Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046623

ARTICLE

Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition

Fatma Harby[1], Mansor Alohali[2], Adel Thaljaoui[2,3,*] and Amira Samy Talaat[4]

[1]Computer Science Department, Future Academy-Higher Future Institute for Specialized Technological Studies, Cairo, 12622, Egypt

[2]Department of Computer Science and Information College of Science at Zulfi, Majmaah University, P. O. Box 66, Al-Majmaah, 11952, Saudi Arabia

[3]Preparatory Institute for Engineering Studies of Gafsa, Zarroug, Gafsa, 2112, Tunisia

[4]Computers and Systems Department, Electronics Research Institute, Cairo, 12622, Egypt

*Corresponding Author: Adel Thaljaoui. Email: adel.t@mu.edu.sa

Received: 09 October 2023 Accepted: 22 December 2023 Published: 27 February 2024

ABSTRACT

Machine Learning (ML) algorithms play a pivotal role in Speech Emotion Recognition (SER), although they encounter a formidable obstacle in accurately discerning a speaker's emotional state. The examination of the emotional states of speakers holds significant importance in a range of real-time applications, including but not limited to virtual reality, human-robot interaction, emergency centers, and human behavior assessment. Accurately identifying emotions in the SER process relies on extracting relevant information from audio inputs. Previous studies on SER have predominantly utilized short-time characteristics such as Mel Frequency Cepstral Coefficients (MFCCs) due to their ability to capture the periodic nature of audio signals effectively. Although these traits may improve their ability to perceive and interpret emotional depictions appropriately, MFCCS has some limitations. So this study aims to tackle the aforementioned issue by systematically picking multiple audio cues, enhancing the classifier model's efficacy in accurately discerning human emotions. The utilized dataset is taken from the EMO-DB database, preprocessing input speech is done using a 2D Convolution Neural Network (CNN) involves applying convolutional operations to spectrograms as they afford a visual representation of the way the audio signal frequency content changes over time. The next step is the spectrogram data normalization which is crucial for Neural Network (NN) training as it aids in faster convergence. Then the five auditory features MFCCs, Chroma, Mel-Spectrogram, Contrast, and Tonnetz are extracted from the spectrogram sequentially. The attitude of feature selection is to retain only dominant features by excluding the irrelevant ones. In this paper, the Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) techniques were employed for multiple audio cues features selection. Finally, the feature sets composed from the hybrid feature extraction methods are fed into the deep Bidirectional Long Short Term Memory (Bi-LSTM) network to discern emotions. Since the deep Bi-LSTM can hierarchically learn complex features and increases model capacity by achieving more robust temporal modeling, it is more effective than a shallow Bi-LSTM in capturing the intricate tones of emotional content existent in speech signals. The effectiveness and resilience of the proposed SER model were evaluated by experiments, comparing it to state-of-the-art SER techniques. The results indicated that the model achieved accuracy rates of 90.92%, 93%, and 92% over the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin Database of Emotional Speech (EMO-DB), and The Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets,

respectively. These findings signify a prominent enhancement in the ability to emotional depictions identification in speech, showcasing the potential of the proposed model in advancing the SER field.

KEYWORDS

Artificial intelligence application; multi features; sequential selection; speech emotion recognition; deep Bi-LSTM

1 Introduction

The significance of expressing emotions in human communication is of significant importance in effectively conveying information to the recipient. The precise recognition of emotions in real-time scenarios possesses a multitude of practical implications across diverse fields, augmenting the functionalities of systems and strengthening interactions between humans and machines. For instance, in virtual therapy, user emotions reorganization can help tailor interventions and replies to better address the emotional needs of individuals. Moreover, emotionally intelligent robots have the potential to offer support to persons with specific requirements by comprehending and appropriately reacting to their emotional signals. Correspondingly, in emergencies, recognition of emotions can be used in call centers to analyze the emotions of customers which assists customer service representatives to empathetically respond and address issues more effectually. In workplace environments, emotion recognition may play a significant role in evaluating the overall well-being of employees through the monitoring of stress levels and job satisfaction. This, in turn, can facilitate the development of more effective organizational initiatives aimed at assisting employees.

The various manifestations of human emotions are dynamic and lively. Nonverbal communication comprises a variety of modalities for conveying information, including but not limited to body language, facial expressions, eye contact, humor, and tone of voice. The linguistic diversity among the world's populations is evident. Nevertheless, it is noteworthy that individuals are capable of grasping certain aspects of the intended message being sent by their communication partner, even in the absence of a shared language. This phenomenon is often attributed to the utilization of emotional expressions, as previously indicated. Of the several modes of human emotional expression, vocal expression has received the most extensive scholarly attention.

The ML algorithms play a vital role in Speech Emotion Recognition (SER) by identifying and categorizing emotions expressed in speech. For real-time applications, models of Machine Learning (ML) can be optimized to let systems respond quickly to emotional state changes during conversations or interactions. However, ML faces numerous challenges in accurately discriminating emotional states from spoken language. Some of the crucial challenges include, emotions frequently being personal and context-dependent, as the utterance of identical words within varying circumstances or accompanied by distinct intonations can express diverse emotional states. Furthermore, the annotation of emotions does not have a universally accepted and established set of labels. The utilization of diverse datasets may include the adoption of varying emotion categories or labels, hence posing challenges in the process of data comparison and integration. The presence of this contradiction may impede the advancement of emotion recognition models that are both robust and capable of generalization.

Besides, emotions may not always exhibit distinct boundaries and sometimes exhibit overlapping characteristics. An individual may manifest a confluence of positive affect and astonishment.

Historically, the main challenge encountered during the feature extraction procedure revolved around the determination of a dependable methodology for obtaining noteworthy and differentiating features from voice signals. The aforementioned characteristics were designed to capture the affective state of a speaker by analyzing the acoustic properties of their speech. In the earlier years, a considerable number of academics have undertaken research endeavors about low-level handmade characteristics in the context of SER. The aforementioned properties comprise various factors, including energy, zero-crossing, pitch, linear predictor coefficient, MFCC, and nonlinear features such as the tiger energy operator. Presently, a considerable proportion of scholars employ deep learning methodologies in the context of SER. In this particular scenario, a frequently utilized input characteristic is a Mel-scale filter bank audio spectrogram. The successful development of Fully Convolutional Networks (FCNs) has been achieved through the use of Convolution Neural Network (CNNs) to effectively tackle the issue of accommodating input data with diverse dimensions. However, despite the remarkable performance of FCNs in tasks of time series classification with stable input variable sizes, they are limited in their ability to adequately capture and acquire temporal information related to this specific difficulty. Hence, the use of Long Short-Term Memory (LSTM) with Recurrent Neural Networks (RNNs) is employed to acquire comprehension of distinctive and temporally-dependent patterns within sequences [1].

Principally, Bidirectional Long Short Term Memory (Bi-LSTM) networks are a type of RNN architecture that are effective for sequential data processing and analysis. In the following, we will mention some of the crucial characteristics and capabilities of Bi-LSTM networks. Mainly, Bi-LSTM networks are designed for sequential data as the order of elements is vital as time series or natural language text. Also, the Bi-LSTM network's basic structure blocks are LSTM cells which can learn long-term dependencies in the data by keeping a memory state, as a result, it helps in catching relationships and patterns within sequences.

Additionally, one of the distinguishing features of Bi-LSTM network**s** is their ability to process bidirectional simultaneously which allows the model to catch information from past and future, which is predominantly valuable for tasks where context substances. Almost, CNNs have been widely employed in combination with LSTM and LSTM-RNNs in the domain of SER to capture latent temporal patterns [2].

Academic researchers are currently engaged in efforts to enhance the precision of SER through the utilization of the CNN-LSTM model. This approach aims to identify significant segments within speech signals and effectively capture temporal information [3]. In their study, they introduced a novel approach to identifying and categorizing emotional states in speech within the context of healthcare environments. The researchers adopt a methodology that leverages CNN features to enhance the ability of intelligent and efficient equipment to accurately perceive an individual's emotional state.

The field of SER has been a subject of active research in recent times. Researchers have increasingly employed deep learning techniques to devise diverse approaches for the identification and classification of emotional states shown by speakers. Researchers commonly employ the CNNs model to acquire significant and distinguishing features, which are subsequently inputted into the LSTM network. This enables the LSTM network to learn concealed temporal cues to identify emotions inside sequences. The utilization of CNNs and artificial intelligence has been shown to enhance recognition accuracy. However, it is crucial to emphasize that the employment of large network weights also leads to a rise in computational expenses. The current CNNs and LSTM architectures did not demonstrate significant improvements in enhancing the accuracy and decreasing the computational complexity of the current SER systems.

In the context of this study, Bi-LSTM networks contribute by effectively capturing dependencies and patterns in sequential data, which enables predicting future values in a time series. As a result, the ability to make informed forecasts or classifications based on sequential input data is improved. In this paper, Bi-LSTM networks are used because of their flexibility in handling sequences of variable lengths they can adapt to input sequences of diverse lengths by regulating their internal memory cells consequently. Furthermore, the hidden layers of a Bi-LSTM network serve as feature extractors as relevant features can be learned automatically from the input sequences. Particularly, this is helpful when dealing with complex patterns in sequential data.

Even though, LSTM networks are proficient in learning long-term dependencies in sequential data, which makes them well appropriate for speech recognition. Nevertheless, training enormous amounts of data with large and complex sequences is not correctly recognized by a simple LSTM network. Hence, in this paper, we propose to utilize a deep Bi-LSTM network to learn and recognize long-term sequences in acoustic data for recognizing emotions. In the deep Bi-LSTM network, the forward and backward pass entails cells, that make the network deeper to calculate the output from the preceding and subsequent sequence regard to time as the network performed in both directions.

This study introduces an innovative methodology for sentiment analysis utilizing a deep Bi-LSTM network. Predominantly, the initial step of the proposed model involves the preprocessing database of the speech samples through the utilization of a 2-D convolutional layer. After the initial stage, a sequence of discriminative multi-features at a high level is retrieved successively. These features include Mel-Spectrogram, MFCC, Contrast, Chroma and Tonnetz. Additionally, both mean and standard deviation technique is utilized on behalf of the purpose of normalizing the features. These normalized features are subsequently fed into a deep Bi-LSTM network to extract temporal information and accurately determine the final state. The Softmax classifier is utilized to generate a probability distribution for speech emotions. The evaluation of the suggested trained model was conducted by the confusion matrix as a means to assess the efficacy of the model.

2 Literature Review

The complex convolutional block attention module (CCBAM) was proposed by Zhao et al. [4] to enhance the capability of representing complex-valued convolutional layers by generating more informative features. The CCBAM was a compact and adaptable component capable of integration within diverse convolutional layers using complex-valued inputs. The researchers enhance their speech augmentation capabilities by integrating CCBAM using the deep complex U-Net and convolutional recurrent network (CRN). Additionally, a mixed loss function was proposed to optimize complicated models in both the temporal-frequency (TF) and time domains. A novel end-to-end (E2E) speech augmentation framework was developed through the integration of CCBAM with the mixed loss technique. The proposed techniques have exhibited higher performance in ablation testing and objective evaluations. The deep complex U-Net DCUnet-MC exhibits an average performance of 97.10 for Short-Time Objective Intelligibility (STOI) metrics. However, this research lacks an examination of the computational complexity, limitations, and applicability of the proposed CCBAM in real-time applications, monaural voice augmentation technologies, and noise and speech sources. The suggested complex speech improvement framework's applicability to varied noise and speech sources is not fully examined.

