Pertanika J. Sci. & Technol. 29 (3): 1915 - 1936 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Speech Emotion Recognition Using Deep Learning LSTM for Tamil Language

Bennilo Fernandes* and Kasiprasad Mannepalli

Department of Electronics & Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

ABSTRACT

Deep Neural Networks (DNN) are more than just neural networks with several hidden units that gives better results with classification algorithm in automated voice recognition activities. Then spatial correlation was considered in traditional feedforward neural networks and which do not manage speech signal properly to it extend, so recurrent neural networks (RNNs) were implemented. Long Short-Term Memory (LSTM) systems is a unique case of RNNs for speech processing, thus considering long-term dependencies Deep Hierarchical LSTM and BiLSTM is designed with dropout layers to reduce the gradient and long-term learning error in emotional speech analysis. Thus, four different combinations of deep hierarchical learning architecture Deep Hierarchical LSTM and LSTM (DHLL), Deep Hierarchical LSTM and BiLSTM (DHLB), Deep Hierarchical BiLSTM and LSTM (DHBL) and Deep Hierarchical dual BiLSTM (DHBB) is designed with dropout layers to improve the networks. The performance test of all four model were compared in this paper and better efficiency of classification is attained with minimal dataset of Tamil Language. The experimental results show that DHLB reaches the best precision of about 84% in recognition of emotions for Tamil database, however, the DHBL gives 83% of efficiency. Other design layers also show equal performance but less than the above models DHLL & DHBB shows 81% of efficiency for lesser dataset and minimal execution and training time.

Keywords: BiLSTM, DNN, Emotional Recognition, LSTM, RNN

ARTICLE INFO

Article history:
Received: 30 January 2021
Accepted: 15 April 2021
Published: 31 July 2021

DOI: https://doi.org/10.47836/pjst.29.3.33

E-mail addresses:
bennij05@gmail.com (Bennilo Fernandes)
mkasiprasad@gmail.com (Kasiprasad Mannepalli)
* Corresponding author

INTRODUCTION

Advanced pipeline consists of various protocols and hand-engineered creation method depend on top speech processing. They define an end-to-end language method,

ISSN: 0128-7680
e-ISSN: 2231-8526

named "Artificial Language" whereby machine learning overrules these steps in the process. The above method, combined with such a classification algorithm, produces improved speed on difficult language processing activities than conventional technique, while still being somewhat easier. Deep Speech performs admirably recently written approaches mostly on phone line. As the result, one should always elaborately design the parts of the network for reliability to enhance effectiveness on such a function like recognizing voice in a loud background. Under comparison, exploiting neural networks, the framework utilizes deeper processing end-to end. To strengthen the overall quality, we reap the benefits of its information represented by deep learning models to train through massive data.

In this article, the characteristics of LSTM, and BiLSTM were analyzed for emotional voice recognition implemented to Tamil emotional information set and used a suitable clustering technique. To classify data sets, distinct user defined classifiers are based end-to-end utilizing CTC. The paper is, as continues to follow, organized. First it describes integrated research in the community of emotional voice recognition and machine learning and then gives the brief introduction about RNN and its layers. Then describes the feature extraction variables that are implemented and details about the dataset collection. Then the research performed, and the outcome are reported. Followed by the conclusion and discussions.

MATERIALS AND METHODS

In deep learning more than 20 years ago, feed-forward neural network analysis was carried were examined (Liu, Z. T. et al., 2018; Cummins et al., 2017; Mustaqeem, & Kwon, 2020; Mannepalli et al., 2016a; Hussain et al., 2019). Recurrent neural networks and fully connected layers models were being used at almost the same time in voice recognition (Huang et al., 2019; Khan et al., 2019; Karim et al., 2019; Khalil et al., 2019). Most recent, with an almost all state-of-the-art speech task comprising a few other types of recurrent neural network, DNNs are becoming a feature throughout the ASR pipeline (Zhang et al., 2019; Tzirakis et al., 2018; Mannepalli et al., 2016b; Kumar et al., 2017). It has also been noticed that convolutionary networks are useful for acoustic systems (Badshah et al., 2019; He et al., 2016; Jiang, 2019; Wang et al., 2018). In nation-of-the-art recognizers, deep neural networks, generally LSTMs (Rao et al., 2018; Khamparia et al., 2019; Navyasri et al., 2017) are now just starting to be implemented and perform well coevolutionary layers for the retrieval of functions (Krizhevsky et al., 2012; Rao & Kishore, 2016; Ocquaye et al., 2019). It has also tested systems in both bidirectional and unidirectional recurrence. End-to-end emotional speech recognition is indeed a popular field of research that, when it is used to score the performances of DNN-HMM (Zeng & Xiao, 2019; Xie et al., 2018; Kishore & Prasad, 2016) shows promising performance. The RNN group performed well enough with graphemic results in emotional speech recognition. It is exposed that the

CTC-RNN system performed well during determining of emotional recognition although a vocabulary will still be required in either case (Sainath et al., 2015; Tzirakis et al., 2017; Ma et al., 2016). In addition, the CTC-RNN process were to be pre-trained with a DNN back propagation network provided by frame-wise formations first from HMM system (Zhang et al., 2018; Liu, B. et al., 2018; Ma et al., 2018; Sastry et al., 2016). In addition, without depending on the frame-wise groupings for preprocessing, we practice the CTC-RNN channels. To date, the exploitation of volume in neural networks has also been important to the field's success.

Recurrent Neural Network (RNN)

The Sequential data theory is used by RNNs. The RNN, a neural network with a recent memory that affects forecasting accuracy besides observations, sequences files contained in RNN memory is being used. In the conventional neural network, the concept of using RNN rather than the convolutional neural network is that any inputs are not dependent on one another. Therefore, use of recurrent neural networks in emotional speech processing is a better idea (Mustaqeem & Kwon, 2020).

LSTM

German researchers identified LSTM or Long Short-Term Memory with in mid-90s mostly as difference of its recurring channel to long short-term memory modules. It is indeed one generation closer with RNN. Whereas the issue of dimensionality reduction and explosion slope has been endured by a recurring channel, these were developed to the two issues described previous section. Analyzing the design methodology of LSTM, the layers of input and output are comparable to others in the RNN. In the center node or the replicating module, the difference lies. In LSTM, the repetition part uses 4 layers rather than one, as in RNN. Such layers come into contact with each other in LSTM, but this communication appears to be part of LSTM's decision-making procedure. The Figure 1 below shows the repetitive subsystem architecture.

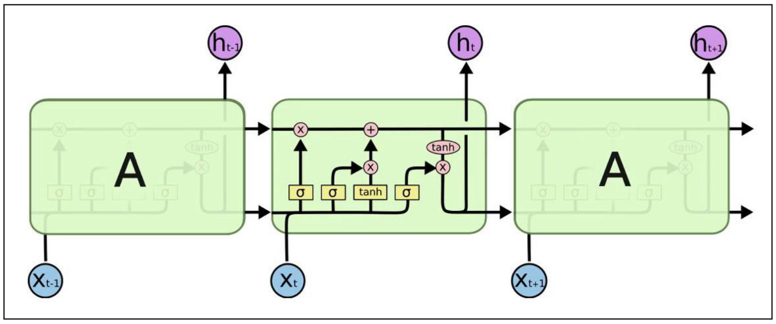


Figure 1. Sequential LSTM layer internal architecture

The line graph which represents plus, and cross symbol show the state of its node and behaves as major factor to LSTM. LSTM can change reflective process throughout the cell. It behaves like a chain which determines the amount of information from all 4 layers to also be handled. The '+' and 'X' represent the gates in each state of the cell. Gates chooses to either allow the data to next phase or not. Individuals are composed of an artificial neural layer including its sigmoid and a point wise procedure of multiplying. The sigmoid part provides the given number between 0 and 1, characterizing how often data to just let through in. Second, it is necessary to determine the latest data is stored throughout the total numbers, that is achieved by referencing the information of the input gate layer and the above-mentioned input training algorithm. Updating the new cell state will be the next input mode. Then it preceded by the objectives have been met computation, completed by the hidden layers of the output.

