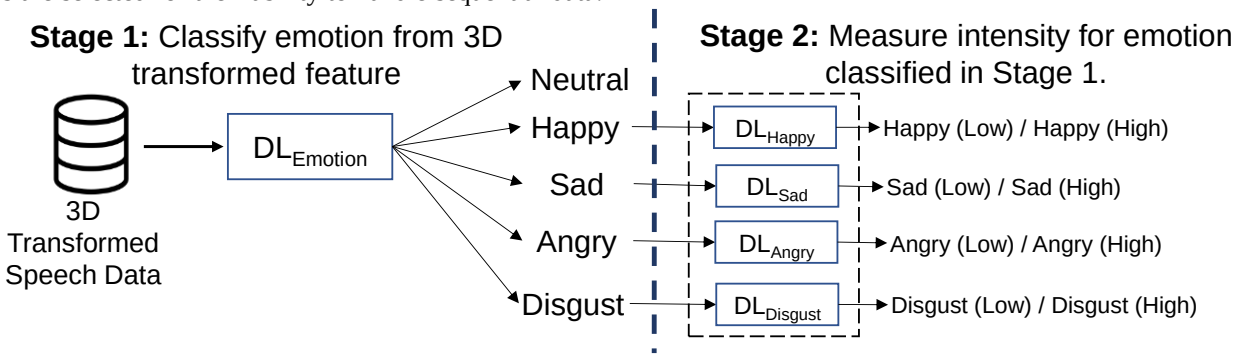
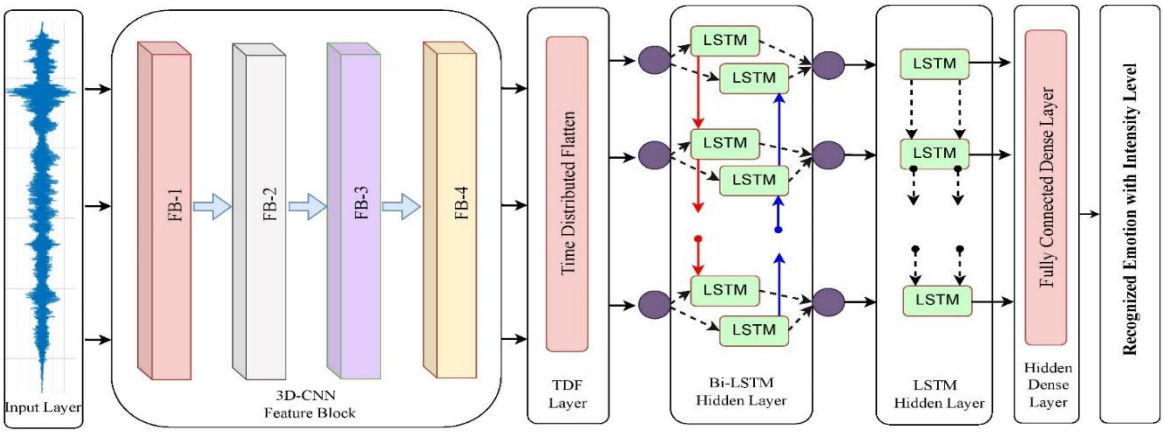
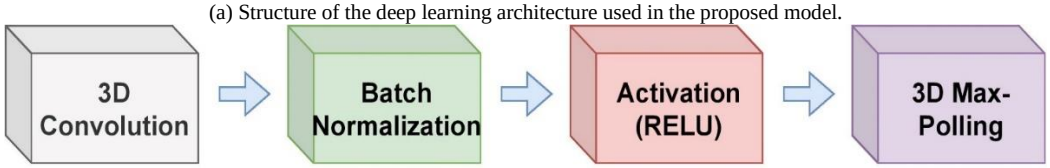


Fig. 3. The proposed cascaded DL model to classify emotion and its intensity from the 3D transformed speech signal.



(a) Structure of the deep learning architecture used in the proposed model.



(b) Structure of the convolutional feature block (FB) in 3D-CNN.

Fig. 4. The architecture of the proposed deep learning model.