For a multichannel speech improvement challenge, Lee et al. [5] developed convolutional time-domain audio separation network (Conv-TasNet). They compared it to the state-of-the art (SOTA) models tested on the same CHiME-3 dataset and found that the suggested model performed more

favorably than the comparative models. In addition, a model with minimal input and output tensor sizes was presented to save memory. The size of the initial temporal convolutional network (TCN) stack was significant for the voice enhancement task; consequently, the downsized model that progressively decreased the output tensor size with increasing stack index performed similarly to the full-sized Conv-TasNet. The CHiME-3 dataset is frequently used in the field; however, it may not reflect all real-world events, restricting generalizability.

Kumar et al. [6] created an intelligent assistant that can recognize emotions in voice communications. They utilized RAVDESS, Texas Instruments Massachusetts Institute of Technology (TIMIT), and Emo-DB. Compared to Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) models, the suggested model is 81.02%, 84.23%, and 85.12% accurate for RAVDEES, TIMIT, and Emo-DB datasets. The study should investigate several NN designs and parameters for voice emotion recognition to increase system accuracy and efficiency.

In a comprehensive analysis conducted by Michelsanti et al. [7], a thorough evaluation of the existing literature was undertaken, with a specific focus on identifying the fundamental attributes that differentiate various systems in this field. These attributes encompass fusion approaches, deep learning techniques, visual and auditory features, training objectives, and objective functions. The researchers also examined deep-learning-based methodologies for voice reconstruction from silent films and non-speech audio-visual sound source separation, as both techniques have the potential to enhance and disentangle audio-visual speech. In light of their crucial contribution to the development of data-driven procedures and evaluation methodologies, the researchers conducted a comprehensive analysis of frequently employed audio-visual speech datasets. These datasets were commonly utilized to assess and ascertain the effectiveness of various systems. There is no extensive examination of deep learning-based audio-visual speech augmentation and separation systems in the paper. It does not compare audio-visual speech datasets and assessment techniques from the literature.

Ye et al. [8] proposed a novel approach for modeling temporal emotions in SER known as the Temporal-aware Bi-direction Multi-scale Network (TIM-Net). The objective of this strategy was to obtain contextual emotional representations from different time scales by employing a multi-scale learning process. The methodology included temporal-aware blocks to construct temporal affective representations. Subsequently, the complementing past and future information was integrated to enhance the contextual representations. Finally, the integration of characteristics from multiple time scales was performed to enhance the model's ability to respond to emotional variations. The researchers conducted experiments on six benchmark sentiment analysis datasets and observed that their proposed method resulted in an average increase of 2.34% in Unweighted Average Recall (UAR) and 2.61% in Weighted Average Recall (WAR) compared to the second-best performing method on each dataset. However, this paper explores how adding contextual information like facial expressions or physiological signals to TIM-Net may improve speech emotion identification and testing the TIM-Net technique for real-time speech emotion identification.

Chen et al. [9] presented SpeechFormer++, a generic framework based on structure for paralinguistic voice processing. A unit encoder was developed to effectively represent both intra- and inter-unit information. The researchers employed merging blocks as a technique for constructing features at different levels of detail, which aligns with the structural pattern observed in the voice stream. Additionally, this study introduces a word encoder that effectively integrates both fine-grained and coarse-grained information into each unit encoder, thereby achieving a harmonious equilibrium between these two types of information. SpeechFormer++ was assessed on tasks such as IEMOCAP and Multimodal EmotionLines (MELD), depression categorization (DAIC-WOZ), and Alzheimer's

disease detection. The results reveal that SpeechFormer++ is the regular Transformer though consuming significantly less processing power. Moreover, it demonstrates superior performance compared to SOTA techniques in terms of outcomes. The lexical information was not employed in the creation of a unified textual-audio framework. They also did not take into account semantic information in SpeechFormer++ when handling speech recognition tasks. However, SpeechFormer++'s computational efficiency had not been compared with other models for paralinguistic speech processing tasks other than emotion identification, depression categorization, and Alzheimer's disease detection.

Wang et al. [10] proposed a multi-feature fusion and multi-lingual fusion SER algorithm based on RNN with an improved local attention mechanism, they used traditional hand-crafted features extracted from GeMAPS and deep automatic features based on the VGGish model. The classification accuracy is improved while the dataset is slight.

In [11], a simple architecture of residual blocks in combination with one-dimensional Dilated CNN (DCNN) finds a correlation between emotional cues and the sequence of learning. They were achieved with IEMOCAP and EMO-DB datasets and a recognition rate of 73% and 90%, respectively.

In [12], an approach based on LSTM network models as a classifier using the MFCC features was introduced. This approach achieved 89% accuracy based on the RAVDEES dataset and only 4 emotion categories were employed (happy, neutral, sad, and angry), some researchers combined various feature types and used them as input for classification models.

Peng et al. [13] combined CNN using sliding RNNs with attention-based auditory heads to limit the auditory system of the human which can easily identify the speaker's intentions by inspection of any utterance intensity and frequency. The researchers noted through several experiments that their model can effectively use temporal modulation cues to identify the state, an accuracy rate of 62.6% for IEMOCAP and 55.7% for the MSP-IMPROV datasets is recorded.

Venkata Subbarao et al. [14] proposed an SER system based EMO-DB database, in the first phase 39 MFCCs were extracted than in the second phase, different types of K-nearest neighbor (KNN) classifiers (Fine, Medium, Cosine, Cubic, and Weighted) were used to classify different emotion categories. The training data were divided into two categories: 90% for training and 10% for testing. The simulation results achieved an overall accuracy of 90.1% which showed the effectiveness of the KNN classifier of type Fine with the Euclidean distance function.

Sowmya et al. [15] implemented a SER system that can recognize four different types of feelings from the RAVDESS dataset. To characterize the feelings, MFCC, chroma, tonnetz, mel, and contrast were extracted then SVM, Multilayer Perception (MLP), Random Forest (RF), and Decision Tree (DT) calculations were utilized for classification. Additionally, when compared to other methods by splitting the dataset into 75% for training and 25% for testing, MLP produced the best results with an accuracy of 85%.

Sajjad et al. [16] proposed a new method for choosing a more effective communication sequence using Radial Basis Function (RBF) based K-mean clustering algorithm and converting it to spectrograms using the Short Time Fourier Transform (STFT) algorithm. They retrieved discriminative and salient characteristics from speech signal spectrograms using the FC-1000 CNN model Resnet layers, then normalized it using mean and standard deviation to reduce variation. Afterward, they input these discriminative traits into a deep Bi-LSTM to learn the concealed information, identify the conclusion of the sequence, and categorize the speakers' emotional states. They tested the proposed system's robustness using three standard datasets: RAVDESS, EMO-DB, and IEMOCAP. They increased the recognition accuracy for the IEMOCAP dataset to 72.25%. Their work emphasizes handcrafted

characteristics and standard CNN models, which may restrict innovative speech emotion identification methods. The suggested method processes critical portions instead of the complete speech, which may lose contextual information. Neutral and happy emotions have low recognition accuracy in speaker-independent evaluation on the IEMOCAP dataset.

Koduru et al. [17] improved feature extraction by using Discrete Wavelet Transform (DWT), pitch, Energy, and Zero CrossingRate (ZCR) to extract the most information from the voice signal and achieve higher accuracy and speech recognition rate. They proposed five distinct feature extraction strategies, as well as a comparison of several classifiers. The simulation results show the accuracy and efficiency of several classifiers, with SVM being 70% accurate, the DTs being 85% accurate, and LDA being 65% accurate. The suggested system outperformed previous work in terms of accuracy, meaning that it pulls the most information from the signal required to describe the properties of the signal and to distinguish emotions from the signal efficiently. In addition, the suggested mechanism takes less processing time than the present one. As a result, the suggested system applies to all types of signals, provides a higher rate of speech recognition, and increases the system's accuracy and efficiency.

Chen et al. [18] addressed single-feature limits using a system that uses various types, layers, and scales of emotion features. A parallel training method feeds separate layer-level characteristics into multiple network designs. A CNN with multi-scale feature representation captures spatio-temporal aspects of multiple frame-level data using an attention mechanism. Using multi-head attention to combine deep representations from different levels while keeping feature-type characteristics. Segment-level speech emotion multiplex decision approach to improve classification robustness and reduce local interference signals. The suggested STRL-SER method Gaining significant performance increases over earlier research, with a weighted accuracy (WA) of 81.60% and an unweighted accuracy (UA) of 79.32% on the IEMOCAP dataset and 88.88% and 87.85% on the RAVDESS dataset. The method should be tested on bigger and more diverse datasets to confirm its efficacy and generalizability. Testing the suggested method in intelligent healthcare, customer service, contact centers, automatic translation systems, and human-computer interaction.

Zhou et al. [19] presented two new attention-based Bi-LSTM architectures for personality detection using emoji and textual input at distinct semantic levels. To create user document representations, the authors extract emoji information from online user-generated material and mix it with word and phrase embedding. Emoji information is valuable in personality recognition tasks, evidenced by the suggested approaches' state-of-the-art performance above baseline models on a real dataset. The work reveals the rich semantics of emoji information and suggests a novel technique to include it in personality detection tests. From the researches above, it can be seen that SER system results depend heavily on the utilized dataset, process of the feature extraction, and the ML method. So, these systems do not perform as well when applied to various datasets. Additionally, those intensive studies of the SER problem presented a huge collection of latent features of speech emotion. However, large feature sets confuse the SER task as the number of analyzed speech patterns becomes lower than the feature order. Therefore, the results classification becomes untrustworthy and meaningless. To overcome this problem, in this paper, we introduced sequential feature selection for reducing feature sets by keeping only dominant features by excluding the insignificant ones.

3 Methodology

This section explains the intricate design of the suggested framework, which was built expressly to identify emotions in speech. The proposed framework proposes to identify emotions based on

EMO-DB by a sequential selection of multi-feature extraction and Bi-LSTM. The diagrammatic representation of the proposed framework is depicted in Fig. 1.