BiLSTM

Schuster and Paliwal (1997) developed bi-directional recurrent neural networks (BRNN) to incorporate 2 different hidden LSTM layers from reverse direction to almost the similar outcome to overcome the drawbacks of a singular LSTM cell which can only collect prior framework but not have the future context. By this architecture, the activation function can use the comparison to previous framework of specific aspects. The sequence input $x = (x_1, x_2, \ldots., x_n)$ is measured by a BiLSTM again from reverse direction with a forward hidden pattern $\vec{h} = (\vec{h}_1, \vec{h}_2, \ldots., \vec{h}_n)$ and a primitive hidden sequence $\overleftarrow{h} = (\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots., \overleftarrow{h}_n)$. The encrypted variable yt, e, is established by the convolution including its finalized outputs forward and backward, $y_t = [\vec{h}_t, \overleftarrow{h}_t]$ (Equation 1-4).

$$\vec{h}_t = \sigma \left(W_{\vec{h}x} x_t + W_{\vec{h}\vec{h}} \vec{h}_{t-1} + b_{\vec{h}} \right) \qquad [1]$$

$$\overleftarrow{h}_t = \sigma \left(W_{\overleftarrow{h}x} x_t + h\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}} \right) \qquad [2]$$

$$h_t = W_{hh}\vec{h}_t + W_{hh}\overleftarrow{h}_t + b_h \qquad [3]$$

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \qquad [4]$$

Where the throughput pattern of the first hidden units is $y = (y_1, y_2, \ldots., y_n)$. In addition, there may be some relevant studies assistance to demonstrate that even a high system structure is much more effective than a simplistic one in portraying a few other functions. Thus, this paper has identified a layered BiLSTM network in which the y_t output again from lower layer will become the higher layer feedback. Figure 2 illustrates the loaded BiLSTM system.

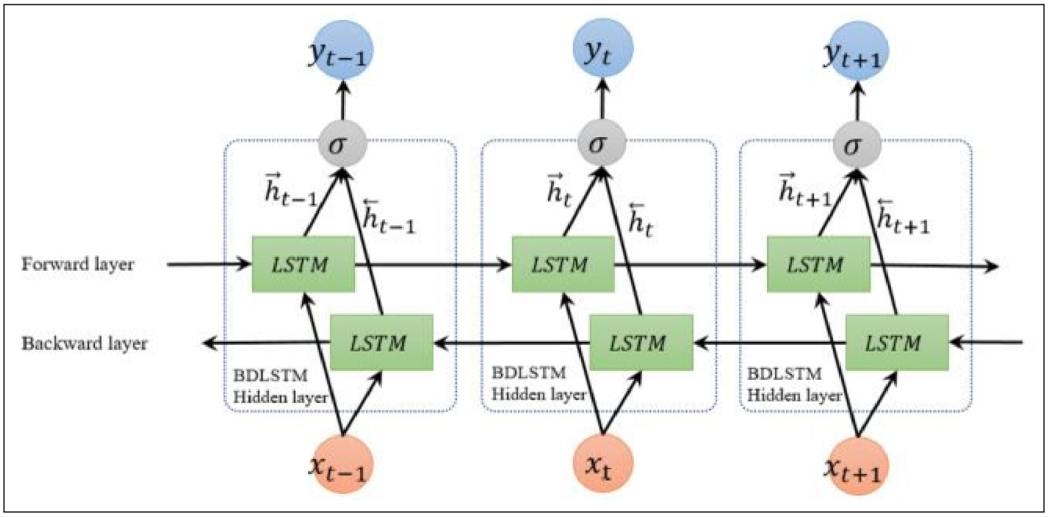


Figure 2. BiLSTM layer internal architecture

Describing $Q = (q_1, q_2, \ldots., q_n)$ as well as $A = (a_1, a_2, \ldots., a_n)$ to portray discussion patterns and response sequential output during which n and m signify the duration of the answers to questions including both q_t and a_t clearly show the responses to questions throughout the t-th sentences. To acquire their previous hidden structures, HQ and HA, a layered BiLSTM over the responses to questions, and the arithmetic is as continues to follow in which *d* seems to be the hidden layer factor (Equation 5-8).

$$h_t^q = sBiLSTM\left(h_{t-1}^q, h_{t+1}^q, q_t\right),\ h_0^q = 0, \qquad [5]$$

$$h_t^a = sBiLSTM\,(h_{t-1}^a, h_{t+1}^a, a_t),\ h_0^a = h_n^q, \qquad [6]$$

$$H_Q = \left[h_1^q, h_2^q, \ldots., h_n^q\right] \epsilon\, R^{d*n} \qquad [7]$$

$$H_A = [h_1^a, h_2^a, \ldots., h_m^a]\ \epsilon\, R^{d*m} \qquad [8]$$

Feature Extraction

Mel Frequency Cepstral Coefficients (MFCC). MFCC is determined by the characteristics of listening in the human ear, which simulates the human auditory system using a nonlinear frequency unit. The Fast Fourier Transform (FFT) technique is optimally used to transform, as explained in, each sample frame from the time domain into the frequency domain (Equation 9).

$$\mathrm{S[k]} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{s[n].e}^{\frac{-\mathrm{j2\pi nk}}{\mathrm{N}}}, 0 \leq \mathrm{k} \leq \mathrm{N} - 1 \qquad [9]$$

The mel filter bank is composed of overlapping triangular filters with the cutoff frequencies determined by the two adjacent filters' center frequencies. The filtration has centre frequencies linearly distributed, and fixed mel scale bandwidth. The logarithm seems to have the impact, mentioned in, of shifting multiplier into addition (Equation 10).

$$\mathrm{F[m]} = \log \sum_{\mathrm{n=0}}^{\mathrm{N-1}} |\mathrm{x[k]}|^2 \mathrm{H_m[k]}, 0 \leq \mathrm{m} \leq \mathrm{M} \qquad [10]$$

Ultimately, to find the MFCC, the Discrete Cosine Transform (DCT) of the log wavelet packet energy is computed (Equation 11).

$$\mathrm{c[n]} = \sum_{\mathrm{m=1}}^{\mathrm{M}} \mathrm{s[n]}.\mathrm{e}^{\frac{-\mathrm{j}2\pi\mathrm{nk}}{\mathrm{N}}}, 0 \leq \mathrm{k} \leq \mathrm{N}-1 \qquad [11]$$

MFCC Delta

MFCC Delta, also known as variance and maximum speed coefficients. The features of MFCC vector explains only the power spectral functions of single frames, but in speech data the information will be obtained in dynamic values and more variation in features, what the trajectories of the MFCC features extracted are done with over time. It gives an out turns that calculating and appending the MFCC trajectories to the vectors of real and original features increases the performance of ASR by quite a bit (if we have 12 MFCC coefficients, we would also get 12 delta coefficients, which would combine to give a length 24 feature vector). The following formula is employed to calculate the delta coefficients (Equation 12):

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad [12]$$

where d_t is a delta coefficient, t frames are computed in terms of the static coefficients c_{t+n} to $c_{t\text{-}n}$. A typical value for N is 2.

The Bark Scale

It is based on the key throughput idea, is predictable underneath 500 Hz. e Bark scale outcomes from portraying an entire band of wavelengths with sequences of critical bands and not allowing to merge them. The Bark 1 to 24 numbers are the 24th critical band in the proceedings. Equivalent Rectangular Bandwidth ERB respective logarithmic and sequential; that every dimension is like Bark scale as it also offers an approximation of bandwidths of high noise filters, and therefore utilizes rectangular (unachievable recognition) band-pass filters to efficiently optimize filter modelling. The case hardening conversion would be between ERB and Hertz.

Spectral Kurtosis

The component associated with the execution of extracting features is spectral kurtosis, but it symbolizes the statistical relationship from both voice samples (Xie et al., 2018). Throughout the transmissions, the spectral kurtosis could still be described as the value of kurtosis of the variables of voice, and therefore is described as Equation 13:

$$F_5 = \frac{a_4\{S^*(m), S^*(m), S^*(m), S^*(m)\}}{a_2\{S^*(m), S^*(m)\}^2} \qquad (13)$$

During which $S^*(m) \in \{S(m), S^c(m)\}$ the complicated conjugate of the process parameters $S^c(m)$ is demonstrated by S(m) as well as the accumulated fourth and second order are stated by a_4 and a_2.

Spectral Skewness

The spectral skewness (Wang et al., 2018) demonstrates the irregularities in the spectrum ‘s distribution of the voice signal on its average rating. The spectral skewness further assumes the energy level of its spectrum via transfer. Unless the energy size is low upon this distribution left side, it will be very strong if the spectral variable of skewness contains its speech signal.

Dataset

The emotional voice signals are recorded through mobile apps for training and research. All inputs are captured in 44KHz frequency mono signal. The samples collected were utilized for the simulation purpose. Speech information is obtained from 10 individual male and female speakers individually. Every speaker has been asked to utter 10 times each sentence in different emotions like anger, happy, sad, fear, disgust, neutral and boredom. The sentence I have taken is “Na nalla iruken ennaku onnum illa”. Both male and female speakers report a total of 1400 emotional speech data samples. These samples were taken into consideration for this design flow analysis. A sentence-based samples were recorded by students of arts. For testing purpose, the samples were collected with co-working faculty to identify their emotions during their counselling period. Totally 50 samples were collected with same 44KHz through same mobile apps. Thus these 50 samples were tested to identify the emotions of working faculty. Since the training data base were collected by the professional actors, taking that Tamil emotional data as base the testing emotion database can be identified with more accuracy and perfection.