TABLE I. THE HYPER-PARAMETERS OF THE PROPOSED DEEP LEARNING MODEL ARCHITECTURE

Feature Blocks	Layers	Input Shape	Output Shape	Filters	Kernel Size	Strides	Units
FB-1	Conv3D	128×513×3	128×513×3 @ 64	64	5×5×5	1×1×1	-
	MaxPolling3D	128×513×3 @ 64	64×256×3 @ 64	-	2×2×1	2×2×1	-
FB-2	Conv3D	64×256×3 @ 64	64×256×3 @ 64	64	5×5×5	1×1×1	-
	MaxPolling3D	64×256×3 @ 64	16×64×3 @ 64	-	4×4×1	4×4×1	-
FB-3	Conv3D	16×64×3 @ 64	16×64×3 @ 128	128	5×5×5	1×1×1	-
	MaxPolling3D	16×64×3 @ 128	4×16×3 @ 128	-	4×4×1	4×4×1	-
FB-4	Conv3D	4×16×3 @ 128	4×16×3 @ 128	128	5×5×5	1×1×1	-
	MaxPolling3D	4×16×3 @ 128	1×4×3 @ 128	-	4×4×1	4×4×1	-
-	TDF	1×4×3 @ 128	1×1536	-	-	-	-
-	Bi-LSTM	1×1536	1×512	-	-	-	256
-	LSTM	1×512	256	-	-	-	256

-	Dense	256	5/2	-	-	-	5/2

The proposed DL architecture has a set of layer parameters outlined in Table I. All CNN layers employ the same kernel size of 5×5×5 and strides of 1×1×1. The first feature block, FB-1 and FB-2, have filters of 64, while the last two feature blocks, FB-3 and FB-4, have filters of 128. The max-pooling layer for the FB-1 has a 2×2×1 sized kernel and a stride size of 2×2×1, whereas the last three max-pooling layers for the feature block have a 4×4×1 sized kernel and a stride of 4×4×1. The output of FB-4 is passed into the TDF layer, which is transformed into 1536 features. The outputs of the TDF layer are served as input to the Bi-LSTM layer. The Bi-LSTM layer outputs are 512 features, which are passed on to the LSTM layer. The outputs of the LSTM layer pass into the FC dense layer. The dense layer was concluded by applying the SoftMax activation function, which normalized the recognition of emotions in the classification process. Finally, the output layer will have five neurons in the DL classifier in Stage 1, and DL classifiers in Stage 2 will have two neurons in their output layers. The elaborated descriptions of CNN, LSTM, and Bi-LSTM are available in the literature [4] [21].

C. Significance of the Proposed Method

Bangla SER with intensity level is the primary concern of this study. This study utilized a 3D form of the speech signal that integrates three 2D transformations to enhance features and improve identification, resulting in significant performance enhancements. A cascaded DL model is considered in this study, where the task is split into two stages. The first stage involves classifying the emotion using a DL (i.e., $DL_{Emotion}$) model, and in the second stage, distinct DL models (i.e., DL_{Happy}, DL_{Sad}, DL_{Angry}, and $DL_{Disgust}$) classify the intensity level of the classified emotion as either Normal or Strong. DL models in both stages use the same 3D feature as input for a particular speech sample. The output of the first stage's DL model identifies the emotion class of the sample based on its 3D features. In contrast, a particular emotion-based DL model measures its intensity level using the same 3D feature in Stage 2. DL architectures in both stages are the same and consist of a 3D CNN, Bi-LSTM, and LSTM. This proposed model differs from the existing methods by using innovative techniques to recognize individual emotions and their intensity level.

IV. Experimental Studies

The effectiveness of the proposed DL method has been evaluated through several series of experiments and compared with other prominent methods. Since intensity level identification is the main attraction of this study for Bangla SER and no existing dataset is appropriate for the task, an appropriate dataset is pre-pared for this study. The following sections provide information regarding the dataset development, experimental setup, result analysis, and performance comparison.

A. Bangla SER Dataset Development

An appropriate dataset is a major issue in developing any ML/DL method. The major concern of this study is not only the categorization of emotions from speech signals but also the intensity of the categorized emotion. The intensity of speech emotion has great impacts, which has already been discussed as the motivation of the present study. A few Bangla SER datasets are available, but no existing dataset considers the intensity level of emotions. Moreover, existing datasets consist of only a few dialogues with a minimal number of actors; such datasets are unsuitable concerning re-al-world scenarios. Therefore, a realistic Bangla emotional speech dataset is used in this study, which is called the KUET Bangla Emotional Speech (KBES) dataset [22]. The dataset consists of 900 audio signals (i.e., speech dialogs) from 35 actors (20 females and 15 males) with diverse age ranges. Sources of the speech dialogs are Bangla Telefilm, Drama, TV Series, and Web Series. There are five emotional categories: Neutral, Happy, Sad, Angry, and Disgust. Except for Neutral, samples of a particular emotion are divided into two intensity levels: Normal and Strong.

