Pertanika J. Sci. & Technol. 29 (4): 2967 - 2992 (2021)

SCIENCE & TECHNOLOGY

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

Enhanced Deep Hierarchical Long Short-Term Memory and Bidirectional Long Short-Term Memory for Tamil Emotional Speech Recognition using Data Augmentation and Spatial Features

Bennilo Fernandes* and Kasiprasad Mannepalli

Department of ECE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, 520002 India

ABSTRACT

Neural networks have become increasingly popular for language modelling and within these large and deep models, overfitting, and gradient remains an important problem that heavily influences the model performance. As long short-term memory (LSTM) and bidirectional long short-term memory (BILSTM) individually solve long-term dependencies in sequential data, the combination of both LSTM and BILSTM in hierarchical gives added reliability to minimise the gradient, overfitting, and long learning issues. Hence, this paper presents four different architectures such as the Enhanced Deep Hierarchal LSTM & BILSTM (EDHLB), EDHBL, EDHLL & EDHBB has been developed. The experimental evaluation of a deep hierarchical network with spatial and temporal features selects good results for four different models. The average accuracy of EDHLB is 92.12%, EDHBL is 93.13, EDHLL is 94.14% & EDHBB is 93.19% and the accuracy level obtained for the basic models such as the LSTM, which is 74% and BILSTM, which is 77%. By evaluating all the models, EDHBL performs better than other models, with an average efficiency of 94.14% and a good accuracy rate of 95.7%. Moreover, the accuracy for the collected Tamil emotional dataset, such as happiness, fear, anger, sadness, and neutral emotions indicates 100% accuracy in a cross-fold matrix. Emotions such as disgust show around 80% efficiency. Lastly, boredom shows 75% accuracy. Moreover, the training time and evaluation time utilised by EDHBL is less when compared with the other models. Therefore, the experimental analysis shows EDHBL as superior to the other models on the collected Tamil emotional dataset. When compared with the basic models, it has attained 20% more efficiency.

Keywords: BILSTM, data augmentation, emotional recognition, LSTM

ARTICLE INFO

Article history:
Received: 07 May 2021
Accepted: 30 July 2021
Published: 28 October 2021

DOI: https://doi.org/10.47836/pjst.29.4.39

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

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

INTRODUCTION

A majority of natural speech processing strategies, such as voice-activated methods and chatbots, call for speech to be considered suggestions. Usually, the basic treatment is to initially transform this speech type into textual content by using Automatic Speech Recognition (ASR) methods and next rub other learning operations or classification followed by the ASR content paper. Convolutional neural network (CNN), in addition to the pre-trained phrase vectors for sentence-level distinction and also accomplish state-of-the-art outcomes on several benchmarks. CNNs are used for text groups, and it has proved similar outcomes towards standard versions as bag full of words, n-grams and their term frequency-inverse document frequency (TF IDF) versions, ConvNets based on the words, and Recurrent Neural Networks (RNN).

Human-computer interactions are used to acquire additional active and personalised as computer systems develop as part of predicting the present mental status of the man speaker, aiding them in distinguishing various contextual meanings of the identical terms. ASR resolves variants in a speech from diverse people using probabilistic acoustic plus words designs, which results in speech transcriptions and makes the speaker impartial. This approach will be acceptable enough for many programs but has an undesired consequence for methods that depend on understanding the planned emotion in the speech to operate properly.

State-of-the-art ASR methods create outputs with good precision but shed a considerable quantity of information that hinted at emotions offered by speech. This specific gap has led Speech-based Emotion Recognition (SER) devices to turn into a location of fascinating exploration for several decades. Speech is among the organic methods for people to convey individual emotions.

The features of Long short-term memory (LSTM) and Bidirectional LSTM (BILSTM) have been investigated in this study for emotional voice recognition using a Tamil emotional information collection and an appropriate clustering technique. Different user-defined classification techniques are used end-to-end to identify data sets using Connectionist temporal classification (CTC). The group in emotional voice recognition and machine learning collaborated for this project. The technique provides a concise overview of RNN and its layers and the function extraction variables used. The dataset array and its specifics were then briefly identified. Finally, the analysis output and results were recorded using five different testing datasets, followed by a hypothesis and analysis compared with the other designs.

MATERIALS AND METHODS

A common SER structure operates on removing options such as spectral characteristics, pitch frequency functions, formant characteristics and effort associated capabilities coming from speech (Hochreiter & Schmidhuber, 1997; Zhou et al., 2016; Krizhevsky et al., 2012; Graves et al., 2013; Sak et al., 2014). These were observed using a distinction process to forecast a variety of instructional classes of emotion such as Bayesian Network Model (BNM), Hidden Markov Model (HMM), Support Vector Machines (SVM), Gaussian Mixture Model and the Multi Classifier Fusion. In addition, there are several methods utilised in conventional category projects (Liu et al., 2018; Cummins et al., 2017; Mustaqeem & Kwon, 2020; Mannepalli et al., 2016a; Hussain et al., 2020). Studying powerful RNN models and their novelty within the machine learning group is an extremely energetic analysis issue. RNNs resemble LSTMs and are presented as topics for some experiments and alterations during the earlier decade (Huang et al., 2019; Khan et al., 2019; Karim et al., 2019).