RESULTS AND DISCUSSION

With sequential data input, the emotional speech database is analyzed in this design layer. The speech signal is converted to LSTM / BiLSTM Layers as sequential vectors and

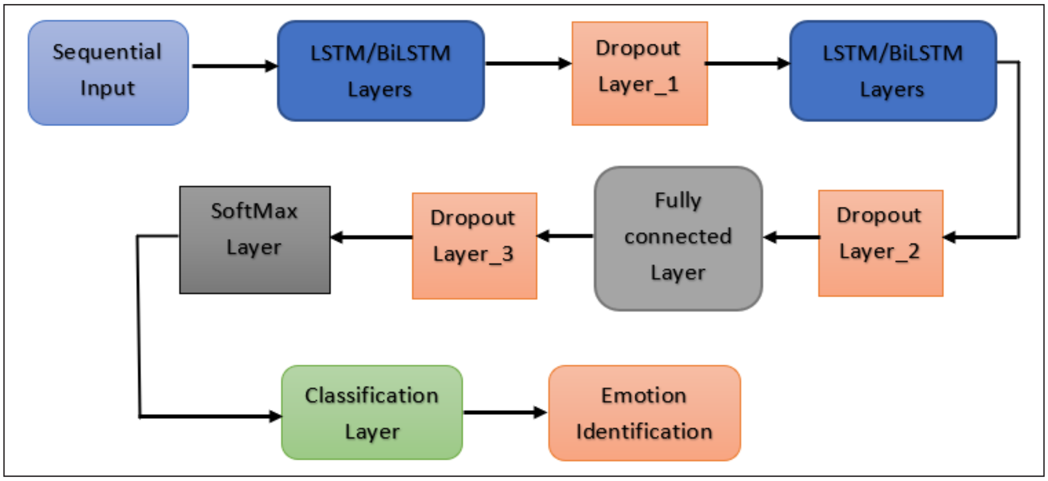


Figure 3. Proposed DNN design layer architecture

then passed to them. MFCC, MFCC delta, Bark spectrum, Spectral kurtosis and Spectral Skewness are the extraction characteristics selected for this design analysis shown in Figure 3. For testing and training, all the characteristics were examined and concatenated for each speech data to identifies its mean and standard deviation. The vector feature per sequence is assigned to 20 and total number of feature overlapping is 10. With these characteristics the evaluation for different design layer structures that have been fixed. Adam is the optimizing algorithm used here. The Adam optimization algorithm is applied to back-propagation, which has been utilized in many areas recently for the analysis of deep learning and other applications like artificial intelligence and computer vision. Integration, some of the common features of Adam, is directly forwarded for experimentation. Effectiveness in computation. Tiny specifications for recollection. Wavelet transform for adjusting patterns diagonal direction. Well suited for problems with information- and/or parameter-size. Good for goals that are non-stationary.

For very noisy/ or scattered gradients, an effective algorithm. Hyper-parameter interpretation is user-friendly and usually includes minor changes. Optionally, during each single era, the data must optimize the training weights numerous times. The volume of material which is included in almost every transformation in sub-epoch weight is known as the size of the batch. For example, with a 50-voice test set, an entire batch size would be 1000, a 500 or 200 or 100 mini batch size, and batch size will define the deep function of training and testing of data, thus mini batch size is set to 250 and for the evaluation, the number of hidden layers is 500 and the initial learning rate is 0.005 and the max epoch is 10. Well after the epoch increases, the iteration can increase the efficiency by continuously training data, but the accuracy and loss during iteration remain the same. The accuracy level of the training dataset after 10 epochs has not been modified. The timeline for the learning rate is piecemeal. Dropout is a method that addresses both problems. This prevents overfitting and provides a way to effectively combine numerous different neural networks exponentially. The word dropout refers to the dropping out of units in a neural network (hidden and visible). In the simplest case, each unit is maintained with a fixed probability p, independent of other units, where p can be selected using a validation set or simply set to 0.5, which seems to be almost optimal for several networks and operations. The optimal retention probability, however, for the input units is generally closer to 1 than to 0.5. For design layer analysis three dropout layers were accomplished after each LSTM. The probability values are 0.5 each. The LSTM / BiLSTM design layer was analyzed by fixing all these parameters.

Deep Hierarchal LSTM&LSTM (DHLL) Model

As mentioned before the input speech signal is converted into the sequential data and processed to the dropout layer. The performance of the models is analyzed to reach a conclusion that DHLL model generates confusion matrix with 10-fold cross validation. Since cross folding is random each evaluation output shows different accuracy level a mean of 5 evaluation was considered for DHLL accuracy rate. Among the average 10 folds cross valuation fold 3 shows 85.7% of accuracy and fold 4 and 6 shows 83.98% of accuracy shown in Figure 4, where other folds also show better performance of accuracy around 70-80%. In the testing phase 50 samples of emotions were given as input for analysis of emotional recognition. From the 5 evaluation the best and higher accuracy level obtained in DHLL is 80.1%.

In the five evaluation the time taken to training and evaluation of classification timings were considered from Figure 5. While taking the mean value for training of DHLL takes around 7.86 Mins and to evaluate the classification it takes around 1.36 mins.