The significant issue of the KBES dataset is that the speech dialogs are almost unique, with a relatively large number of actors. In contrast, existing datasets (such as SUBESCO and BanglaSER) hold samples with repeatedly spoken few pre-defined dialogs by a few actors/research volunteers in the lab environment. Finally, the KBES dataset is exposed as a nine-class problem to classify emotions into nine categories: Neutral, Happy (Low), Happy (High), Sad (Low), Sad (High), Angry (Low), Angry (High), Disgust (Low) and Disgust (High). However, the dataset is kept symmetrical, holding 100 samples for each of the nine classes; 100 samples are also gender balanced, having 50 samples for male/female actors. The speech signals in the KBES dataset have a length of three seconds and a sample rate of 48 kHz. The developed dataset seems a realistic one when compared with the existing SER datasets. The developed KBES dataset is publicly available and recently published data article holds its description [22].

B. Experimental Setup

The Python Librosa library [23] was used to trans-form speech signals of the KBES dataset. To make speech signals (having a length of three seconds and a sample rate of 48 kHz) in the KBES dataset suitable for DL input, the signal samples were resampled to 42.66 kHz. As a result, a speech signal was transformed into 127,980 bit-vectors. By adding 20 bits of zeros with it, 128,000 bit-vectors were used as input for the DL model. The DL models were developed using open-source Python libraries, specifically Keras [24] and Tensorflow [25].

The experiments were executed using a Jupyter Notebook hosted on Google Colaboratory, a cloud service that provides GPU capability [26]. The personal computer was an HP Pavilion laptop with an Intel(R) Core i5-7200U CPU @ 2.50 GHz processor, 8 GB of RAM, and an NVIDIA GeForce 940MX Graphics Card.

We utilized the Adam optimizer with a learning rate of 10−4 and employed the categorical cross-entropy as the loss function for compiling the model. The model aims to reduce the categorical cross-entropy loss, a widespread loss function for multi-class classification problems [27]. During the training process, the model was trained for 60 epochs with a batch size of 16.

The KBES dataset was separated into a training set and a test set, with 80% of the samples used for training and the remaining 20% used for testing. The training set consisted of 720 samples collected from each of the nine categories and was utilized for training the model. The test set, made up of the remaining 180 samples, was kept aside to assess the model's generalizability after training. Since the major concern of any ML/DL model is to perform well beyond the training data (i.e., on unseen data), test set performance (generalization) is a key performance measure of the developed model.

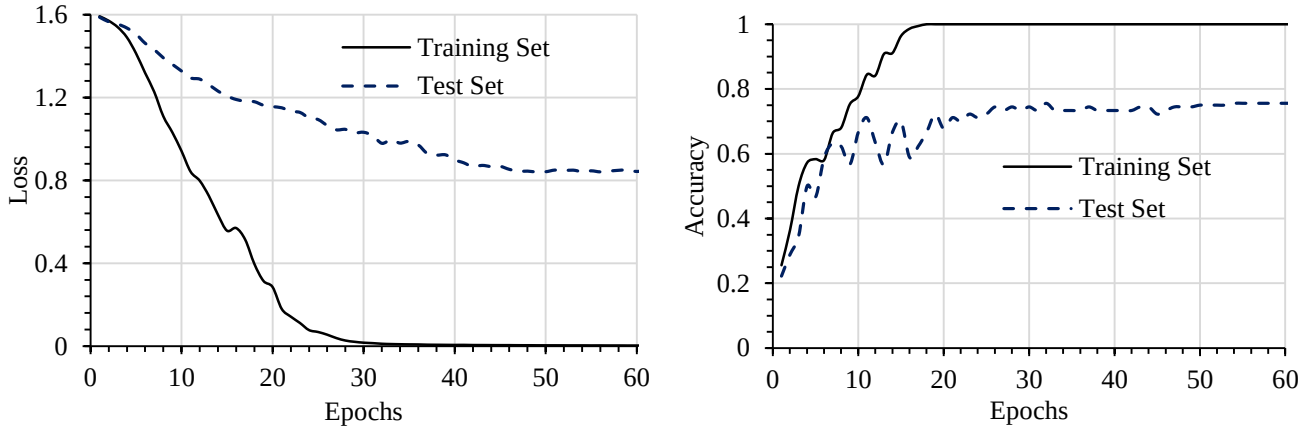
C. Outcomes of the Proposed Cascaded DL Model

The proposed cascaded model is carried out in two stages. In Stage 1, a DL model is trained with 720 training samples and tested with 180 samples for emotional classification. Fig. 5 shows the loss and accuracy curves of both the training and test sets in Stage 1 (i.e., for DLEmotion). The training set loss reduces smoothly and reaches close to zero after 30 epochs, while the test set loss reduces initially and then stabilizes after 45 epochs. Consequently, accuracy reached 100% for the training set through smooth improvement, while the best test set accuracy was close to 76%.