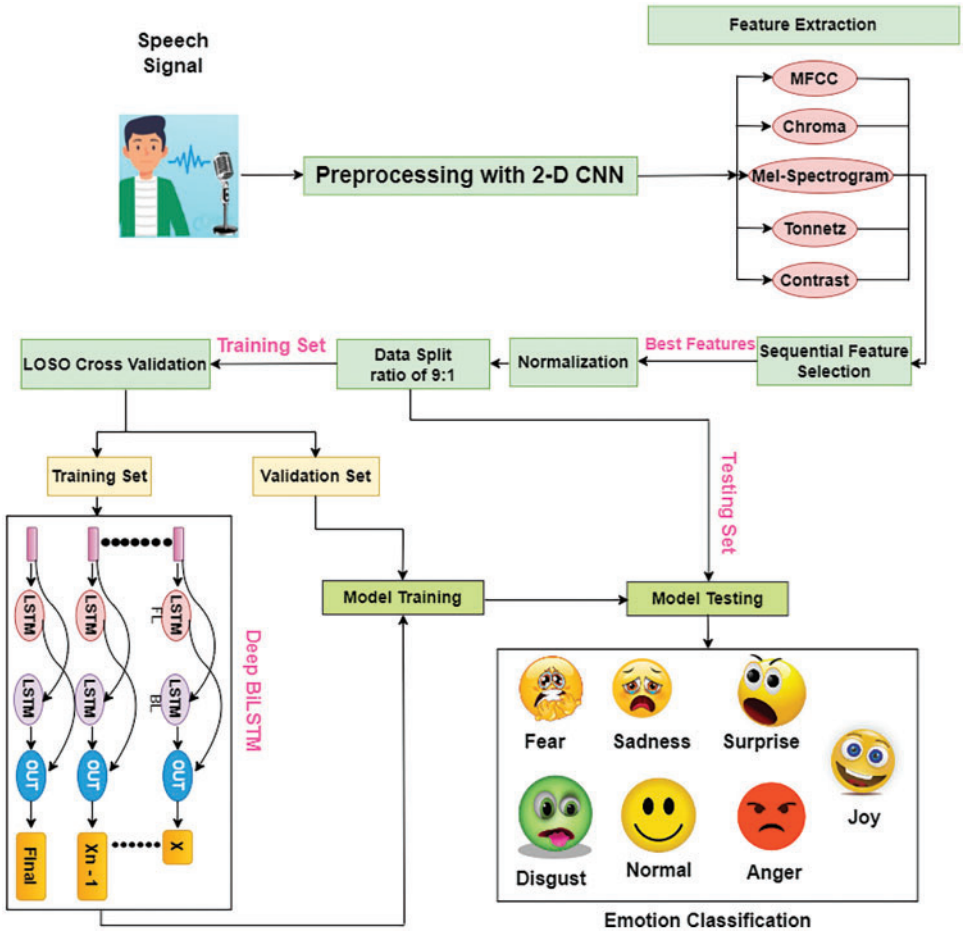


Figure 1: Overall structure of the proposed framework

From the above figure, it can be seen that the pre-processing is performed on the input speech signal. Then, the EMO-DB dataset will employ several audio extraction algorithms to retrieve its audio features. Next, the extracted features undergo sequential feature selection to assess the influence of data reduction on the resultant model.

Following the completion of the pre-processing stage, the Bi-LSTM model will utilize the training set as the reference data for training purposes. Once the training results model has been acquired, the subsequent phase involves conducting testing using test data. The suggested framework employs leave-one-speaker-out (LOSO) cross-validation for training and validation.

The emotion prediction results will be collected through the testing procedure using the trained Bi-LSTM model. These results will then be converted into a confusion matrix table, which will serve as a reference for evaluating the performance of the proposed model. The aforementioned parameters are employed to assess the precision of the suggested model. The succeeding sections provide a comprehensive analysis of each block inside the framework, with a detailed description of each.

3.1 Pre-Processing

The input and preparation stages include collecting audio data and removing any noise or other undesired signals.

Initially, the audio signal undergoes a conversion to the frequency domain with the use of the STFT. This transformation employs a window length of 256 samples, a 75% overlap, and a Hamming window. To achieve a reduction in the size of the spectral vector, it is advisable to decrease it to a dimension of 129 by removing the frequency samples associated with negative frequencies. The rationale for this modification is supported by the observation that the speech stream in the time domain is current, hence guaranteeing the preservation of information integrity.

The predictor and target network signals are representative of the magnitude spectra of the noisy and clean audio signals, respectively. The input to the predictor consists of a series of eight consecutive STFT vectors that have been corrupted by noise. The calculation of each output estimate in STFT entails the utilization of the current STFT vector, which contains noise, in addition to the preceding seven STFT vectors that also contain noise.

The de-noising network incorporated a 2-D convolutional layer which employed sliding filters to do the processing of the input. The convolutional layer performs a convolution operation on the input data through the systematic displacement of filters both for the vertical and horizontal direction across the input. The process involves the dot product calculation between the filter weights and the associated input values, and subsequently incorporating a bias term.

In practical applications, it has been observed that convolutional layers have a reduced parameter count in comparison to fully connected layers. The FCNs consists of a cumulative sum of 16 convolutional layers. The initial set of 15 convolutional layers is comprised of clusters of three layers, which are repeated five times. The convolutional layers in question have filter widths of 9, 5, and 9, and correspondingly employ 18, 30, and 8 filters. The last convolutional layer is comprised of a single filter with a width of 129.

Convolution operations in this network are solely executed in a unidirectional manner, specifically inside the frequency dimension. Furthermore, the filter width in the temporal dimension remains consistent at a value of 1 across all levels, except for the initial layer. In the architectural design of CNNs, the sequential arrangement of Rectified Linear Unit (ReLU) and batch normalization layers is typically observed after the convolutional layers.

The output of the network is represented as the magnitude spectrum of the de-noised signal. Consecutively, the predictor input is used by the regression network to minimize the mean square error between the output and target input. The conversion of the de-noised audio signal from the frequency domain to the time domain entails the use of both the output magnitude spectrum and the noisy signal phase [20]. The basic deep learning training scheme for denoising audio signal is shown in Fig. 2.

3.2 Dataset

Mainly, in this paper, EMO-DB which includes Berlin Emotional Speech Database recordings is used for training and testing [21]. The EMO-DB has 535 utterances said by ten actors (5 females, 5 males), each utterance expresses anxiety/fear, contempt, boredom, anger, happiness, neutrality, or sadness. The EMO-DB dataset has various advantages, including carefully controlled recording conditions that provide high-quality recordings and decrease speech signal acoustic variation. The

dataset includes a variety of emotions and numerous speakers, enabling the construction of more emotion recognition models which enable better robustness and generalizability.

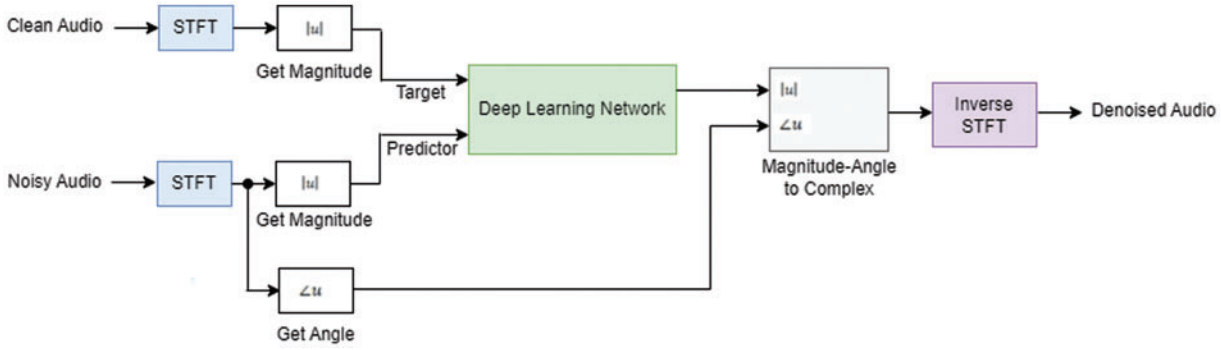


Figure 2: Speech de-noising using deep learning network

3.3 Feature Extraction

Human speech is a multifaceted and dynamic signal which carries a prosperity of emotional information. Numerous acoustic features such as pitch, rhythm, intensity, and spectral content can be revealing of distinct emotions. Consequently, feature extraction and analysis are crucial for recognizing emotions accurately as they enable the system to express the nuanced and multifaceted nature of human emotions uttered in speech. Significantly, auditory provides valuable contextual information, such as intonation, tone, and prosody that can contribute to the emotional content understanding. These contextual cues are vital for distinguishing similar acoustic patterns that are associated with diverse emotions.

Meanwhile, relevant information extraction from acoustic inputs allows models to consider a wider set of features, it improves models' capability to recognize complex emotional states so it is significant for SER system success. The speech signal includes parameters that define speech. However, speech features' dynamic nature and temporal modifications make understanding difficult [22].

In theory, speech recognition can be achieved directly from the signal. However, due to voice signal variation, feature extraction techniques are beneficial to mitigate its effects [23]. The recognition process relies on feature extraction, which transforms or combines the original feature set to create new features.

In speech recognition, the most widely used features are mfccDeltaDelta, spectralCentroid, spectralFlux, pitch, and harmonicRatio, additional features and audio features extractor pipeline are observed in Fig. 3.

Deep learning approaches are commonly employed by researchers for SER, with the Mel-scale filter bank voice spectrogram being utilized as the input feature. To enhance the ability to identify human emotions, it is advisable to take into account multiple acoustic characteristics inside our model. The proposed framework employs a comprehensive extraction approach that integrates many audio features, including MFCC, Chroma, Mel-Spectrogram, Tonnetz, and Contrast. The amalgamation in question is designed to effectively decrease computing complexity, while simultaneously preserving energy and bandwidth resources.

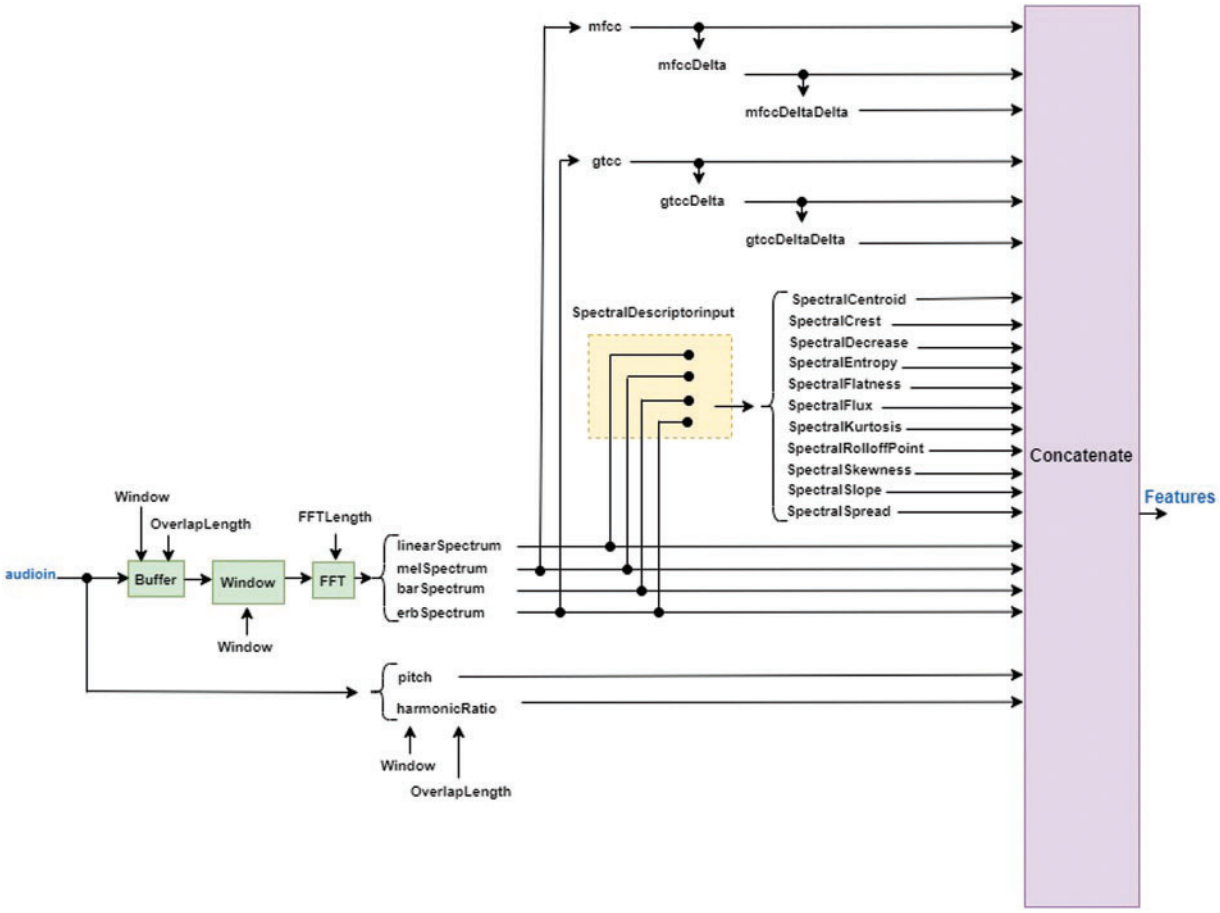


Figure 3: Audio feature extraction pipeline

The MFCC has become the predominant feature in the field of voice recognition [24]. The primary purpose of employing MFCC is to replicate the audio characteristics of the human ear. According to previous research [25], the MFCC has been suggested as a reliable method for identifying monosyllabic audio, eliminating the requirement for speaker identification.