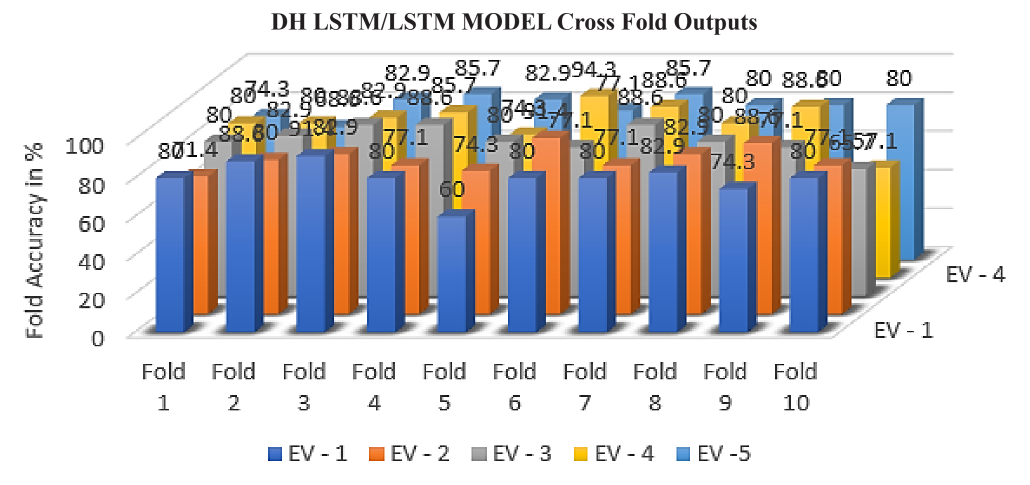


Figure 4. DHLL Cross fold output for multiple evaluation

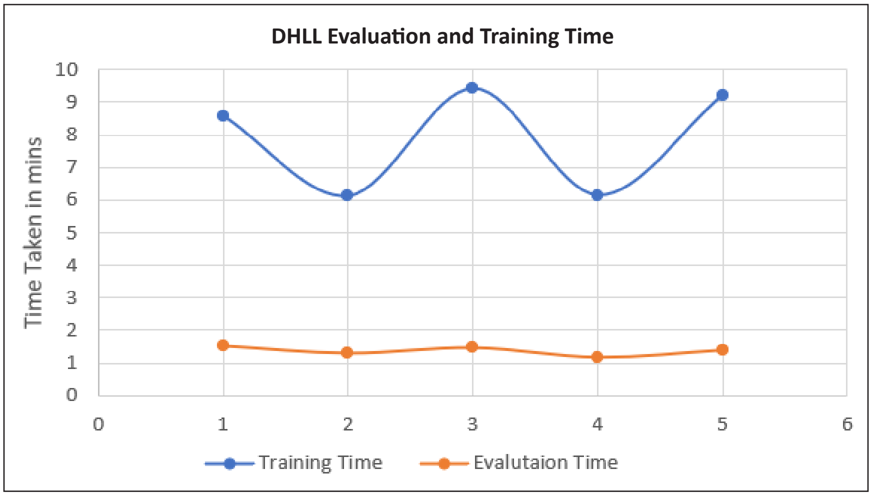


Figure 5. DHLL performance of evaluation time and training time

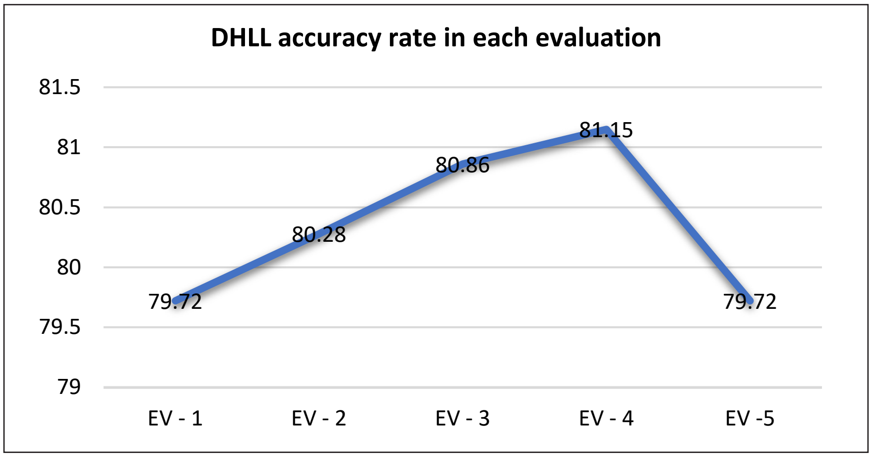
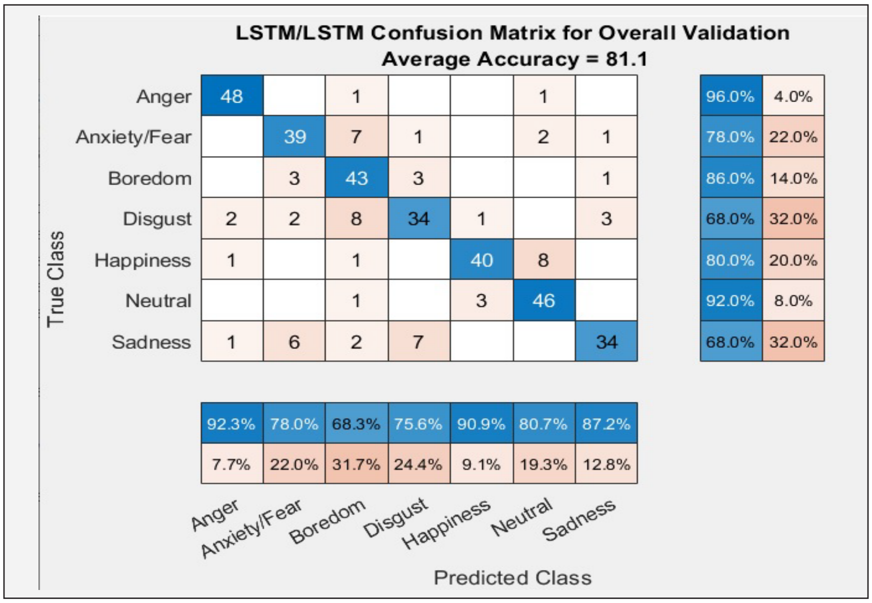


Figure 6. DHLL accuracy rate for 5 evaluations

Finally, by considering the higher accuracy level iteration it shows time taken is 6.14 and 1.18 for training and evaluation (Figure 6). Since the data set is recorded with 44khz of mono signal the evaluation time for training the dataset extend to 9.41mins, but the accuracy level is low than the best accuracy rate.

Thus, from Figure 6 its concluded that best accuracy obtained for Tamil emotional dataset in DH LSTM/LSTM model is 81.1%, but disgust is lagging at higher rate. From Figure 7, it is clear that still emotions like disgust are mapped or overlapped with other emotions like boredom and sadness. The confusion matrix shows emotions like anger and neutral shows better accuracy rate of 96% and 92%. Whereas other emotions have some average performance towards their own parent class. Emotion like sadness, disgust and happiness has only 68% of accuracy and 32% of loss can be seen in row normalization.



LSTM/LSTM Confusion Matrix for Overall Validation
Average Accuracy = 81.1

True Class	Anger	Anxiety/Fear	Boredom	Disgust	Happiness	Neutral	Sadness		
Anger	48		1			1		96.0%	4.0%
Anxiety/Fear		39	7	1		2	1	78.0%	22.0%
Boredom		3	43	3			1	86.0%	14.0%
Disgust	2	2	8	34	1		3	68.0%	32.0%
Happiness	1		1		40	8		80.0%	20.0%
Neutral			1		3	46		92.0%	8.0%
Sadness	1	6	2	7			34	68.0%	32.0%
	92.3%	78.0%	68.3%	75.6%	90.9%	80.7%	87.2%		
	7.7%	22.0%	31.7%	24.4%	9.1%	19.3%	12.8%		

Predicted Class

Figure 7. Cross fold confusion matrix for DHLL

Deep Hierarchal LSTM & BiLSTM (DHLB) Model

The performance of DHLB models is analyzed to reach a conclusion that DHLB model generates confusion matrix with 10-fold cross validation as final classification output. AS like DHLL max 5 times the simulation is evaluated to find the consistent in accuracy level. Among 5 simulation the average 10 folds cross valuation fold 3 shows 88.58% of accuracy and fold 2 and 9 shows 84.56% of accuracy shown in Figure 8, where other folds also show better performance of accuracy around 70-80%. In the testing phase same 50 samples of data used in DHLL is utilized for analysis. Each iteration there is a small variation in identification of emotional recognition. The average accuracy level for 5 evaluation is around 81.54%.

Now by analyzing the time factor the five evaluation the time taken to training and evaluation of classification timings were considered from Figure 9. While taking the mean value it is clear that for training of DHLB takes around 5.63 Mins and to evaluate the classification it takes around 1.05 mins.

From Figure 10 the accuracy level in each simulation is established. As the cross-validation folds are random the accuracy level changes randomly. But it lies in the range of 80 to 84. In each simulation the training time and the evaluation time also varies, but only seconds of variation can be identified. Among the 5 simulation results, in second iteration higher range of results is identified i.e., 83.43%.

Finally, by considering the higher accuracy level iteration it shows that for the given input Tamil database DHLB model gives 83.4% of efficiency (Figure 11).