Table II shows the confusion matrix on the test set on emotion classification only (without intensity) in Stage 1. For example, for 40 happy samples, the model correctly identified 28. The remaining 12 happy samples were misclassified as Neutral, Sad, Angry, and Disgust in cases 5, 1, 2, and 4, respectively. There is a total of 45 misclassifications, which are in category misses. With a true classification of 135 samples, the DL model shows an accuracy of 75.56% in emotion classification, regardless of the intensity.

In Stage 2 of the proposed DL model, the samples classified in Stage 1 were used to determine the emotional intensity with a DL model for a particular emotion category. In this stage, four DL models are trained, and a particular DL is trained with the samples for a particular emotion category. As an example, there are a total of 160 training samples for Happy emotion (80 samples for each of Normal and Strong intensity cases) in 720 training samples. These 160 happy samples were used to train the DL model (i.e., DLHappy) to classify the intensity of Happy emotion. The intensity level of Happy emotion was tested using 40 samples from the test set.

Similarly, the other three different DL models (i.e., DLSad, DLAngry, and DLDisgust) were trained and tested for intensity level classification of Sad, Angry, and Disgust emotions. In Stage 2, the intensity of a sample will be classified by the designated DL model for the classified emotion in Stage 1. If a sample is misclassified in Stage 1 (i.e., a Category Miss), there is no remedy in Stage 2, and treated as Category Miss. On the other hand, a truly classified sample in Stage 1 may be wrongly classified in the case of Intensity Level in Stage 2, which is called Intensity Miss. Therefore, due to the inclusion of intensity miss in Stage 2, the classification accuracy of emotion with intensity is lower than the accuracy of emotion classification by Stage 1.



(a) Loss vs epochs (b) Accuracy vs epochs

Fig. 5. Loss and accuracy curves of training and test sets of the DL model in Stage 1 in the proposed cascaded DL model.

TABLE II. TEST SET CONFUSION MATRIX ON EMOTION CLASSIFICATION (WITHOUT INTENSITY) IN STAGE 1 OF THE CASCADED DL

Emotion	Neutral	Happy	Sad	Angry	Disgust	Correctly Classified	Category Miss	Total Samples
Neutral	13	1	1	3	2	13	7	20
Happy	5	28	1	2	4	28	12	40
Sad	2	2	32	1	3	32	8	40
Angry	0	0	0	35	5	35	5	40
Disgust	1	0	3	8	28	28	12	40
Total						**136**	**44**	**180**

TABLE III. THE TEST SET CONFUSION MATRIX ON EMOTION RECOGNITION WITH INTENSITY LEVEL BY THE CASCADED DL MODEL

Emotion	Emotion Intensity	Neutral	Happy		Sad		Angry		Disgust		Category Miss	Intensity Miss	Total Samples
			Low	High	Low	High	Low	High	Low	High			
Neutral	-	13	1	0	1	0	2	1	2	0	7	0	20
Happy	Normal	3	14	1	0	0	0	0	2	0	5	1	20
	Strong	2	0	13	0	1	2	0	0	2	7	0	20
Sad	Normal	2	1	0	15	0	0	0	1	1	5	0	20
	Strong	0	0	1	1	16	1	0	0	1	3	1	20
Angry	Normal	0	0	0	0	0	15	2	2	1	3	2	20
	Strong	0	0	0	0	0	0	18	0	2	2	0	20
Disgust	Normal	1	0	0	2	0	3	1	11	2	7	2	20
	Strong	0	0	0	0	1	0	4	1	14	5	1	20
Total											**44**	**7**	**180**

Table III depicts the confusion matrix of the pro-posed cascaded model with Stage 1 + Stage 2. The ma-trix reveals both category and intensity miss for the four emotions. As Neutral emotion has no intensity measure, it holds only category miss, and the number of misclassifications is seven as of Table II. Again, out of the 28 correctly classified Happy emotion samples in Stage 1, one sample of Happy (Low) was misclassi-fied as Happy (High) in Stage 2; thus, Intensity Miss of Happy is one. The total category miss of the Happy was 12 in Table II, which just distributed for Happy (Low) and Happy (High) in Table III. Among category miss cases, three Happy (Low) were misclassified as Neutral, which is the highest misclassification as Neutral. Such misclassification of Happy samples is logical as it is generally accepted that Happy (Low) is close to Neu-tral. On the other hand, no Angry samples are misclas-sified as Neutral or Happy, and five Angry samples are misclassified as Disgust. Including proficiency in Sad and Disgust samples, the performance of the proposed cascaded DL model on an unseen test set of unique KBES datasets is remarkable. With a total of 51 misclassifications (=44 Category Miss +7 Intensity Miss) out of 180 test speeches, the resulting test set accuracy is 71.67% (i.e., (180-51)/100).