This specific evolution has very recently resulted in a novel structure known as Gated Recurrent Unit (GRU), which simplifies the complicated LSTM mobile layout. In addition, the Deep Learning methods used in earlier years have contributed breakthroughs in natural speech understanding (NLU) (Alías et al., 2016; Sastry et al., 2016; Mannepalli et al., 2016b). Nevertheless, the applicability of RNN is minimal based on two factors. First, the CTC technique has become beneficial in positioning among feedback and the resulting labelling that is unfamiliar (Zhang et al., 2016; Kumar et al., 2017; Li et al., 2014; Rao et al., 2018).

For any long-term dependencies in information where the gap involves the pertinent information and the location exactly where it is required is large, and RNNs have minimal consumption (Srivastava et al., 2014; Ioffe & Szegedy, 2015; Park et al., 2019). As a result, a specific criterion in RNN, the LSTM architecture networks, are released. LSTMs are usually created to focus on long-term dependencies on collected datasets. LSTM has proved in many situations to be good at speech identification responsibilities in which the specific recollection of LSTMs cells is utilised to determine longer dependencies. GRUs can also be created for long-range dependencies as it works well with sequential details as do LSTMs (Liu et al., 2014; Rao & Kishore, 2016; Schwarz et al., 2015). Deep Belief Networks (DBN) for emotional speech recognition has displayed a tremendous enhancement above basic designs which do not make use of deep learning, and that implies nonlinear with high order associations, which were much more effective when prepared for emotion speech recognition (Ravanelli et al., 2016; Kishore & Prasad, 2016; Ravanelli et al., 2017). Deep neural network extreme learning machine (DNN ELM) uses utterance level features from segment level chances distributions plus an individual hidden level neural net to recognise utterance amount feelings. However, enhancement of accuracies had been restricted. Bidirectional LSTM designs are meant to instruct the characteristic sequences. They have also accomplished an emotion recognition precision of 62.8% within the interactive emotional dyadic motion capture (IEMOCAP) dataset, a tremendous enhancement across

DNN ELM. CNNs in deep conjunction with LSTMs managed to attain good outcomes within the IEMOCAP dataset. Recently, scientists have already commenced exploring the usage of multimodal functions for emotion recognition.

Data Augmentation

In order to procedure information, waveform sound changes to spectrogram and nourishes neural networking to produce output. The standard way to do data augmentation is generally given waveform and other strategies that adjust spectrogram (Park et al., 2019). It presented a spectrogram, which can see it as a picture in which the x-axis is normally the period while the y-axis is considered the frequency. Understandably, it obtains a better training rate as it conveys information transformation among waveform information to spectrogram information and augments spectrogram information. Later SpecAugment for information augmentation were found in speech recognition (Park et al., 2019). There are three standard methods to augment information.

First, Time Warping is a unique factor that will likely be identified and warping to either right or left with a distance chosen from consistent distribution from zero on the moment warp parameter *W* of that particular series.

Second, Frequency Masking is where the frequency stations [f0, f0 + f) will be masked and f can be selected using a consistent division by zero on the frequency conceal parameter F, and also f0 might be selected through (0, ν’ f) wherein ν is the number of frequency networks.

Third, Moment Masking with t consecutive period measures [t0, t0 + t) will be masked, and t can be selected from a consistent division from zero on the moment mask parameter T, and t0 is selected from [0, τ’ t).

Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs). These are a parametric representation of the speech signal, widely used around automated speech recognition, though they have turned out to achieve success for remaining functions too; some of them are speaker identification and emotion recognition (Mannepalli et al., 2016a). It is recognised for being robust to all of the features for virtually any kind of speech activity. A Mel will be a product of measuring recognised frequency or pitch associated with an overall tone. Mapping upon the Mel-scale, which can be an adaptation on the Hertz scale for frequency to the man’s feeling of hearing, MFCCs identify a signal characterisation closer to man’s belief. They are estimated using a Mel scale filtration bank on the Fourier transform associated with a windowed signal. As a result, a Discrete Cosine Transform (DCT) converts the logarithmical spectrum directly into a cepstrum using Equation 1 given as follows:

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

Mel filtering banks are composed of overlapping triangular screens from the cut of wavelengths based on the middle wavelengths of the two adjacent filters. The air filters have been arranged linearly with spaced middle wavelengths and restored band breadth over the Mel dimensions. The logarithms have the impact of modifying multiplication to the inclusion of addition.