For several decades in speech processing, MFCCs have been used as a benchmark feature representation, owing to their ability to capture human auditory system characteristics and their effectiveness for discriminative feature extraction from audio signals. As the human auditory system does not distinguish all frequencies equally, the Mel scale is a perceptual scale that approximates how humans perceive sound pitch. It is nonlinear and emphasizes lower frequencies, which are more applicable for processing speech. MFCCs are derived from the Mel-scale filter banks which imitate the human ear's response to distinct frequencies. MFCCs afford a compact representation of speech signal spectral characteristics. So, via a subset of coefficients selection, feature space dimensionality is reduced, enhancing computing efficiency and mitigating potential overfitting in ML models. MFCCs are relatively robust to variations in speakers and environmental conditions. Over and above, MFCCs are relatively robust to the diversity of speakers and environmental conditions.

Even though MFCCs are appreciated for capturing confident aspects of speech, there are some limitations to using them. MFCCs are disposed to the background noise impact, which can disguise

or distort emotional cues in speech. So, pre-processing, such as noise reduction should be applied to decrease noise influence.

The process for extracting audio MFCC features, as described in reference [26], consists of multiple stages. The process commences with the pre-emphasis stage, wherein the audio stream is amplified at higher frequencies. Following that, the processes of frame and windowing are executed. The initial phase of audio processing involves dividing the audio signal into distinct temporal segments, often spanning from 20 to 30 milliseconds. The application of the windowing approach serves to alleviate disruptions occurring at the initial and final segments of the audio. The windowed results are transformed into MFCC using the Fast Fourier Transform, Discrete Cosine Transform, and Mel Filter Bank, under the aforementioned phases.

The MFCC, utilized as a feature in the field of SER, provides the benefit of effectively capturing the acoustic characteristics of human speech. The utilization of the mel scale is advantageous due to its near alignment with the frequency perception capabilities of the human auditory system. The utilization of the logarithm of the power spectrum and its subsequent conversion to cepstrum in the process of MFCC serves to effectively decrease the dimensionality of features and simplify the computational complexity involved. Moreover, the utilization of short-duration frame segmentation techniques in MFCC allows for the effective capture of temporal information in speech signals. This enables the capturing of fluctuations related to temporal emotional shifts.

However, MFCCs do not explicitly capture prosodic features such as pitch and intonation which play an essential role in conveying emotions over speech, which can limit their capability to characterize certain emotional characteristics [27].

To address this limitation, prosodic features are integrated with MFCCs to provide a more inclusive representation of emotional cues. In this paper, multiple feature sets that capture different aspects of speech are combined to improve the emotion recognition system's overall performance. In this study, Chroma, Contrast, Mel-Spectrogram, and Tonnetz will be used with MFCCs as emotion recognition features.

Chroma, conversely, is a method of extracting features that primarily concentrate on auditory tones related to music [28]. The representation of tonal differences in audio as a basic characteristic is facilitated by this technology. The outcome of implementing the Chroma feature extraction technique yields a Chroma gram that is constructed based on a scale consisting of 12 distinct tone levels [29]. The application of Chroma is intended to identify and differentiate between high and low-pitch attributes in voice audio, as pitch might offer indications of particular emotional states.

Mel-Spectrogram, another audio feature extraction method, addresses the limitations of human auditory perception in discerning high-frequency values [29]. In this study, the Mel-Spectrogram is employed to extract information about frequency variations, particularly in the identification of different emotional states. Tonnetz, derived from Chroma, is an additional feature extraction technique that centers on acoustic harmony and classes of tone [30]. Contrastingly, a feature extraction approach in audio proves valuable in estimating sound energy averages by considering each sub-bands peak and values of the valley spectrum [31].

3.4 Sequential Feature Selection for Audio Features

After multiple features are extracted from the audio signals, the selection of features is systematically done based on the sequential feature selection technique. The approach employed for feature selection in our framework is sequential feature selection.

In this study, we used this technique to improve the performance and efficiency of the proposed learning model by selecting a subset of relevant features successively.

This technique entails training the network using a specific feature set and iteratively adding or removing features until the highest level of accuracy is attained [32]. This process aims to optimize the accuracy of the network by systematically adjusting the set of features utilized.

Selection starts forwarded with an empty set of features then the most relevant feature iteratively added until reaching the optimal accuracy or until further improvements in accuracy are no longer observed this is called SFS. Sequentially, training starts with all features and iteratively removes the least relevant features until no enhancement in accuracy is met this is known as SBS. Finally, a selected subset of features through SFS or SBS turns into an input set for the Deep Bi-LSTM model to learn the temporal information for the final state of emotion identification [32].

3.4.1 Forward Selection

A basic instance of forward selection involves a group of four features. During the initial forward selection iteration, each of the four features is individually tested by training a network and comparing their respective validation accuracies. The feature that yields the highest validation accuracy is recorded. In the subsequent forward selection iteration, the best feature from the first iteration is combined with each of the remaining features. This results in pairs of features that are used for training. If the accuracy in the second iteration does not surpass the accuracy achieved in the first iteration, the selection process concludes. However, if the accuracy does improve, a new set of features that perform the best is chosen. The forward selection loop continues until the accuracy no longer demonstrates any further improvement. In Fig. 4, the phase of features forward selection is demonstrated.

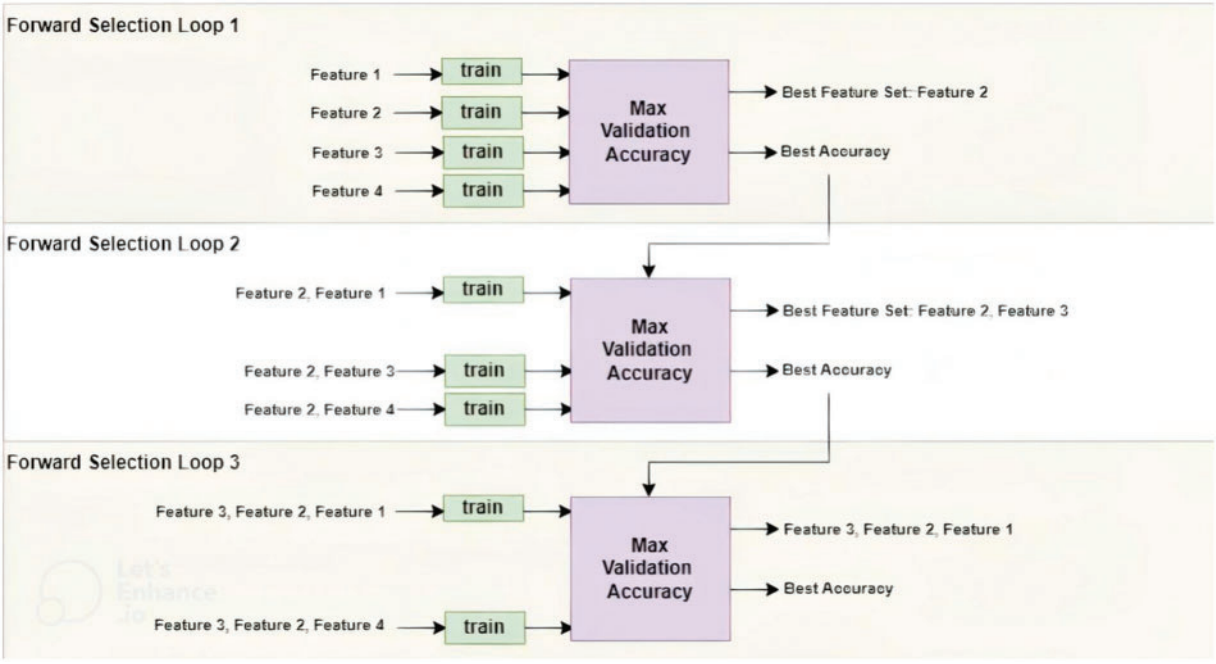


Figure 4: Forward selection for the features

3.4.2 Backward Selection

Backward selection for the features involves an initial training phase using a feature set that encompasses all available features. The subsequent step involves systematically removing individual features and assessing whether the accuracy of the model improves, Fig. 5 exhibits bachward selection phase.