D. Preference for Cascaded DL Model over Single DL Model

The proposed cascaded model consists of five hybrid CNN-LSTM DL components where individual compo-nent performs different tasks. The individual DL com-ponents classified the samples into nine categories through task division. It is also possible to perform the task in a single stage by a single DL model having nine neurons in its output layer. Therefore, it is a common query to select a cascaded model with five DL components over the single DL model. Additional experiments with a single DL have been conducted to observe the effectiveness of the proposed cascaded model over a single DL. Experiments were also conducted with samples of male and female actors separately for better observations.

Table IV compares the performance of the proposed cascaded model with a single DL model for a full test set, male and female actors' samples. While full set training and test samples are 720 and 180 as of previous experiments, as the KBES dataset is balanced for male and female cases, training and test samples for males or females are 360 and 90, respectively. Ac-cording to the table, performances for male and female cases are competitive in both single DL and cascaded DL models. Furthermore, both DL models (i.e., single or proposed cascaded) show competitive performance in case of category miss. As an example, both the models misclassified equal 21 samples for the female case. The remarkable performance distinction between the two models is observed in the case of intensity classifica-tion, and the cascaded model remarkably outperformed the single DL model. The intensity missed cases by single DL model for male, female, and full set cases are 14, 13, and 23, respectively. The intensity missed cases for male, female, and full set by the cascaded model are 4, 5, and 7, respectively. Due to such outperformance in intensity classification, the overall accuracy of the cascaded DL model is much better than the single DL model. At a glance, the proposed cascaded DL model showed a full test set accuracy of 71.67% (as also discussed in the previous section), whereas the single DL model's accuracy on a full test set is 61.11% (= (180-47-23)/180). The outperformance of the cascaded DL model reflects the effectiveness of its task division in two stages as well as validates the appropriateness of selecting the cascaded model.

E. Performance Comparison with Different Speech Signal Transformations

The integrated 3D form of feature formation using MFCC, STFT, and CSTFT from the input speech signal is a crucial task of the proposed method. To observe the effectiveness of using different feature forms, additional experiments have been conducted with different signal transformations individually and combining two or more among MFCC, STFT, and CSTFT. Table V provides performance results for a cascaded DL model using three individual speech transformation features (MFCC, STFT, and CSTFT), three double combinations (MFCC + STFT, STFT + CSTFT, and MFCC + CSTFT), and finally, proposed triple combination MFCC + STFT+ CSTFT. Appropriate DL architectures were con-sidered to classify emotions using distinct feature transformations. Among the individual methods, MFCC achieved the highest accuracy of 60.56%. However, the dual combination of MFCC and STFT has an accuracy of 61.67%, which is slightly better than the accuracy of MFCC. Finally,

the integrated 3D form of transformed MFCC, STFT, and CSTFT (i.e., MFCC + STFT+ CSTFT) achieved the best accuracy of 71.67%. The performance comparison justifies the use of a 3D form of a feature in this study.

TABLE IV. PERFORMANCE ON FULL TEST SET, MALE AND FEMALE ACTORS FOR SINGLE DL MODEL AND PROPOSED CASCADED DL MODEL

Particular of Samples	Training & Test Samples	Test Set Performance by the Single Deep Learning Model			Test Set Performance by the Cascaded Deep Learning Model		
		Category Miss	Intensity Miss	Accuracy (%)	Category Miss	Intensity Miss	Accuracy (%)
Male	360 & 90	18	14	64.44	20	4	73.30
Female	360 & 90	21	13	62.22	20	5	72.22
Full Set (= Male + Female)	720 & 180	47	23	61.11	44	7	71.67

TABLE V. PERFORMANCE COMPARISON ON THE TEST SET WITH DIFFERENT SIGNAL TRANSFORMATIONS FOR CASCADED DEEP LEARNING MODEL

Combination	Signal Transformation	Sequences × Frames ×Dimension	Category Miss	Intensity Miss	Truly Classify	Accuracy (%)
Individual	Mel-Frequency Cepstral Coefficient (MFCC)	128×513×1	50	21	109	60.56
	Short-Time Fourier Transform (STFT)	128×513×1	53	27	100	55.56
	Chroma STFT (CSTFT)	128×513×1	62	33	85	47.22
Double	MFCC+STFT	128×513×2	53	16	111	61.67
	STFT+CSTFT	128×513×2	57	27	96	53.33
	MFCC+CSTFT	128×513×2	50	23	107	59.44
Proposed Triple	MFCC+STFT+CSTFT	128×513×3	44	7	129	71.67