Spectral Centroid. The spectral centroid is a degree utilised within electronic signal processing to represent a spectrum. It recommends where the centre of mass on the spectrum is located. Perceptually, it has a strong relationship which gives a feeling of enhancement of a noise. It is used to relate the median on the spectrum, where it represents another statistic, and the gap between them is the same as the real difference between the unweighted median and mean reports. Since each of these tends to be actions of main inclination, the addition of certain circumstances that exhibit some identical behaviour. However, the regular acoustic spectra are randomly distributed, and the two parameters usually provide strong and distinct values. The analysis reveals that the mean will be a greater match compared to the median. It is estimated that since the weighted hostile is estimated with the help of Fourier transform, with the magnitudes of their weights (Equation 2):

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)\, x(n)}{\sum_{n=0}^{N-1} x(n)} \qquad [2]$$

Here x(n) belongs to the weighted frequency level or maybe magnitude of bin quantity n, and then f(n) belongs to the middle frequency of that bin.

Spectral Crest. The crest element is a variable of any wave function, such as alternating sound or current, which displays good standards' correlation with the actual value. The crest element signifies precisely how severely the peaks are represented in a waveform. Crest point one specifies the number of peaks, such as immediate current or even a square wave. Greater crest variables suggest peaks, such as audio waves generally, which have higher crest elements. The Crest factor may be the highest amplitude on the waveform split through the RMS benefit of the waveform (Alías et al., 2016). It corresponds to the ratio on the L_∞ standard on the L_2 norm on the parameters within the waveform (Equation 3):

$$C = \frac{|x_{peak}|}{x_{rms}} = \frac{\|x\|_\infty}{\|x\|_2} \qquad [3]$$

Spectral Entropy (SEN). It is a distributive type of Shannon's entropy, and it has the energy spectrum amplitude parts on the time frame sequence required for entropy analysis. It quantifies the spectral intricacy of the EEG signal. Shannon's Entropy (ShEn) would measure the collection of relational variables that change linearly, together with the logarithm belonging to the number of options. It is also a degree of information spread and is most often accustomed to evaluating the dynamic purchase. SEN is normally acquired by multiplying the strength in every frequency by the logarithm of the very same energy, so the item is multiplied by one. The SEN is provided by Equation 4,

$$SEN = \sum_{f} p_f log\left(\frac{1}{p_f}\right) \qquad [4]$$

Spectral Flatness. It is a degree utilised majorly in electronic signal processing to analyse a sound spectrum, and it is measured in decibels. It also provides the means to identify just how to tone such a sound instead of becoming noise-like parameters. The significance of the tonal within this context is within the experience of volume in peaks or maybe a resonant framework on a strength spectrum instead of the dull spectrum associated with white noise. A substantial spectral flatness suggests the spectrum has a comparable quantity of energy in most spectral bands, and this also would seem much like white noise. Therefore, the graph on the spectrum would seem to be at a perfect level and sleek. A reduced spectral flatness suggests that the spectral strength is concentrated in essentially a few of the bands, and this will usually seem like a blend of sine waves. The spectrum would seem sharply spiked. The spectral flatness is estimated by getting a ratio of the geometric norm of the power spectrum through the arithmetic median on the power spectrum (Equation 5), i.e.

$$Flatness = \frac{\exp\left(\left(\frac{1}{N}\right)\sum_{n=0}^{N-1} ln\, x(n)\right)}{\left(\frac{1}{N}\right)\sum_{n=0}^{N-1} ln\, x(n)} \qquad [5]$$

Here x(n) belongs to the magnitude of bin quantity n. It is necessary to be aware that when an individual (or more) cleans out a bin will yield a flatness of zero, and therefore the measure is most helpful when receptacles are not null.

Spectral Flux. It is the method for spectral modification among two successive frames. First, the squared distinction between the normalised magnitudes and the spectra on the two successive short-term windows is calculated: the 'ith' normalised DFT coefficient in the 'ith' frame. Then, the spectral flux will continue to be applied within the following Equation 6:

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_l}(EN_i(k) - EN_{i-1}(k))^2 \qquad [6]$$

The histograms represent the mean valuation of spectral flux progression of segments through two classes: i) speech and ii) music. It may be observed that the values of spectral flux are bigger with the speech type. It is anticipated that the local spectral adjustments are more regular using speech data due to the fast conversion with phonemes, and several of them are quasi-periodic. However, others are associated with a loud nature.