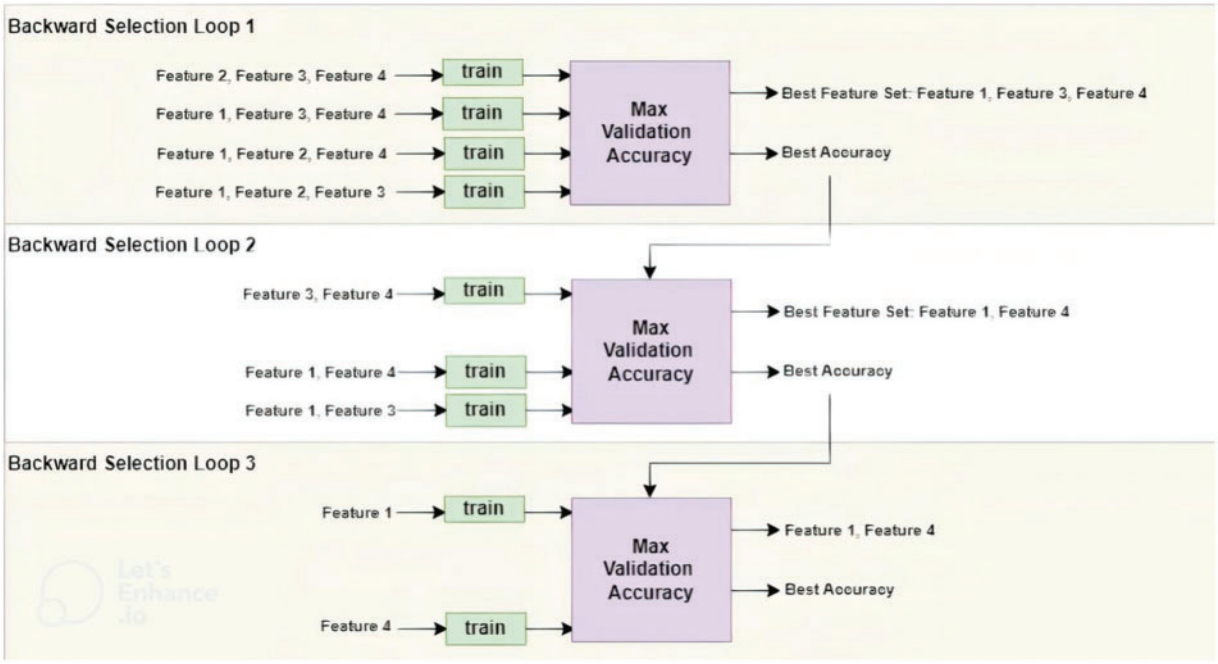


Figure 5: Backward selection for features

3.5 Deep Bi-LSTM Classifier

Practically, several standard classifiers are utilized for recognizing the emotion of humans from their speech, including the SVM, KNN and Kernel Deterioration, Maximum Likelihood Bayes Classifier (MLBC), Hidden Markov Model (HMM), Naive Bayes Classifier [33], GMM and NN [34]. Among these, the LSTM classifier has gained popularity and finds wide applications in numerous fields, like speech recognition, machine translation, and language modeling. LSTM is an RNN feature that offers effective solutions for a broad spectrum of problems. It is commonly employed to mitigate the issue of vanishing gradients, although it does not eliminate this problem. However, LSTMs can become inefficient, and demand increased bandwidth for training data.

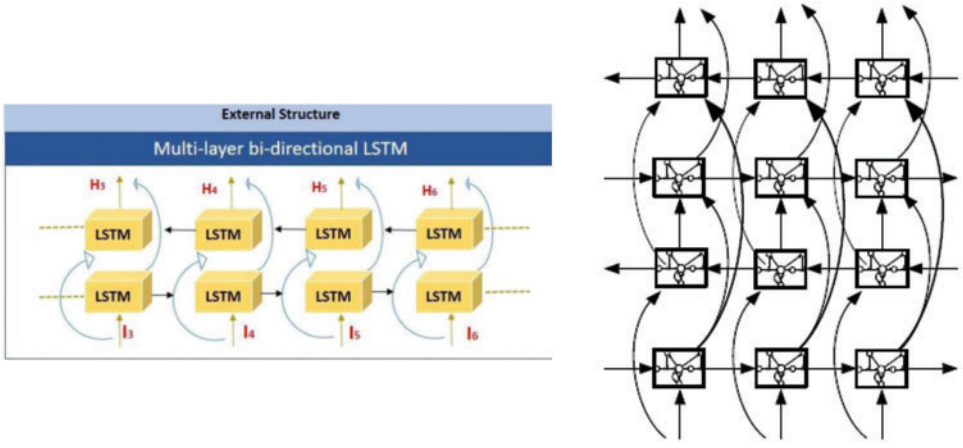
The utilization of a variant of LSTM known as Bi-LSTM is employed to incorporate an improved functionality. In practical terms, this approach involves the integration of two LSTM models to effectively represent bidirectional long-term relationships within sequential data or across distinct time steps. The aforementioned dependencies demonstrate their significance when RNN is assigned the duty of acquiring knowledge from the complete time series at every time step.

The Bi-LSTM networks involve training each signal in both the forward and backward directions, which allows for the disentanglement of recurrent networks. Both sequences are connected to the

identical output layer. These models are capable of capturing and incorporating extensive contextual information for each element within a sequence, including both preceding and subsequent elements.

The input sequence is initially subjected to processing by a forward LSTM layer. Subsequently, the input sequence is provided to the LSTM network in reverse form through a backward layer. The utilization of LSTM twice enhances the acquisition of long-term dependencies, hence leading to an enhanced accuracy of the model. In the concealed layers, data undergoes bidirectional processing, where bidirectional LSTM models gather relevant information from both preceding and succeeding directions. This facilitates the retrieval of extensive contextual information. The aforementioned information is subsequently directed toward the identical output layer in Bi-LSTM models. However, the output of the Bi-LSTM model at time "t" is not solely determined by a single segment of the sequence, but rather impacted by both the preceding and subsequent segments [35]. Bidirectional RNNs are composed of a dual set of RNNs that are combined in a stacked configuration. In the proposed model, the sequence of the input is processed through two RNNs, with one RNN operating in the forward direction and the other RNN operating in the backward direction. The combined output is computed by these two RNNs using their respective hidden states. As training a huge amount of data with large and complex sequences is not correctly recognized by a simple LSTM network, in this study, we have utilized a deep Bi-LSTM network to learn temporal signals and recognize the sequential information in a sequence and analyze the emotional state of the speaker in speech cues. Deep architectures denote the stacking of multiple layers of Bi-LSTM units which allow the model to learn the input data hierarchical representations, as each layer catches different abstraction levels.

The basic structure block of a Bi-LSTM is the LSTM cell, the LSTM presents memory cells and gates (input, forget, output) to adjust the flow of information. The LSTM equations include operations for input, forget, and output gates, in addition to updating the memory cell are explained in detail in [36,37]. This research utilizes a multi-layer deep Bi-LSTM model to extract information and detect prolonged temporal patterns in audio data, with a specific focus on the task of recognizing emotions. The forward and backward passes utilized a two-layer LSTM which makes the model a deep Bi-LSTM. The overarching architecture of the multi-layer Bi-LSTM and deep BILST is depicted in Fig. 6.



(a) (b)

Figure 6: Deep Bi-LSTM network (a) External architecture (b) Internal architecture

The above visual representation in Fig. 6a showcases the external structure, specifically highlighting the training stage of the bidirectional RNN. The architectural design incorporates the hidden states

derived from both the forward and backward passes, which are subsequently combined in the output layer. Afterward the output layer, both cost and validation metrics are calculated. Subsequently, the weights and biases are adjusted through the process of backpropagation. Stacking the Bi-LSTM by joining the output of the lower layer to the higher layer input provides the internal structure of deep Bi-LSTM as illustrated in Fig. 6b.

In the training phase, the model is fed with input sequences through the first layer of the Bi-LSTM, creating hidden states for each time step. From the previous forward pass, hidden states are concatenated with the backward pass hidden states to form the final hidden states for each time step. Sequentially backpropagation through time is applied to update network weights to reduce the difference between predicted and true output. Commonly, a loss function, like categorical cross-entropy, is used to enumerate this difference. Practically, deep Bi-LSTM network training involves various hyper-parameters adjusting, like rates of learning, dropout rates, and LSTM layer numbers. To update the parameters of the model to minimize the loss, an optimization algorithm like Adam, is used. Especially in deep architecture, regularization techniques, like dropout, can be applied to prevent overfitting. As we trained the proposed model on different datasets, the model should be fit to training data by adjusting epochs number and batch size based on the used dataset size. During the testing phase, the trained deep Bi-LSTM is utilized for sequence prediction. for each time step. Then, the evaluation of the trained model on the testing dataset is accomplished to assess its generalization performance. The proposed model's predictions are compared against the true output to calculate metrics like accuracy, precision, recall, and F1 score. Experimentally, we found that the deep architecture enhances the model's capacity to capture complex patterns and representations in the data. The deep Bi-LSTM network utilizes cells for both the forward and backward propagation, allowing the network to generate output by considering preceding and subsequent sequences in a temporal context, as the network functions bi-directionally. Algorithm 1 describes the complete training and testing algorithm based on Deep Bi-LSTM and multi-features.

Algorithm 1: Proposed speech emotion recognition based Deep Bi-LSTM and multi-features.

Input: Selected speech samples of EMO-DB dataset.

Output: Classification accuracy

1. Loading the selected speech samples of EMO-DB database $\boldsymbol{T}_{unpreprocessed}$.
2. For each Sample $\boldsymbol{T} \in \boldsymbol{T}_{unpreprocessed}$, Divide the samples into frames of 30 msec.
3. For each frame $\boldsymbol{f} \in$ sample $\boldsymbol{T}$, perform denoising and Append $\boldsymbol{f}$ to new pre-processed sample $\boldsymbol{T}_{preprocessed}$.
4. Split the set $\boldsymbol{T}_{preprocessed}$ into training $\boldsymbol{T}_{Train}$, validation $\boldsymbol{T}_{Valid}$, and testing $\boldsymbol{T}_{Test}$ datasets.
5. For each sample $\boldsymbol{T}_{Train}$, employ Librosa to extract the MFCC, Chroma, Spectral Contrast and Tonnetz features.
6. Sort the $\boldsymbol{T}_{Train}$ and $\boldsymbol{T}_{Valid}$ datasets by sequence length.
7. Build the Deep Bi-LSTM network.
8. Initialize the training hyper-parameters and identify the training options.
9. Train the Deep Bi-LSTM network with $\boldsymbol{T}_{Train}$ dataset.
10. Validate the Deep Bi-LSTM network with $\boldsymbol{T}_{Valid}$ dataset.
11. If Deep Bi-LSTM network is not optimized, then reinitialize the hyper-parameters from step 8.
12. Load the testing dataset $\boldsymbol{T}_{Test}$ and Classify the samples using a pre-trained Deep Bi-LSTM model.

(Continued)

Algorithm 1 (continued)

13. Match the similarity between the test and predicted labels.
14. Evaluate the accuracy of the model classification. If accuracy is optimal, then output classification accuracy
    else
    the model is re-build from step 7.

4 Experiment Result and Discussion

4.1 Perform Speech Emotion Recognition

Emotions encompass alterations in both physical and psychological sensations that have an impact on human behavior and cognition. They are directly related to temperament, personality, mood, motivation, and vitality.

In this paper, we introduced a deep Bi-LSTM-based SER system. The development of a Structural Equation Modelling (SEM) requires a substantial amount of annotated training data, a process that is both time-consuming and costly. To tackle this difficulty, a thorough evaluation was conducted using the publicly available EMO-DB dataset [21] to assess the proposed approach's productivity. The utilization of this dataset facilitated a comparative analysis between the suggested model and established methodologies, thereby substantiating its superior performance, as it is extensively employed in studies about speech emotion recognition.

The initial attempt at training yielded a 10-fold cross-validation accuracy of approximately 60% due to a lack of enough training data. The model, when trained on an inadequate amount of data, exhibits overfitting in certain folds and underfitting in others. The dataset's size can be expanded by employing audio augmentation to enhance the overall suitability. Various augmentation strategies can be systematically or stochastically employed through cascading or parallel augmentation procedures.

A range of audio augmentation methods, including pitch shifting, time shifting, time-scale alteration, noise addition, and loudness control, can be utilized to increase the size of an audio dataset. This study employs simple data augmentation techniques, including the addition of noise and pitch shifting, to enrich the data.