F. Performance Comparison Varying DL Architecture

The DL architecture employed in this study to classify emotions from 3D feature maps is 4 3D-CNN + TDF + Bi-LSTM + LSTM, which is a hybrid CNN and LSTM model. In the model, 3D CNN with four layers is the main component to extract features to classify the input 3D feature sample. It is also possible to perform classification from CNN features using one/two FC layers. Distinct experiments with TDF, LSTM, and Bi-LSTM options have been conducted, and Table VI compares performance on the test samples classification by the considered DL architectures. The comparison in Table VI indicates that our proposed model achieves the highest accuracy over all other models. It is seen from the table that the 4 3D-CNN + 2 FC architecture is the simplest one and showed the lowest ac-curacy, which is 62.78%. The architecture holds two FC layers on 3D CNN, while other architectures hold one FC layer by default on top of the relatively complex DL architectures. In general, accuracy seems to be improved with architectural complexity. Finally, the considered hybrid architecture of CNN, LSTM, and Bi-LSTM shows the best accuracy of 71.67%. The performance comparison justifies the employment of the hybrid DL architecture 4 3D-CNN + TDF + Bi-LSTM + LSTM to achieve better emotion recognition using the KBES dataset.

G. Performance Comparison with Existing SER Meth-ods

According to our knowledge, this study is the pioneer for Bangla SER with intensity. The barrier was the speech emotion dataset with intensity measure, which is resolved in this study by developing the KBES dataset. Recent Bangla SER studies employed BanglaSER and SUBESCO datasets for emotion classification regardless of the intensity issue [16] [17]. On the other hand, the RAVDESS dataset for the English language has intensity identification and used SER with intensity in the recent study by Islam et al. [4]. In the following subsections, the proposed model (i.e., Stage 1) is com-pared with several existing methods for emotional classification only on BanglaSER, SUBESCO, RAV-DESS, and KBES datasets. Then, a rigorous comparison is performed for SER with intensity considering RAV-DESS and KBES datasets. Notably, the SER datasets hold different numbers of samples, and there are 1467, 7000, 1440, and 900 samples in BanglaSER, SUBESCO, RAVDESS, and KBES datasets, respectively. Although the KBES dataset holds the lowest number of samples, it is the most unique among other datasets regarding the number of individual speech dialogs, the number of individual actors, and the realistic manner of the samples.

Table VII compares emotional classification regard-less of intensity using Stage 1 of the proposed cascaded method with other existing methods on KBES and three other SER datasets: BanglaSER, SUBESCO, and RAVDESS. The test set accuracy for each method was evaluated using the 20% of reserved test speech samples for the individual datasets. The results presented in the tables show that the proposed method outper-forms all other methods for any dataset. For instance, the 7 CNN+2 FC [20] method showed an accuracy of 70.55% for BanglaSER, 78.14% for SUBESCO, 83.04% for RAVDESS, and 67.22% for KBES. In contrast, the proposed cascaded method achieved the highest accu-racy of 83.56% for BanglaSER, 90.14% for SUBESCO, 92.98% for RAVDESS, and 75.56% for KBES. Another observation from the table is that performance on the KBES dataset is the lowest, and the RAVDESS dataset is the highest among the datasets for any individual method. For example, the 4 CNN+TDF+Bi-LSTM [14] model showed the lowest accuracy of 72.78% for KBES and 88.89% for RAVDESS. BanglaSER, SU-BESCO, and RAVDESS hold a fixed number of speeches with repetition, whereas the KBES dataset

consists of a distinct number of speeches to form a realistic environment. Therefore, the lowest performance on the KBES dataset for Bangla SER is logical for such a realistic environment with challenging samples.