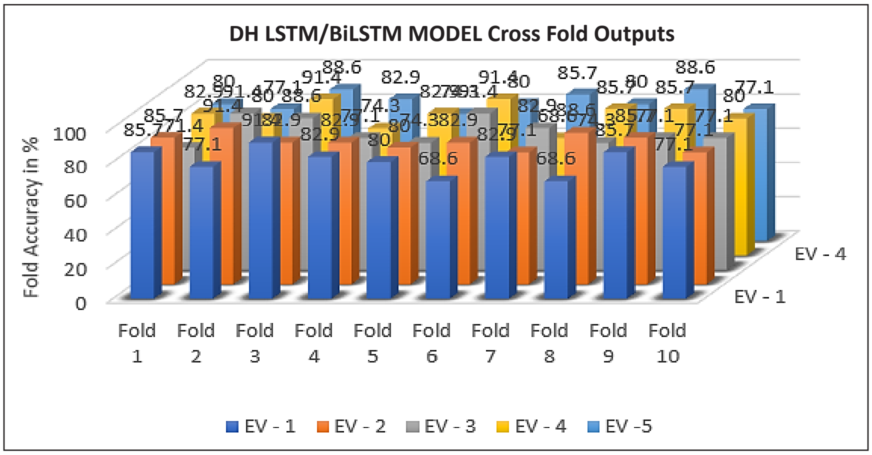


Figure 8. DHLB cross fold output for multiple evaluation

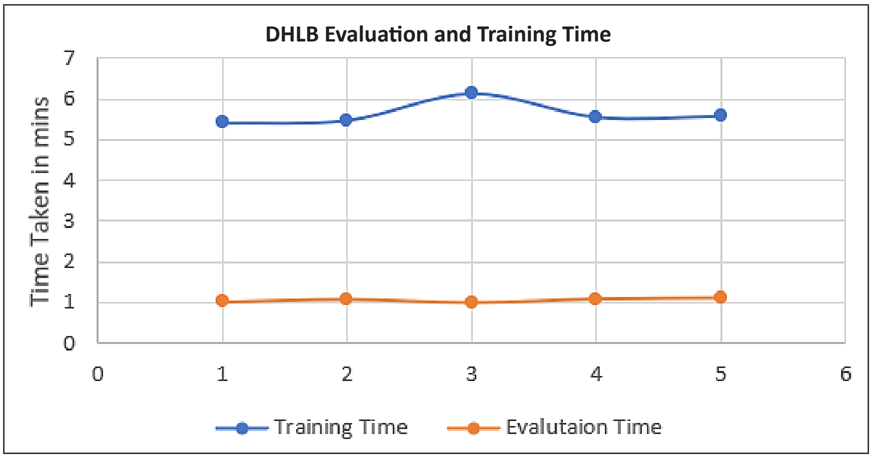


Figure 9. DHLB performance of evaluation time and training time

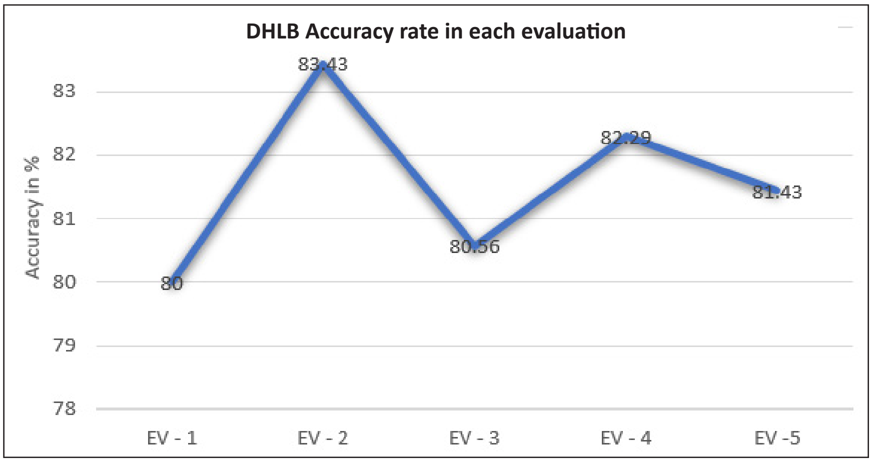
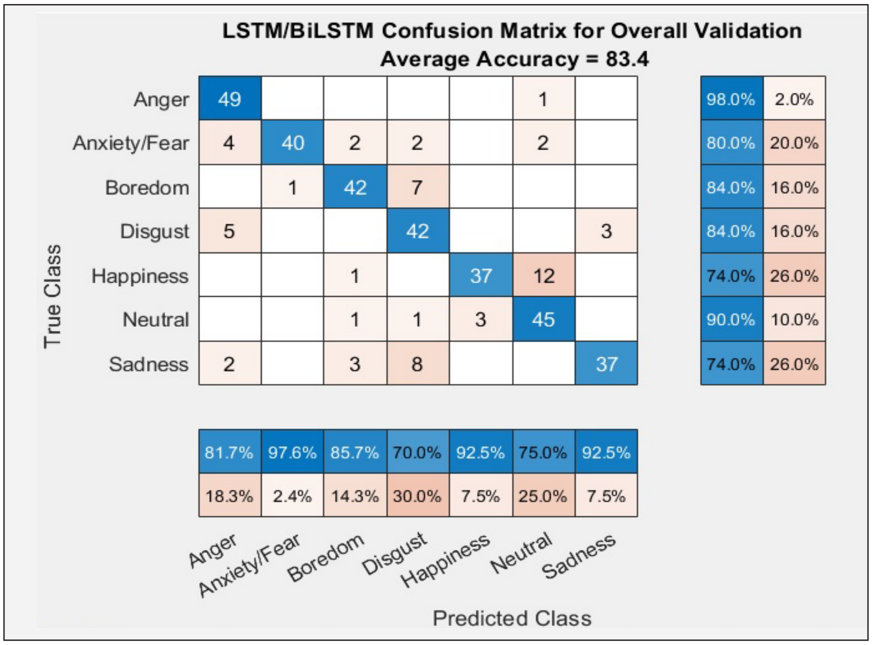


Figure 10. DHLB accuracy rate for 5 evaluations



LSTM/BiLSTM Confusion Matrix for Overall Validation
Average Accuracy = 83.4

True Class	Anger	Anxiety/Fear	Boredom	Disgust	Happiness	Neutral	Sadness		
Anger	49					1		98.0%	2.0%
Anxiety/Fear	4	40	2	2		2		80.0%	20.0%
Boredom		1	42	7				84.0%	16.0%
Disgust	5			42			3	84.0%	16.0%
Happiness			1		37	12		74.0%	26.0%
Neutral			1	1	3	45		90.0%	10.0%
Sadness	2		3	8			37	74.0%	26.0%
	81.7%	97.6%	85.7%	70.0%	92.5%	75.0%	92.5%		
	18.3%	2.4%	14.3%	30.0%	7.5%	25.0%	7.5%		

Predicted Class

Figure 11. Cross fold confusion matrix for DHLB

In the confusion matrix, emotions like anger and neutral gives higher rate of 98% and 90%. As like DHLL this model also lags in other emotional states. Happy and Neutral emotions lags in DHLB model. Only 74% of accuracy is obtained in both states and shows lowest of all emotion recognition. Fear, Boredom and Disgust shows 80% and 84% of accuracy.

Deep Hierarchal BiLSTM & LSTM (DHBL) Model

The DHBL models is analyzed to reach a conclusion that DHBL technique generates confusion matrix with 10-fold cross validation as final classification output. As like DHLB max 5 times the simulation is evaluated to find the consistent in accuracy level. Among 5 simulation the average 10 folds cross valuation fold 6 shows 88.56% of accuracy and fold 5, 3, and 8 shows 84% and 82% of accuracy shown in Figure 12, where other folds also show better performance of accuracy around 70-80%. In the testing phase same 50 samples of data used in DHLL is utilized for analysis. Each iteration there is a small variation in identification of emotional recognition. The average accuracy level for 5 evaluation is around 81.3%.

Now by analyzing the time factor the five evaluation the time taken to training and evaluation of classification timings were considered from Figure 13. While taking the mean value it is clear that for training of DHBL takes around 10.4 Mins and to evaluate the classification it takes around 0.54 mins.

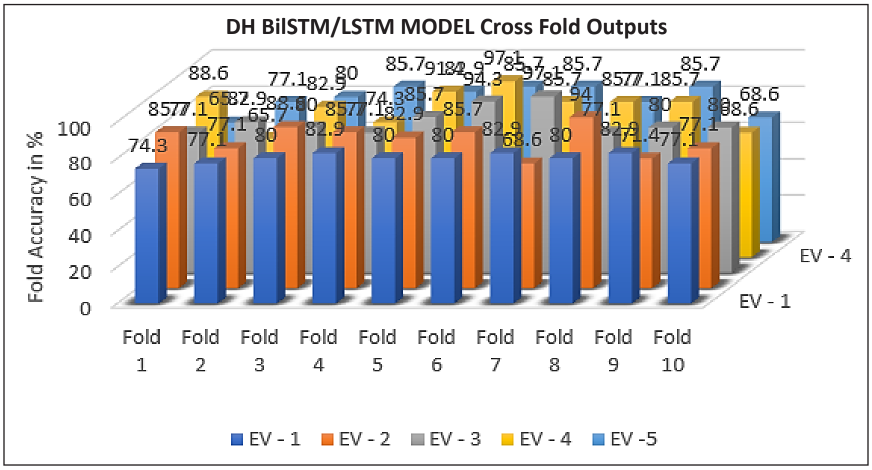


Figure 12. DHBL cross fold output for multiple evaluation

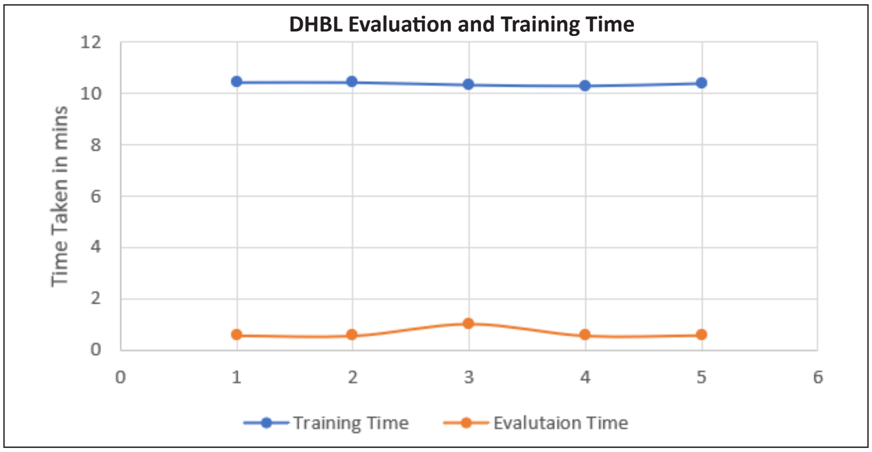


Figure 13. DHBL performance of evaluation time and training time

From Figure 14 the accuracy level in each simulation is established. As the cross-validation folds are random the accuracy level changes randomly. But it lies in the range of 79 to 84. In each simulation the training time and the evaluation time also varies, but only few seconds of variation can be identified. Among the 5 simulation results, in third iteration shows higher range of results is identified i.e., 83.13%.