In the conducted experiment, a deliberate selection of 55 augmentations per file was made, taking into consideration the trade-off between processing time and the improvement in accuracy. First and foremost, the generation of noisy speech signals involves the multiplication of the maximum value of the speech signal, denoted as s, by the noise ratio. This operation results in the determination of the noise amplitude. In previous iterations, the amplitude of noise was integrated with the original signal to produce a speech signal containing additional noise. The utilization of the Fourier transform on the primary vocal signal is employed to achieve pitch shifting. Upon reflection, it can be observed that the signal in the frequency domain undergoes multiplication by a logarithmic phase factor, which possesses a numerical value of 0.7. To complete the process of pitch shifting, the signal is ultimately subjected to a Fourier inverse transform, which allows it to revert to the domain of the time.

The chance of pitch shifting is configured to 0.5, and the default range is employed. Next, establish the probability of using time shifting as 1 and utilize a time range spanning from -0.3 to 0.3 s. Ultimately, the probability of introducing extraneous signals is established at a value of 1, while the signal-to-noise ratio (SNR) is configured within the range of -20 to 40 decibels (dB). Based on empirical evidence, it is recommended to increase the dataset size to 29,425 instances, incorporating

55 augmentations, to improve the overall fitting performance. Table 1 presents the outcomes of the data augmentation performed on the EMO-DB database.

Table 1: The data quantity for each emotion in the EMO-DB database before and after augmentation

Emotion	Count	Data set after augmentation (55 augmentation)
Anger	127	6985
Anxiety/Fear	69	3795
Boredom	81	4455
Disgust	46	2530
Happiness	71	3905
Neutral	79	4345
Sadness	62	3410

Following a sequential process, the various characteristics were derived after the preprocessing of the emotional signals. In the proposed framework, multi-feature extraction is performed with the help of Librosa [38] which is a python package developed for the analysis of music and audio, it provides the building blocks needed to generate music information for retrieval systems.

In terms of the outcomes of the extraction process, a comprehensive set of 182 features was acquired. This set comprised 40 features based on the MFCC and 10 features depend on the Chroma method. Additionally, the Mel-Spectrogram, Contrast, and Tonnetz methods yielded a total of 120, 7, and 5 features, respectively. Sequential feature selection is then employed to select the features. Earliest, the network is trained on a specific feature set, and then features are added or removed incrementally until the highest accuracy is achieved. Foremost, a 30 ms Hamming window with a 20 ms overlap is employed.

The feature sequences are sequentially inputted into the network to perform mean prediction and estimate and predict the standard deviation. The characteristics are normalized using the mean and standard deviation to ensure accurate recognition performance.

Subsequently, the feature is transformed into sequences consisting of 20 elements, wherein each sequence overlaps with the preceding and succeeding sequences by 10 elements. This arrangement corresponds to windows of approximately 600 milliseconds in duration, with an overlap of 300 milliseconds. Ultimately, the feature vector array transforms into sequences.

Formerly, the feature vectors were organized into sequences consisting of 20 feature vectors apiece, by using 10 feature vectors overlaps. To ensure a one-to-one correspondence between the sequences and their respective labels in the training and validation sets, it is necessary to replicate the labels. To ensure that the categorization array includes all emotions, it is necessary to construct an empty array encompassing all emotional classes and afterward add it to the validation tags. This approach accounts for the possibility that not all speakers possess utterances for every emotion.

Finally, the recovered normalized features are inputted into a deep Bi-LSTM model to capture temporal patterns, identify the sequential information inside a sequence, and determine the ultimate speaker's emotional state in speech signals. The deep Bi-LSTM network employed a total of four bidirectional hidden layers, with respective LSTM cell sizes of 100, 100, 50, and 30. Furthermore, dropout layers were incorporated within the training process to mitigate the risk of overfitting. These

dropout layers were assigned probabilities of 0.5, 0.5, 0.3, and 0.3, respectively. Lastly, the Softmax function is employed as an activation function for the output layer to achieve the classification.

4.2 Implementation Environment

To assess the suggested model, the methodology recommended in this study was employed, employing various hardware and software configurations. Table 2 provides a thorough summary of the components assembled to carry out the proposed model. This public inventory of resources used ensures replicability and further research by providing insight into the computational backbone of our methods. We successfully executed the model and acquired the results by using the provided hardware and software.

Table 2: Implementation specifications

Model implementation	
RAM	16 GB
GPU	NVIDIA GeForce MX450
CPU	11th Gen Intel (R) Core (TM) i7-11370H @ 3.30 GHz
OS	Windows 10 Pro
Programming environment	Matlab
Cross-validation	10-fold

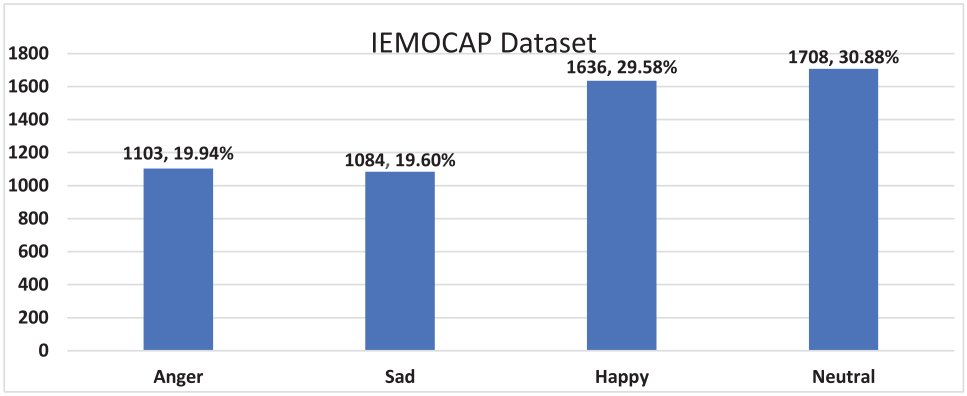
4.3 Model Optimization

During the training phase, the Adam optimizer was used to optimize the model, and the most effective bias correction method was picked to enhance performance. The experiments were conducted by employing both un-normalized and normalized features while considering various batch sizes and learning rates to determine the most optimal approach.

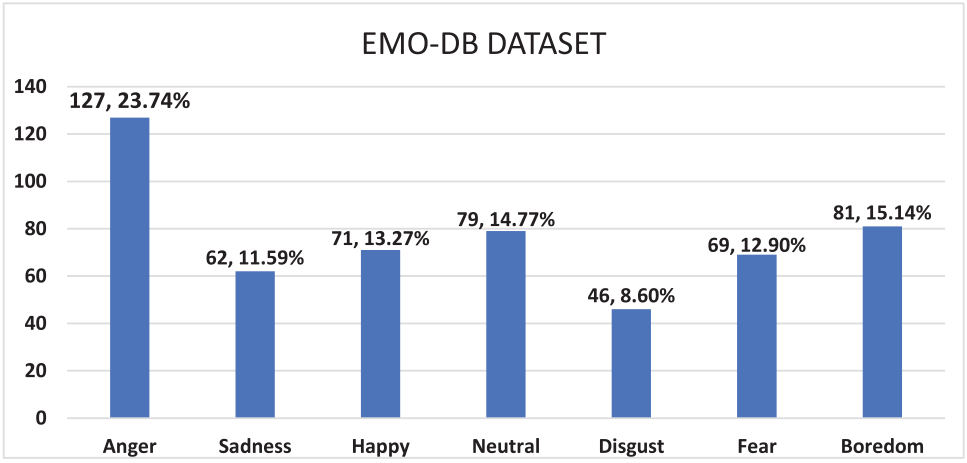
To be able to evaluate the reliability of the suggested framework, the Scikit-Learn package is utilized to partition 10% of the dataset to perform tests. The remaining 90% is allocated randomly to create training data and validation data, with a ratio of 9:1.

The model's empirical validation is contingent upon conducting complete experiments utilizing three distinct speech-emotional datasets, namely IEMOCAP, EMO-DB, and RAVDESS. The emotions distribution description in detail of different classes and participation of each class in percentage for the three used datasets is described in Fig. 7.

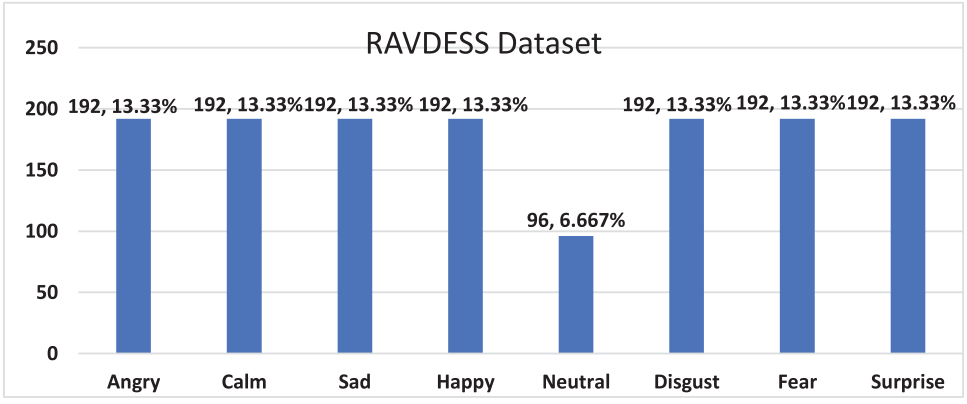
Tables 3 and 4 give a comprehensive depiction of various parameters and their accompanying outcomes as obtained from the proposed model. The tables presented in this study encompass both un-normalized and normalized features, providing a comprehensive examination of the data. The determination of the superlative learning rate and batch size for the entire model is conducted before running full tests on all datasets.



(a)



(b)



(c)

Figure 7: Emotion distribution for each dataset: (a) IEMOCAP (b) EMO-DB (c) RAVDESS

Table 3: Proposed model performance using un-normalized features on the IEMPOCAP dataset, RAVDESS dataset, and EMO-DB dataset

Dataset	Batch-size	Learning rate	Accuracy (%)
IEMPOCAP	256	0.01	75.22
		0.001	75.86
		0.0001	75.14
	512	0.01	77.34
		0.001	**77.97**
		0.0001	77.02
	1024	0.01	76.42
		0.001	76.88
		0.0001	76.45
RAVDESS	256	0.01	80.52
		0.001	80.65
		0.0001	80.23
	512	0.01	82.44
		0.001	**82.78**
		0.0001	82.05
	1024	0.01	81.36
		0.001	81.22
		0.0001	80.16
EMO-DB	256	0.01	88.24
		0.001	88.92
		0.0001	88.17
	512	0.01	90.22
		0.001	**91.65**
		0.0001	91.27
	1024	0.01	89.02
		0.001	89.97
		0.0001	89.41

Based on the findings presented in Tables 3 and 4, it is evident that the implementation of feature normalization leads to enhanced identification accuracy across different datasets. The utilization of feature normalization led to a 0.5% increase in recognition accuracy for the IEMOCAP dataset, a 0.3% increase for the EMO-DB dataset, and a 0.2% increase for the RAVDESS dataset, in comparison to using un-normalized features.