TABLE VI. TEST SET PERFORMANCE VARYING DEEP LEARNING ARCHITECTURES IN THE PROPOSED CASCADED MODEL

Deep Learning Architecture	Category Miss	Intensity Miss	Truly Classified	Accuracy (%)
4 3D-CNN+2 FC	52	15	113	62.78
4 3D-CNN+LSTM	51	13	116	64.44
4 3D-CNN+Bi-LSTM	48	14	118	65.56
4 3D-CNN+TDF+LSTM	47	9	124	68.89
4 3D-CNN+TDF+Bi-LSTM	46	10	124	68.89
4 3D-CNN+TDF+Bi-LSTM+LSTM	44	7	129	71.67

TABLE VII. COMPARISON OF EMOTION CLASSIFICATION WITHOUT INTENSITY LEVEL (STAGE 1) ON FOUR SER DATASETS

Work Ref., Year	Feature	DL Architecture	Test Set Accuracy (%)			
			BanglaSER	SUBESCO	RAVDESS	KBES
[20], 2019	Log-Mel-Spectrograms	7 CNN+2 FC	70.55	78.14	83.04	67.22
[18], 2019	Log-Mel-Spectrograms	4 CNN+LSTM	75.69	81.57	86.55	70.56
[16], 2022	Log-Mel-Spectrograms	4 CNN+TDF+Bi-LSTM	78.77	87.21	88.89	72.78
[4], 2022	MFCC + STFT + CSTFT	4 3D-CNN+TDF+Bi-LSTM	80.48	89.43	91.23	73.88
Proposed Method	MFCC + STFT + CSTFT	4 3D-CNN+TDF+Bi-LSTM + LSTM	83.56	90.14	92.98	75.56

TABLE VIII. COMPARISON OF EMOTION CLASSIFICATION WITH INTENSITY LEVEL ON RAVDESS AND KBES DATASETS

Work Ref., Year	Feature	DL Architecture	RAVDESS		KBES	
			Single DL Model	Cascaded DL Model	Single DL Model	Cascaded DL Model
[20], 2019	Log-Mel-Spectrograms	7 CNN+2 FC	69.01	74.27	51.11	60.56
[18], 2019	Log-Mel-Spectrograms	4 CNN+LSTM	73.68	79.53	53.89	64.44
[16], 2022	Log-Mel-Spectrograms	4 CNN+TDF+Bi-LSTM	77.19	83.04	57.22	67.78
[4], 2022	MFCC + STFT + CSTFT	4 3D-CNN+TDF+Bi-LSTM	80.12	86.55	58.89	69.44
Proposed Method	MFCC + STFT + CSTFT	4 3D-CNN+TDF+Bi-LSTM+LSTM	84.21	88.30	61.11	71.67

Table VIII compares emotion classification with intensity among the different DL architectures for both single DL and cascaded DL models on RAVDESS and KBES datasets as they hold intensity marked. The 7 CNN+2 FC [20] architecture showed accuracies for the RAVDESS dataset are 69.01% and 74.27% by the single DL and the cascaded DL models, respectively. For the KBES dataset, the same architecture achieved 51.11% and 60.56% by the single DL and the cascaded DL models, respectively. At a glance, the proposed cascaded DL model outperformed single DL for both datasets. Again, the proposed SER method with 3D features and hybrid DL architecture achieved higher accuracy than other existing methods. The proposed cascaded method achieved the best accuracy of 88.30% for the RAVDESS dataset and 71.67% for the KBES dataset. The achieved performances on the RAVDESS dataset are also better or more competitive than the reported results in the recent study [4]. The outcomes for KBES are also remarkable due to its complexity and uniqueness in the samples. Finally, the proposed cascaded model using four 3D-CNN+TDF+Bi-LSTM+LSTM is the most suitable approach for Bangla SER with intensity.

V. CONCLUSIONS

This study has proposed automated recognition of emotion with intensity level from speech signals in Bangla, a globally popular and primary spoken language in South Asia with growing speakers worldwide in the era of trade and labor globalization. Since di-verse online social media and platforms have become substantial modes of verbal communication and emotional expression, SER is necessary to detect unusual or harmful human behaviors automatically. Therefore, emotion identification and its intensity of expression are significantly discussed in this study for Bangla through developing a novel DL-based cascaded model. The major steps of the proposed model are the trans-formation of speech signal 3D form integrating three different 2D transformations and then employing a cascaded DL model having hybrid DL architectures (consisting of CNN, Bi-LSTN, and LSTM) in its two stages for emotional classification and intensity level identification accordingly. For developing a realistic model, this study considered the KBES dataset, which holds natural speech samples, and its development is also a significant contribution to this study. Rigorous experiments have been conducted with KBES and other SER datasets in

Bangla and English. The proposed cascaded model has been identified as the best-performing SER method compared to several main-stream DL methods. At a glance, the proposed cascaded DL model showed its superiority over the existing Bagla SER methods, performing KBES test set accuracy of 71.67% for emotion classification with intensity and 75.56% regardless of speech intensity (i.e., excluding intensity miss cases).