Finally, by considering the higher accuracy level iteration it shows that for the given input Tamil database DHBL model gives 83.13% of efficiency (Figure 15). In the confusion matrix, emotions like anger and neutral gives higher rate of 98% and 92%. As like DHLL this model also lags in other emotional states. Disgust and Sadness emotions lags in DHBL model. Only 64% of accuracy is obtained in both disgust and sadness states and shows lowest of all emotion recognition. Fear, Boredom and Disgust shows 80% and 86% of accuracy.

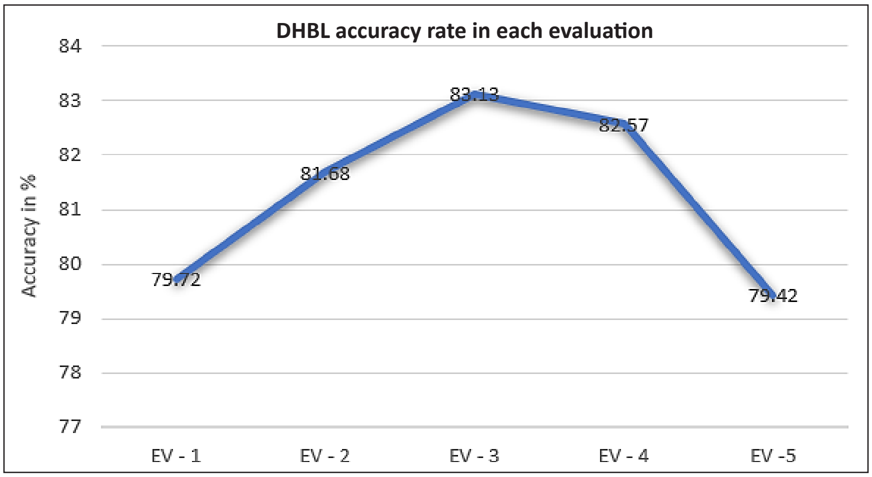


Figure 14. DHBL accuracy rate for 5 evaluations

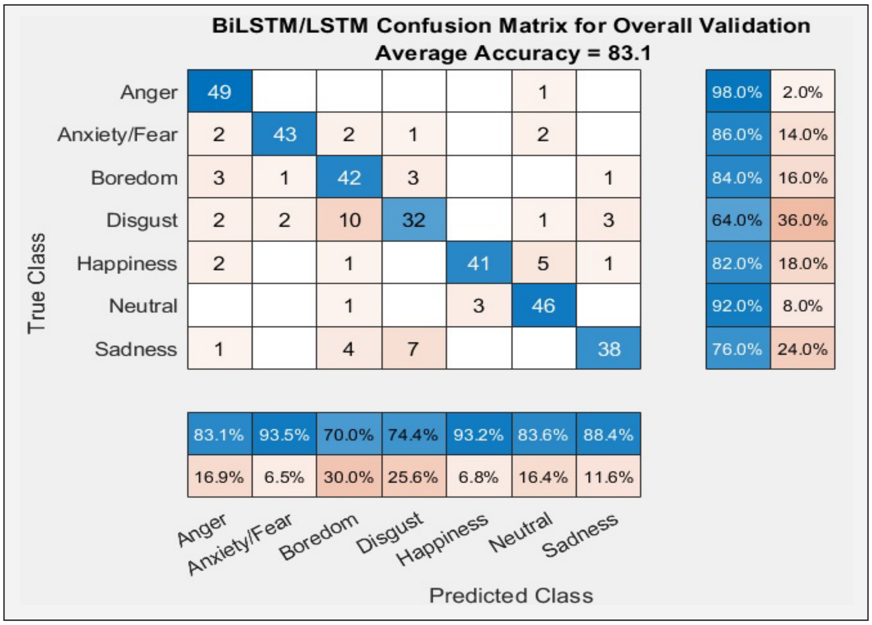


Figure 15. Cross fold confusion matrix for DHBL

Deep Hierarchal BiLSTM & BiLSTM (DHBB) Model

The DHBB models is analyzed to reach a conclusion that DHBB technique generates confusion matrix with 10-fold cross validation as final classification output. As like other models max 5 times the simulation is evaluated to find the consistent in accuracy level. Among 5 simulation the average 10 folds cross valuation fold 3 shows 86.26% of accuracy and fold 4 and 8 shows 80% of accuracy shown in Figure 16, where other folds also show better performance of accuracy around 70-80%.

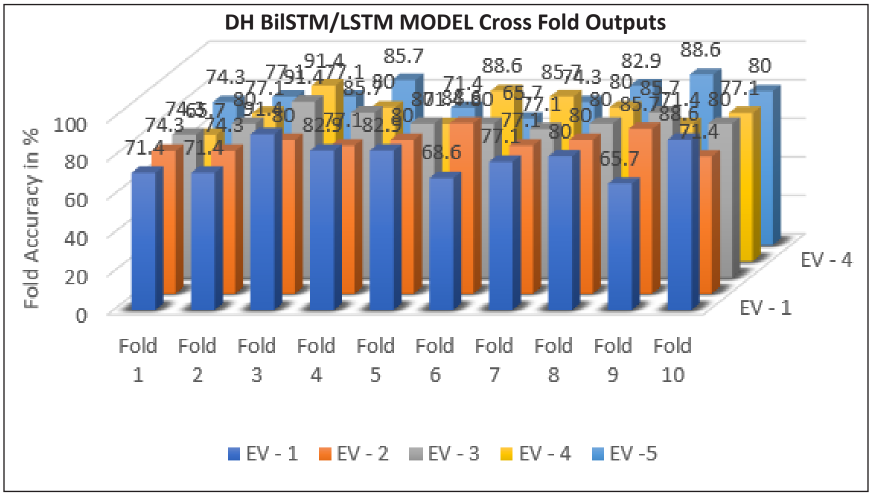


Figure 16. DHBB cross fold output for multiple evaluation

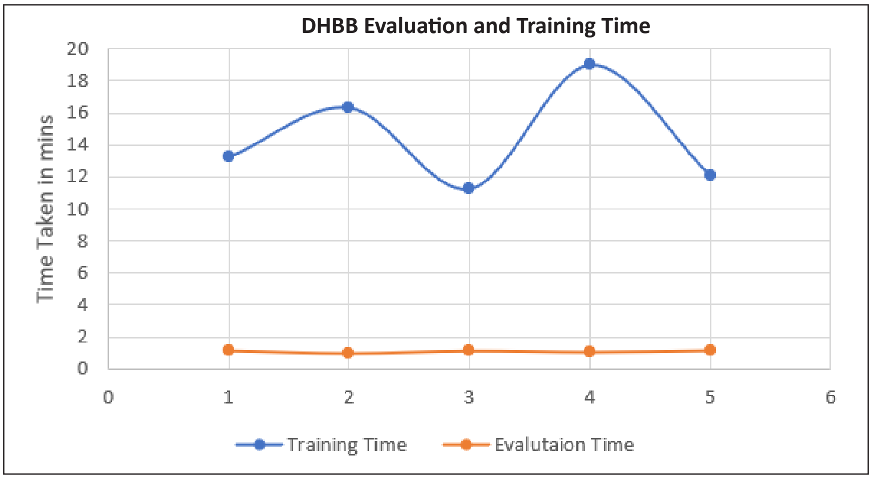


Figure 17. DHBB performance of evaluation time and training time

Each iteration there is a small variation in identification of emotional recognition. The average accuracy level for 5 evaluation is around 79.42%. Now by analyzing the time factor the five evaluation the time taken to training and evaluation of classification timings were considered from Figure 17. While taking the mean value it is clear that for training of DHBB takes around 14.3 Mins and to evaluate the classification it takes around 1.14 mins.

From Figure 18 the accuracy level in each simulation is established. As the cross-validation folds are random the accuracy level changes randomly. But it lies in the range of 78 to 82. In each simulation the training time and the evaluation time also varies, but only few seconds of variation can be identified. Among the 5 simulation results, in third iteration shows higher range of results is identified i.e., 83.13%.