Table 4: Proposed model performance using normalized features on the IEMPOCAP dataset, RAVDESS dataset, and EMO-DB dataset

Dataset	Batch-size	Learning rate	Accuracy (%)
IEMPOCAP	256	0.01	82.43
		0.001	82.67
		0.0001	82.11
	512	0.01	77.34
		0.001	**77.97**
		0.0001	77.02
	1024	0.01	76.42
		0.001	76.88
		0.0001	76.45
RAVDESS	256	0.01	84.22
		0.001	84.55
		0.0001	84.13
	512	0.01	83.44
		0.001	**83.78**
		0.0001	83.05
	1024	0.01	82.32
		0.001	82.24
		0.0001	82.10
EMO-DB	256	0.01	90.61
		0.001	90.94
		0.0001	90.24
	512	0.01	92.77
		0.001	**92.51**
		0.0001	92.60
	1024	0.01	90.34
		0.001	90.55
		0.0001	90.10

As environmental issues can present variability that may not be interrelated to the emotional state of the speaker, feature normalization may aid in producing feature vectors that are more robust to variations in the environment.

5 Results

This section aims to assess the performance of the proposed model in speaker-dependent emotion recognition, to showcase the superiority of our proposed model compared to other existing models. In the majority of approaches employed in SER systems, the model is trained using an extensive amount of the available data, though reserving a substantial proportion for evaluation purposes. The dataset is divided into two subsets, namely the training set and the test set. The training set is allocated 80% of the samples, while the test set is allocated the remaining 20% of the samples. In the study referenced by [39], a total of 428 samples from the EMO-DB dataset were designated for training, whereas 107 samples were put aside for testing.

Recently, academics have made significant advancements in the development of different methodologies aimed at discerning the emotional state of individuals through the analysis of their verbal utterances. The classifiers known as GMM and HMM were able to obtain an accuracy rate of 77% in the presence of SNR of 40 dB, as reported in reference [40]. Based on the analysis of multiple methodologies, it is seen that the majority of procedures yield a performance ranging from 75% to 85% when applied to the EMO-DB dataset. The comparative analysis of the proposed model's performance, concerning other established approaches for SER that utilize spectral information for classification, is presented in Table 5. The performance analysis was conducted using the EMO-DB dataset, which was divided into two sets with an 80/20 split. Specifically, 80% of the data was utilized for training the model, while the remaining 20% was reserved for testing purposes.

Table 5: Comparison of SER models performance on EMO-DB

Models	Features	Classifier	Accuracy (%)
Basu et al. [41]	Acceleration,Velocity	CNN-LSTM	80
Ogawa et al. [35]	MFCCs	CNN-DNN	88.22
Pulatov et al. [39]	MFCC	Bi-LSTM	91.4
	GTCC		92.1
Proposed	Multi features with sequential selection	Deep Bi-LSTM	93.58

Table 5 presents the outcomes of the comparison between proficient SER assignment schemas, including the one provided in this research work. However, in the execution of sentence encoding and retrieval SER tasks, our system demonstrated superior performance compared to the chosen models, especially when incorporating semantic data. This accomplishment demonstrates the sophisticated expertise of our model and suggests it is probable to outperform existing methodologies in the domain of recognition of voice emotion.

To enhance the comprehensiveness of the study, additional experiments were conducted using natural emotional data from the IEMOCAP, EMO-DB, and RAVDESS datasets. The IEMOCAP and EMO-DB corpora consist of 10 actors, while the RAVDESS dataset comprises 24 players. The LOSO cross-validation method is employed to train and validate the suggested model. The k-fold cross approach involves doing training using $k-1$ speakers, followed by validation with the speaker that was not included in the training set. The aforementioned procedure is executed k times. Finally, validation accuracy is determined by calculating the average of the k-folds.

To assure the dependability of the proposed model, a random selection process is employed using the Scikit-Learn package. This process involves dividing the dataset into three subsets: a testing set including 90% of the data, a train set, and a validation set, each containing 5% of the remaining data. The division of the remaining data into the train and validation sets is completed randomly, with a ratio of 9:1. The comprehensive evaluation of the suggested model is demonstrated in terms of precision, recall, and F1 score at the class level for each emotion. Each statistic within the context of the model represents a unique facet of its efficacy. Tables 6–8 provide a comprehensive description of the numerical results for each dataset.

Table 6: The performance of the proposed model using the IEMOCAP dataset

Emotion	Precision	Recall	F1 score
Anger	0.89	0.94	0.91
Sadness	0.86	0.92	0.89
Happiness	0.99	0.87	0.93
Neutral	0.89	0.92	0.90

Table 7: The performance of the proposed model using the EMO-DB dataset

Emotion	Precision	Recall	F1 score
Anger	0.98	0.96	0.97
Anxiety/Fear	0.91	0.91	0.91
Boredom	0.94	0.96	0.95
Disgust	0.93	0.89	0.91
Happiness	0.96	0.90	0.93
Neutral	0.88	0.91	0.89
Sadness	0.92	0.97	0.94

Table 8: The performance of the proposed model using the RAVDESS dataset

Emotion	Precision	Recall	F1 score
Anger	0.90	0.98	0.94
Calm	0.91	0.94	0.92
Sadness	0.94	0.92	0.93
Happy	0.94	0.83	0.88
Neutral	0.90	0.83	0.86
Disgust	0.93	0.93	0.93
Fear	0.91	0.90	0.90
Surprise	0.90	0.94	0.92

The evaluation of the performance for the proposed model's classification was accomplished on three datasets: IEMOCAP, EMO-DB, and RAVDESS. The confusion matrix for emotion prediction for each dataset is presented in Figs. 8–10, respectively. The confusion matrix displays the observed and projected emotional outcomes for each category.

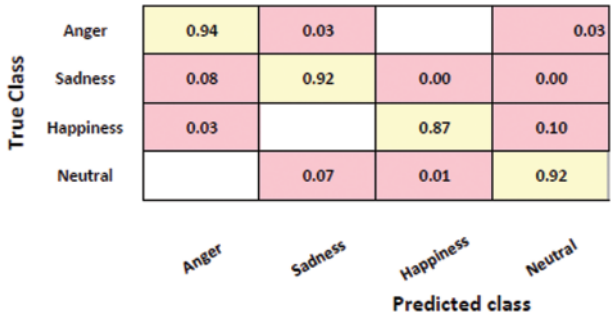


Figure 8: Confusion matrix of emotion prediction on IEMOCAP dataset with 90.92% accuracy

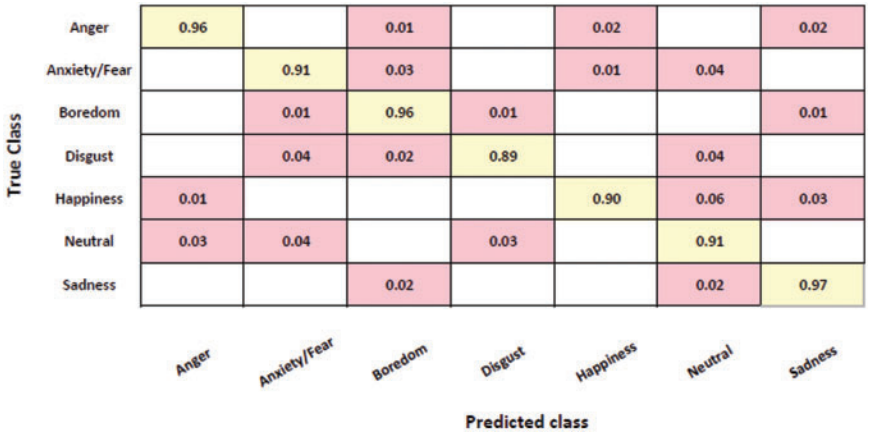


Figure 9: Confusion matrix of emotion prediction on EMO-DB dataset with 93% accuracy

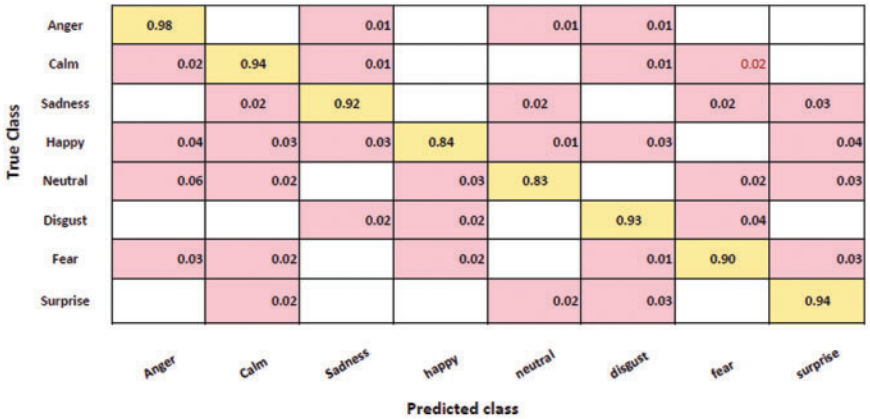


Figure 10: Confusion matrix of emotion prediction on RAVDESS dataset with 92% accuracy

The figure above displays the accuracy of recognition for the proposed model on the IEMOCAP dataset. In the conducted experiment, the acquired accuracy rates for the recognition of various emotions were as follows: 94% for anger, 92% for sadness, 87% for happiness, and 92% for neutrality. The experiment's ability to accurately identify emotions surpasses the current leading outcomes. Fig. 9 displays the outcomes of the EMO-DB dataset.

The suggested model demonstrated superior performance in experimental trials conducted on the EMO-DB dataset, with an average recall rate of 93% in accurately recognizing emotions. In this experimental study, the system demonstrated a high level of recognition for emotions such as anger, boredom, and melancholy, ranking them prominently. Additionally, the system achieved a recognition rate of over 90% for emotions such as fear, happiness, and neutrality. The emotion of disgust was recognized at a slightly lower rate of 89%. The system exhibited confusion in distinguishing between happy and neutral emotions, often misclassifying mostly happy feelings as neutral, much like a speaker-independent model. The suggested approach exhibits superior performance compared to previous baseline techniques in terms of effectiveness and significance. The performance and generalization of our model for SER are demonstrated by evaluating it on the REVDESS dataset. The results of this evaluation can be observed in Fig. 10.

According to the data presented in Fig. 10, the suggested model demonstrated a high level of accuracy in recognizing emotions. Specifically, the recognition rates for anger, calmness, sorrow, disgust, fear, and surprise were found to be 98%, 94%, 92%, 90%, 93%, and 94%, respectively. However, it is important to note that the rates of happy and neutral feelings were relatively low in comparison to other emotions, even though their recognition rates were superior to those achieved by state-of-the-art approaches.

Furthermore, based on Table 9, a comparative analysis of the proposed system with other baseline SER methods is held respectively on IEMOCAP, EMO-DB, and RAVDESS datasets.

Table 9: Comparison of the proposed method with baseline methods works on IEMOCAP, EMO-DB, and RAVDESS datasets

Dataset	Baseline method	Overall accuracy
IEMOCAP	[42]	66.5
	[43]	64.74
	[44]	69.32
	[16]	72.25
	Proposed method	**90.92**
EMO-DB	[16]	85.75
	[45]	76.74
	[46]	65.30
	[47]	69.72
	Proposed method	**93.0**
RAVDESS	[16]	77.02
	[45]	73.26
	[47]	65.67

(Continued)

Table 9 (continued)

Dataset	Baseline method	Overall accuracy
	[48]	56.5
	Proposed method	**92.0**

From the above table, it can be seen that the proposed method assessed on three standard datasets outperformed and demonstrated noteworthy results compared with previous studies which proved the proposed method's effectiveness and robustness.