Several prospective research scopes have been revealed from this study. KBES dataset is the only Bangla SER with intensity measure, which seems realistic with respect to other datasets. The best-achieved performance with the dataset (i.e., 75.56%) is lower than the other datasets. Therefore, it remained an open challenge to develop better realistic SER methods based on it, especially for Bangla. For simplicity, the same hybrid DL architecture and training process is used in both stages of the cascaded model; different DL architectures for emotion classification and intensity measures and individual fine tunings of different DL architectures might give better performance. Further-more, other speech signal transformations and integration techniques might perform better.

REFERENCES

[1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Communication, vol. 116. 2020, doi: 10.1016/j.specom.2019.12.001.

[2] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," Neurocomputing, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.

[3] "Bengali language," Britannica, 2023. [Online]. Available: https://www.britannica.com/topic/Bengali-language. [Accessed: 01-May-2023].

[4] M. R. Islam, M. A. H. Akhand, M. A. S. Kamal, and K. Yamada, "Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning," Electronics, vol. 11, no. 15, p. 2362, Jul. 2022, doi: 10.3390/electronics11152362.

[5] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in 2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings, 2017, doi: 10.1109/PlatCon.2017.7883728.

[6] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017, vol. 2017-Augus, doi: 10.21437/Interspeech.2017-200.

[7] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation," 2018, doi: 10.21437/smm.2018-5.

[8] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition," IEEE Signal Process. Lett., vol. 25, no. 10, 2018, doi: 10.1109/LSP.2018.2860246.

[9] K. Manohar and E. Logashanmugam, "Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm," Knowledge-Based Syst., vol. 246, p. 108659, Jun. 2022, doi: 10.1016/j.knosys.2022.108659.

[10] G. Wen et al., "Self-labeling with feature transfer for speech emotion recognition," Knowledge-Based Syst., vol. 254, p. 109589, Oct. 2022, doi: 10.1016/j.knosys.2022.109589.

[11] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in ASMMC-MMAC 2018 - Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and 1st Multi-Modal Affective Computing of Large-Scale Multimedia Data, Co-located with MM 2018, 2018, doi: 10.1145/3267935.3267948.

[12] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018 - Proceedings, 2019, doi: 10.23919/APSIPA.2018.8659587.

[13] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2019, doi: 10.18653/v1/d19-1015.

[14] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019, vol. 2019-Septe, doi: 10.21437/Interspeech.2019-1649.

[15] Z. Zhao et al., "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2928625.

[16] 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, 2022, doi: 10.1109/ACCESS.2021.3136251.

[17] 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, Apr. 2021, doi: 10.1371/journal.pone.0250173.

[18] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomed. Signal Process. Control, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.

[19] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, "Accurate EEG-Based Emotion Recognition on Combined Features Using Deep Convolutional Neural Networks," IEEE Access, vol. 7, pp. 44317–44328, 2019, doi: 10.1109/ACCESS.2019.2908285.

[20] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," Sensors (Switzerland), vol. 20, no. 1, Jan. 2020, doi: 10.3390/s20010183.

[21] M. A. H. Akhand, Deep Learning Fundamentals - A Practical Approach to Understanding Deep Learning Methods. Dhaka: University Grants Commission of Bangladesh, 2021.

[22] M. M. Billah, M. L. Sarker, and M. A. H. Akhand, "KBES: A Dataset for Realistic Bangla Speech Emotion Recognition with Intensity Level," Data Br., p. 109741, Oct. 2023, doi: 10.1016/j.dib.2023.109741.

[23] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," Proc. 14th Python Sci. Conf., no. Scipy, pp. 18–24, 2015, doi: 10.25080/majora-7b98e3ed-003.

[24] T. B Arnold, "kerasR: R Interface to the Keras Deep Learning Library," J. Open Source Softw., vol. 2, no. 14, p. 296, 2017, doi: 10.21105/joss.00296.

[25] M. Abadi, "TensorFlow: learning functions at scale," ACM SIGPLAN Not., vol. 51, no. 9, pp. 1–1, 2016, doi: 10.1145/3022670.2976746.

[26] F. R. V. Alves and R. P. Machado Vieira, "The Newton Fractal's Leonardo Sequence Study with the Google Colab," Int. Electron. J. Math. Educ., vol. 15, no. 2, 2019, doi: 10.29333/iejme/6440.

[27] Y. Zhou, X. Wang, M. Zhang, J. Zhu, R. Zheng, and Q. Wu, "MPCE: A Maximum Probability Based Cross Entropy Loss Function for Neural Network Classification," IEEE Access, vol. 7, pp. 146331–146341, 2019, doi: 10.1109/ACCESS.2019.2946264.