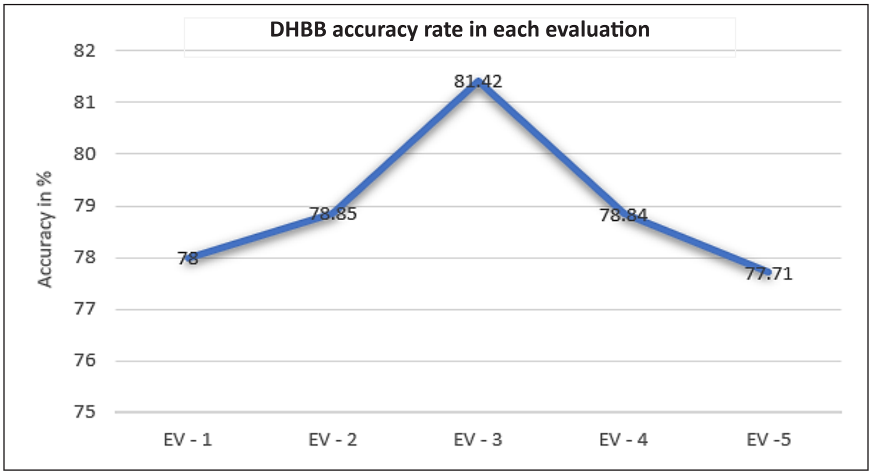
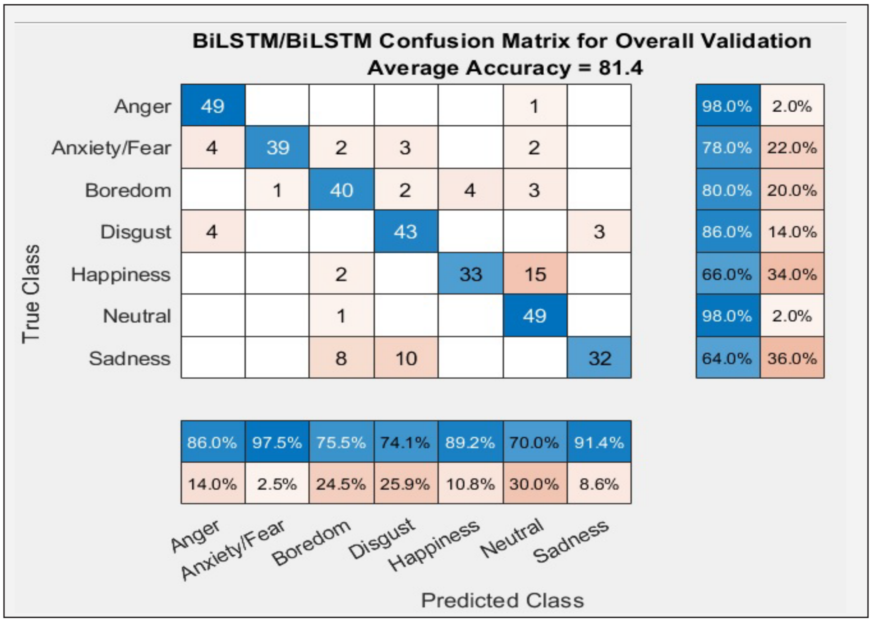


Figure 18. DHBB accuracy rate for 5 evaluations


BiLSTM/BiLSTM Confusion Matrix for Overall Validation
Average Accuracy = 81.4
True Class
Predicted Class
Anger
Anxiety/Fear
Boredom
Disgust
Happiness
Neutral
Sadness

	Anger	Anxiety/Fear	Boredom	Disgust	Happiness	Neutral	Sadness		
Anger	49					1		98.0%	2.0%
Anxiety/Fear	4	39	2	3		2		78.0%	22.0%
Boredom		1	40	2	4	3		80.0%	20.0%
Disgust	4			43			3	86.0%	14.0%
Happiness			2		33	15		66.0%	34.0%
Neutral			1			49		98.0%	2.0%
Sadness			8	10			32	64.0%	36.0%
	86.0%	97.5%	75.5%	74.1%	89.2%	70.0%	91.4%		
	14.0%	2.5%	24.5%	25.9%	10.8%	30.0%	8.6%		

Figure 19. Cross fold confusion matrix for DHBB

Finally, by considering the higher accuracy level iteration it shows that for the given input Tamil database DHBB model gives 81.4% of efficiency (Figure 19).

In the confusion matrix, emotions like anger and neutral gives higher rate of 98%. As like DHLL this model also lags in other emotional states. Happiness and Sadness emotions lags in DHBB model. Only 64% of accuracy is obtained in both happiness and sadness states and shows lowest of all emotion recognition. Fear, Boredom and Disgust shows 80% and 86% of accuracy.

Table 1
Cross fold accuracy of DH LL/LB/BL/BB layers

Fold Accuracy/Methodology	DHLL	DHLB	DHBL	DHBB
Fold 1	80	81.7	77.1	74.3
Fold 2	80	97.4	82.9	80
Fold 3	82.9	82.9	80	91.4
Fold 4	85.7	82.9	77.1	85.7
Fold 5	74.3	80	85.7	80
Fold 6	94.3	82.9	94.3	80
Fold 7	88.6	77.1	97.1	77.1
Fold 8	80	86.6	77.1	80
Fold 9	88.6	85.7	80	85.7
Fold 10	57.1	77.1	80	80

Table 2
Overall performance of DH LL/LB/BL/BB models

Overall Performance (5 Iteration)	DHLL	DHLB	DHBL	DHBB
Best Accuracy	81.15	83.43	83.13	81.42
Average accuracy	80.346	81.542	79.42	78.964
Best Training Time	6.12	5.41	10.27	11.26
Average Training Time	7.89	5.634	10.378	11.388
Best Evaluation Time	1.18	1.01	0.56	1.01
Average Evaluation Time	1.368	1.056	0.656	1.114

Thus, Tables 1 and 2 show the overall performance of the entire designs. Comparing with all the models DHLB shows better performance than the other models. Also, DHBL also achieves equal performance to DHLB. Both the models give accuracy of 84% for the collected Tamil emotional database. Now comparing the training time for all models DHLB acts better than DHBL. Even though DHBL shows equal performance towards DHLB, it takes more time for training the database.

More than half of the time is reduced in DHLB. Also, from Figures 20 and 21, it is clear that in cross fold 2 achieves highest percentage of accuracy of 97.4%. Among all model only 5.41 mins were taken to train the database in DHLB model, whereas DHBL & DHBB takes around 11 mins to complete the training. After training the testing is done to identify the different emotional classification for the input 50 samples. In testing also DHLB shows better evaluation time than other models, it takes only 1.05 mins to complete the evaluation.

In most efficient DHLB gives better performance in RNN followed by DHBL it is most. The LSTM model takes the least time for training and evaluation, where other techniques take slightly more time and DHBB takes the highest time. The results obtained

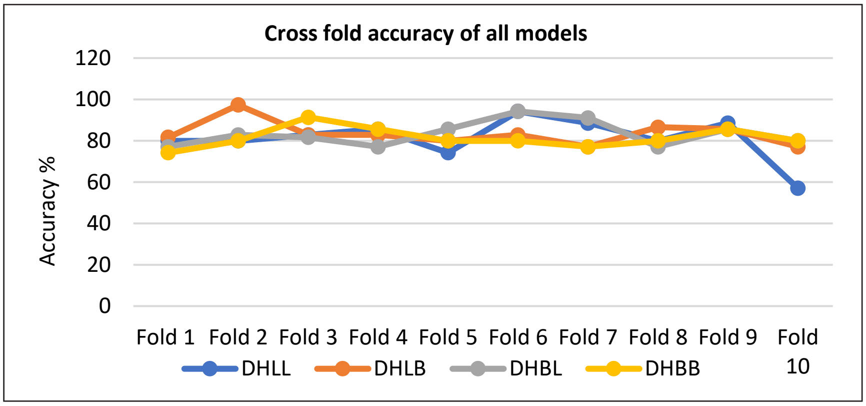


Figure 20. Cross fold accuracy of all different models

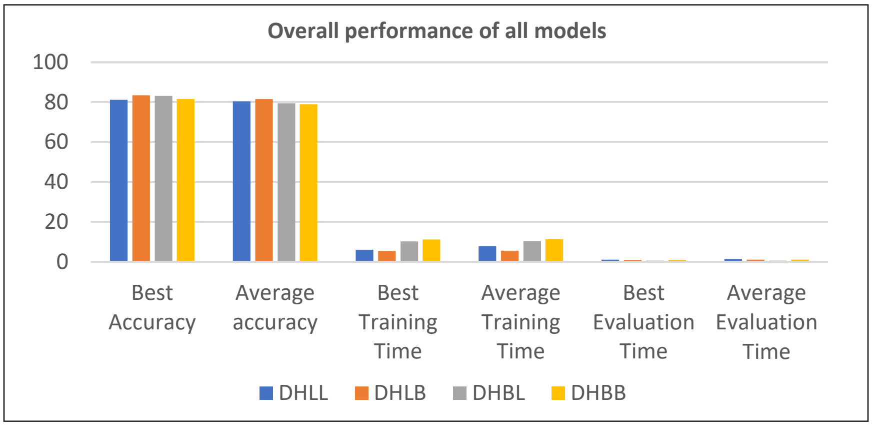


Figure 21. Overall Performance of all models

from different models are generated and presented effectively in this paper. We believe that further research can enhance this model and optimize it with lots of computation and data.