Due to the use of sequential selection-based multi-features which keep only dominant features by excluding the irrelevant ones, the performance of the proposed model's accuracy in emotion type determination has been significantly improved. In addition, using salient and normalized features with a deep Bi-LSTM network also delivers an increase in pattern correlation between features so that the proposed model can demonstrate performance improvements, especially accuracy.

In summary, integrating SER using deep Bi-LSTM into the domains of mathematics and machine learning holds the potential to provide more emotionally intelligent systems, improve user experience, and offer valuable insights for both educational and research endeavors.

The proposed model can be integrated into mathematics teaching systems to acclimate the style of education based on the emotional state of the student. For example, provide supplementary assistance or alternate elucidations if it detects annoyance or bewilderment.

Furthermore, emotional information incorporation into ML algorithms can enhance their performance as understanding the emotional context can aid predictions or recommendations improvement.

6 Conclusion

The present CNN technology employed by the SER system is plagued by numerous difficulties, notably the lack of model accuracy. Consequently, we have proposed a methodology for SER that aims to improve emotion detection by extracting multi-discriminative features from voice signals using deep Bi-LSTM.

The initial training of the network involved utilizing a specified set of features. Subsequently, the network iteratively incorporated or removed individual features until a point was reached where further adjustments no longer increased accuracy. The utilization of the mean and standard deviation in the process of normalizing recovered characteristics efficiently mitigates the presence of variation. The discriminative features are fed into a deep Bi-LSTM model following the normalization step. The objective of this model is to ascertain the ultimate state of the sequence, extract the latent information, and categorize the speakers' emotional states. To assess the resilience of the suggested methodology, a series of experiments were done employing three widely utilized datasets: IEMOCAP, EMO-DB, and RAVDESS. The emotion recognition accuracy was improved to 90.92% for the IEMOCAP dataset, 93% for the EMO-DB dataset, and 92% for the RAVDESS dataset. The findings obtained from the testing of the suggested system have demonstrated the importance and strength of SER in accurately identifying the speaker's emotional state.

While MFCCs are treasured for confident aspects of speech capturing, MFCCs are computed over short, fixed-length frames of speech, which may not capture the dynamic changes in emotional

expression that occur over longer time scales. Nevertheless, MFCCs principally capture acoustic features it miss semantic content direct representation. In future work, longer analysis windows or overlapping frames may be used to capture more temporal context.

On the other hand, in deep bidirectional models there are two topologies, the forward and backward directions can be combined after each layer or only at the output layer. In this paper, we merely use one topology in which the training phase of the bidirectional RNN, forward and backward are combined in the output layer only. In future work, both topologies of deep Bi-LSTM can be applied in our framework and comparing them. Furthermore, audio data usage in conjunction with visual modalities can also be an exhilarating area for research in the future, by way of visual cues, such as body language and facial expressions, which can express appreciated emotional information. Moreover, the proposed framework can be further used for an aspiration for recognition and identification of the speaker which is used in many real-world problems.

Acknowledgement: The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project Number R-2023-757.

Funding Statement: Not applicable.

Author Contributions: Study conception and design: F. Harby, M. Alohali, and A. S. Talaat; data collection: F. Harby, A. theljeoui, and A. S. Talaat; analysis and interpretation of results: F. Harby, M. Alohali, and A. S. Talaat; draft manuscript preparation: F. Harby, A. theljeoui, and A. S. Talaat. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Zhang, W. Zhu, and J. Li, "Spiking echo state convolutional neural network for robust time series classification," *IEEE Access*, vol. 7, pp. 4927–4935, 2018. doi: 10.1109/ACCESS.2018.2887354.

[2] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Calgary, AB, Canada, IEEE, 2018, pp. 5089–5093.

[3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019. doi: 10.1109/ACCESS.2019.2936124.

[4] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *ICASSP 2021–2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, IEEE, 2021, pp. 6648–6652.

[5] D. Lee, S. Kim, and J. W. Choi, "Inter-channel Conv-TasNet for multichannel speech enhancement," arXiv preprint arXiv:2111.04312, 2021.

[6] S. Kumar *et al.*, "Multilayer neural network based speech emotion recognition for smart assistance," *Comput., Mater. Contin.*, vol. 75, no. 1, pp. 1523–1540, 2023. doi: 10.32604/cmc.2023.028631.

[7] D. Michelsanti *et al.*, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368–1396, 2021. doi: 10.1109/TASLP.2021.3066303.

[8] J. Ye, X. C. Wen, Y. Wei, Y. Xu, K. Liu and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *ICASSP 2023–2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, IEEE, 2023, pp. 1–5.

[9] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "Speechformer++: A hierarchical efficient framework for paralinguistic speech processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 775–788, 2023. doi: 10.1109/TASLP.2023.3235194.

[10] C. Wang, Y. Ren, N. Zhang, F. Cui, and S. Luo, "Speech emotion recognition based on multi-feature and multi-lingual fusion," *Multimed. Tools Appl.*, vol. 81, pp. 4897–4907, 2022. doi: 10.1007/s11042-021-10553-4.

[11] S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Syst. Appl.*, vol. 167, pp. 114177, 2021. doi: 10.1016/j.eswa.2020.114177.

[12] S. U. Bhandari, H. S. Kumbhar, V. K. Harpale, and T. D. Dhamale, "On the evaluation and implementation of LSTM model for speech emotion recognition using MFCC," in *Proc. Int. Conf. Comput. Intell. Data Eng.: ICCIDE*, Singapore, Springer Nature Singapore, 2022, pp. 421–434.

[13] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang and M. Akagi, "Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020. doi: 10.1109/ACCESS.2020.2967791.

[14] M. Venkata Subbarao, S. K. Terlapu, N. Geethika, and K. D. Harika, "Speech emotion recognition using k-nearest neighbor classifiers," in *Recent Adv. Artif. Intell. Data Eng.: Select Proc. AIDE*, Singapore, Springer Singapore, 2021, pp. 123–131.

[15] G. Sowmya, K. Naresh, J. D. Sri, K. P. Sai, and D. V. Indira, "Speech2Emotion: Intensifying Emotion detection using MLP through RAVDESS dataset," in *Int. Conf. Electron. Renew. Syst. (ICEARS)*, IEEE, 2022, pp. 1–3.

[16] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep Bi-LSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020. doi: 10.1109/ACCESS.2020.2990405.

[17] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, 2020. doi: 10.1007/s10772-020-09672-4.

[18] Z. Chen, M. Lin, Z. Wang, Q. Zheng, and C. Liu, "Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms," *Knowl.-Based Syst.*, vol. 281, pp. 111077, 2023. doi: 10.1016/j.knosys.2023.111077.

[19] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, "Attention-based Bi-LSTM models for personality recognition from user-generated content," *Inf. Sci.*, vol. 596, pp. 460–471, 2022. doi: 10.1016/j.ins.2022.03.038.

[20] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Fifteenth Annu. Conf. Int. Speech Commun. Assoc.*, pp. 2685–2689, 2014.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[22] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proc. 2011 Int. Conf. Electron. Mech. Eng. Inf. Technol.*, Harbin, China, IEEE, 2011, vol. 2, pp. 621–625.

[23] J. G. Wilpon and D. B. Roe, *Voice Communication between Humans and Machines*. USA: National Academies Press, 1994.

[24] R. M. Hanifa, K. Isa, and S. Mohamad, "Comparative analysis on different cepstral features for speaker identification recognition," in *2020 IEEE Student Conf. Res. Dev. (SCOReD)*, Batu Pahat, Malaysia, IEEE, 2020, pp. 1–6.

[25] S. A. Alim and N. K. Rashid, "Some commonly used speech feature extraction algorithms," in *From Natural to Artificial Intelligence-Algorithms and Applications*, 2018, pp. 2–19. doi: 10.5772/intechopen.80419.

[26] M. Kalpana Chowdary and D. Jude Hemanth, "Deep learning approach for speech emotion recognition," in *Data Anal. Manage.: Proc. ICDAM*, Springer, 2021, pp. 367–376.

[27] W. Bao, Y. Li, M. Gu, J. Tao, L. Chao and S. Liu, "Combining prosodic and spectral features for Mandarin intonation recognition," in *9th Int. Symp. Chinese Spoken Lang. Process.*, IEEE, 2014, pp. 497–500.

[28] J. T. Abraham, A. N. Khan, and A. Shahina, "A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients," *Int. J. Speech Technol.*, vol. 26, no. 3, pp. 1–9, 2021. doi: 10.1007/s10772-021-09888-y.

[29] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, "Prediction of emotions from the audio speech signals using MFCC, MEL and Chroma," in *2020 12th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Bhimtal, India, IEEE, 2020, pp. 87–91.

[30] S. Sen *et al.*, "Speech processing and recognition system," *Audio Process. Speech Recognit.: Concepts, Tech. Res. Overviews*, pp. 13–43, 2019. doi: 10.1007/978-981-13-6098-5_2.

[31] S. Bhattacharya, S. Borah, B. K. Mishra, and A. Mondal, "Emotion detection from multilingual audio using deep analysis," *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 41309–41338, 2022. doi: 10.1007/s11042-022-12411-3.

[32] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, 1997. doi: 10.1109/34.574797.

[33] S. Gupta and A. Mehra, "Gender specific emotion recognition through speech signals," in *2014 Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Noida, India, IEEE, 2014, pp. 727–733.

[34] A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Trans. Neur. Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1421–1432, 2013. doi: 10.1016/j.specom.2013.09.010.

[35] A. Ogawa and T. Hori, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks," *Speech Commun.*, vol. 89, pp. 70–83, 2017. doi: 10.1016/j.specom.2017.02.009.

[36] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017. doi: 10.1109/ACCESS.2017.2778011.

[37] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE Workshop Auto. Speech Recognit. Understanding*, IEEE, 2013, pp. 273–278.

[38] R. S. Sudhakar and M. C. Anil, "Analysis of speech features for emotion detection: A review," in *2015 Int. Conf. Comput. Commun. Control Autom.*, Pune, India, IEEE, 2015, pp. 661–664.

[39] I. Pulatov, R. Oteniyazov, F. Makhmudov, and Y. I. Cho, "Enhancing speech emotion recognition using dual feature extraction encoders," *Sens.*, vol. 23, no. 14, pp. 6640, 2023. doi: 10.3390/s23146640.

[40] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimed.*, vol. 20, no. 6, pp. 1576–1590, 2017. doi: 10.1109/TMM.2017.2766843.

[41] S. Basu, J. Chakraborty, and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *2nd Int. Conf. Commun. Electron. Syst. (ICCES)*, Coimbatore, India, IEEE, 2017, pp. 333–336.

[42] Z. Zhao *et al.*, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019. doi: 10.1109/ACCESS.2019.2928625.

[43] 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. Letters*, vol. 25, no. 10, pp. 1440–1444, 2018. doi: 10.1109/LSP.2018.2860246.

[44] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019. doi: 10.1109/ACCESS.2019.2938007.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[46] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *2019 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Cambridge, UK, IEEE, 2019, pp. 732–737.

[47] J. Parry *et al.*, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 1656–1660. doi: 10.21437/Interspeech.2019.
[48] S. A. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sens.*, vol. 20, no. 1, pp. 183, 2019. doi: 10.3390/s20010183.