CONCLUSION

Since standard feedforward neural networks cannot handle speech data well (due to lacking a way to feed information from a later layer back to an earlier layer), thus, RNNs have been introduced to take the temporal dependencies of speech data into account. Furthermore, RNNs cannot handle the long-term dependencies due to vanishing/exploding gradient problem very well. Therefore, LSTMs and BiLSTM were introduced to overcome the shortcomings of RNNs. This paper evaluated RNN with hierarchal of LSTM and BiLSTM with dropout layers are compared their performances on interchanging the layers for a reduced Tamil emotional speech data set. Four different architectures were evaluated; DHLL, DHLB, DHBL and DHBB with dropout layers and the evaluation measures used were accuracy, loss, training time and evaluation time. The results show that the DHLB performs better than other models. Accuracy rate of 84% is achieved with minimum loss in each seven basic emotions and time taken for training and evaluation is also less than the other models. Thus, the recommendation for the reduced Tamil emotional speech data set is to use DHLB since it returned good efficiency of recognition values within an acceptable running time. Future work will include parameter optimization to investigate the influence on different parameter settings. Furthermore, the learning rate, dropout rate as well as higher numbers of neurons in the hidden layers will be experimented with to get more better performance.

ACKNOWLEDGEMENT

Authors would like to acknowledge the management of KL University, AP, India for supporting this work and the faculty members who supported them for making the data available for this research.

REFERENCES

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., & Baik, S.W. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications, 78*(5), 5571-5589. https://doi.org/10.1007/s11042-017-5292-7.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. W. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 478-484). Association for Computing Machinery. https://doi.org/10.1145/3123266.3123371.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). Computer Vision Foundation. https://doi.org/10.1109/CVPR.2016.90.

Huang, J., Chen, B., Yao, B., & He, W. (2019). ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. *IEEE Access, 7*, 92871-92880. https://doi.org/10.1109/ACCESS.2019.2928017.

Hussain, T., Muhammad, K., Ullah, A., Cao, Z., Baik, S. W., & De Albuquerque, V. H. C. (2019). Cloud-assisted multiview video summarization using CNN and bidirectional LSTM. *IEEE Transactions on Industrial Informatics, 16*(1), 77-86. https://doi.org/10.1109/TII.2019.2929228.

Jiang, S. (2019). Memento: An emotion-driven lifelogging system with wearables. *ACM Transactions on Sensor Networks (TOSN), 15*(1), 1-23. https://doi.org/10.1145/3281630.

Karim, F., Majumdar, S., & Darabi, H. (2019). Insights into LSTM fully convolutional networks for time series classi_cation. *IEEE Access, 7*, 7718-67725. https://doi.org/10.1109/ACCESS.2019.2916828.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access, 7*, 117327-117345. https://doi.org/10.1109/ACCESS.2019.2936124

Khamparia, A., Gupta, G., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound classi_cation using convolutional neural network and tensor deep stacking network. *IEEE Access, 7*, 7717-7727. https://doi.org/10.1109/ACCESS.2018.2888882.

Khan, S. U., Haq, I. U., Rho, S., Baik, S. W., & Lee, M. Y. (2019). Cover the violence: A novel Deep-Learning-Based approach towards violence detection in movies. *Applied Sciences, 9*(22), Article 4963. https://doi.org/10.3390/app9224963.

Kishore, P. V. V., & Prasad, M. V. D. (2016). Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networ. *International Journal of Software Engineering and its Applications, 10*(2), 149-170. https://doi.org/10.1109/IACC.2016.71

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105. https://doi.org/10.1145/3065386.

Kumar, K. V. V., Kishore, P. V. V., & Kumar, D. A. (2017). Indian classical dance classification with adaboost multiclass classifier on multi feature fusion. *Mathematical Problems in Engineering, 20*(5), 126-139. https://doi.org/10.1155/2017/6204742.

Liu, B., Qin, H., Gong, Y., Ge, M., Xia, W., & Shi, L. (2018). EERA-ASR: An energy-efficient reconfigurable architecture for automatic speech recognitionbwith hybrid DNN and approximate computing. *IEEE Access, 6*, 52227-52237. https://doi.org/10.1109/ACCESS.2018.2870273.

Liu, Z. T., Wu, M., Cao, W. H., Mao, J. W., Xu, J. P., & Tan, G. Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing, 273*, 271-280. https://doi.org/10.1016/j.neucom.2017.07.050.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2018, September 2-6). Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. In *Interspeech* (pp. 3683-3687). Hyderabad, India. https://doi.org/10.21437/Interspeech.2018-2228.

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 35-42). Association for Computing Machinery. https://doi.org/10.1145/2988257.2988267.

Mannepalli, K., Sastry, P. N., & Suman, M. (2016a). FDBN: Design and development of fractional deep belief networks for speaker emotion recognition. *International Journal of Speech Technology, 19*(4), 779-790. https://doi.org/10.1007/s10772-016-9368-y

Mannepalli, K., Sastry, P. N., & Suman, M. (2016b). MFCC-GMM based accent recognition system for Telugu speech signals. *International Journal of Speech Technology, 19*(1), 87-93. https://doi.org/10.1007/s10772-015-9328-y

Mustaqeem, & Kwon, S. (2020). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors, 20*(1), Article 183. https://doi.org/10.3390/s20010183.

Navyasri, M., RajeswarRao, R., DaveeduRaju, A., & Ramakrishnamurthy, M. (2017). Robust features for emotion recognition from speech by using Gaussian mixture model classification. In *International Conference on Information and Communication Technology for Intelligent Systems* (pp. 437-444). Springer. https://doi.org/10.1007/978-3-319-63645-0_50.

Ocquaye, E. N. N., Mao, Q., Song, H., Xu, G., & Xue, Y. (2019). Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition. *IEEE Access, 7*, 93847-93857. https://doi.org/10.1109/ACCESS.2019.2924597.

Rao, G. A., & Kishore, P. V. V. (2016). Sign language recognition system simulated for video captured with smart phone front camera. *International Journal of Electrical and Computer Engineering, 6*(5), 2176-2187. https://doi.org/10.11591/ijece.v6i5.11384

Rao, G. A., Syamala, K., Kishore, P. V. V., & Sastry, A. S. C. S. (2018). Deep convolutional neural networks for sign language recognition. *International Journal of Engineering and Technology (UAE), 7*(Special Issue 5), 62-70. https://doi.org/10.1109/SPACES.2018.8316344

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4580-4584). IEEE Conference Publication. https://doi.org/10.1109/ICASSP.2015.7178838.

Sastry, A. S. C. S., Kishore, P. V. V., Prasad, C. R., & Prasad, M. V. D. (2016). Denoising ultrasound medical images: A block based hard and soft thresholding in wavelet domain. *International Journal of Measurement Technologies and Instrumentation Engineering (IJMTIE), 5*(1), 1-14. https://doi.org/10.4018/IJMTIE.2015010101

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673-2681. https://doi.org/10.1109/78.650093.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1301-1309. https://doi.org/10.1109/JSTSP.2017.2764438.

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5089-5093). IEEE Conference Publication. https://doi.org/10.1109/ICASSP.2018.8462677.

Wang, H., Zhang, Q., Wu, J., Pan, S., & Chen, Y. (2018). Time series feature learning with labeled and unlabeled data. *Pattern Recognition, 89*, 55-66. https://doi.org/10.1016/j.patcog.2018.12.026

Xie, Y., Liang, R., Tao, H., Zhu, Y., & Zhao, L. (2018). Convolutional bidirectional long short-term memory for deception detection with acoustic features. *IEEE Access, 6*, 76527-76534. https://doi.org/10.1109/ACCESS.2018.2882917.

Zeng, M., & Xiao, N. (2019). Effective combination of DenseNet and BiLSTM for keyword spotting. *IEEE Access, 7*, 10767-10775. https://doi.org/10.1109/ACCESS.2019.2891838.

Zhang, A., Zhu, W., & Li, J. (2019). Spiking echo state convolutional neural network for robust time series classi_cation. *IEEE Access, 7*, 4927-4935. https://doi.org/10.1109/ACCESS.2018.2887354.

Zhang, S., Zhang, S., Huang, T., & Gao, W. (2018). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia, 20*(6), 1576-1590. https://doi.org/10.1109/TMM.2017.2766843.