Spectral Skewness. The level of asymmetry of the frequency division of band energy talks about the spectrum skewness. A skewness worth zero will show that the spectrum band power is equally dispersed earlier and beneath the spectrum centroid frequency. Adverse skewness suggests that much more band power will occur over the centroid. Excellent skewness suggests that much more band power occurs beneath the centroid (Equation 7).

$$\tilde{\mu}_3 = \frac{[(X-\mu)^3]}{(E[X-\mu]^2)^{3/2}} \qquad [7]$$

Where μ represent the mean, the standard deviation as σ, the μ_3 moment of the third centroid, and E is the expected operator.

Spectral Slope. It is a degree of dependency on the reflectance over the wavelength. Several organic acoustic impulses hold the inclination of a lot less power during the fact and high frequencies, which this particular pitch has and is connected to the dynamics of an audio source of energy. One method is to quantify the use of linear regression on the Fourier magnitude spectrum on the signal. That will create a single amount indicating the incline on the line-of-best-fit from the spectral information. In signal processing, it is a degree of just how efficiently the spectrum of an audio sound tail from towards the high wavelengths, estimated utilising a linear regression (Equation 8).

$$S = \frac{R_{F1} - R_{F0}}{\lambda_1 - \lambda_0} \qquad [8]$$

It is where R_{F1}, R_{F0} gives the reflectance value and filters F_0, F_1 and $\lambda_{0,}$ λ_1 as central wavelength.

Dataset Collection

Mobile applications are used to monitor psychological speech signals for research and training. Both signals are mono signals with a 44KHz frequency. The data recorded has been used to achieve the prediction goal. The speech data comes from ten different male and female speakers. Every performer must repeat each expression ten times in various emotions such as anger, disgust, fear, sorrow, happiness, normal and boredom. Female and male speakers claim a sample of 1400 cognitive speech data sets. As it is the concept flow study, these samples were analysed. In addition, learners of skills have taken samples based on sentences. The samples were obtained with co-working spaces staff to assess their feelings throughout the therapy period to obtain the research objective.

A total of 350 samples were collected, and five datasets were made randomly with fifty samples in each dataset with about the same 44KHz that use the same mobile apps. As a result, these five datasets with fifty samples in each dataset were used to try and figure out how faculty members felt about functioning together. Since professional actors gathered the coaching knowledge base, the emotional appraisal database could be identified with great sensitivity and efficiency using Tamil emotional specifics as a base.

LSTM & BILSTM Network Architecture

A standard recurrent neural community (RNN) iterates the following formulae from $t = 1$ to T to compute the hidden vector sequence $h = (h_1, \ldots, h_T)$ and the paper vector sequence $y = (y_1, \ldots, y_T)$ given an input sequence $x = (x_1, \ldots, x_T)$ (Equations 9 & 10) (Mustaqeem et al. 2020).

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \qquad [9]$$

$$y_t = W_{hy}h_t + b_y \qquad [10]$$

The W requirements signify weights matrix multiplication (for example, W_{xh} is the input hidden weight matrix). At the same time, the b conditions denote bias vectors (for example, b_h is the main bias vector H) and could be the embedded level feature (Chen et al., 2015; Weninger et al., 2015; Erdogan et al., 2015; Eyben et al., 2013; Pascanu et al., 2013).

In most cases, H is an elementwise program with a hidden layer. It was learned that its Long Short-Term Memory (LSTM) architecture stores information in purpose-built

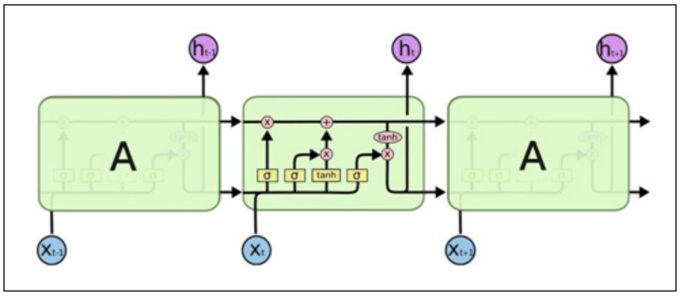


Figure 1. Sequential LSTM layer internal architecture

memory cells, which are good to find and exploit for longer range meaning. Figure 1 displays a one-of-a-kind LSTM memory cell. The subsequent recursive method is used to apply [twelve] *H* to the LSTM version added to this paper (Equations 11-15).

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \qquad [11]$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \qquad [12]$$

$$c_t = f_t c_{t-1} + i_t\, tanh\,(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \qquad [13]$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \qquad [14]$$

$$h_t = o_t \tanh(c_t) \qquad [15]$$

Where σ is the logistic sigmoid function, and *i*, *f*, o and *c* are the forms for the *input gate, forget gate, output gate and cell* activation in the vectors, are nearly the same measurements as the hidden parameter *h*. Since the weight matrices from modular to gate functions (e.g. W_{si}) are diagonal, each gate vector's component m only takes messages from object m, mostly on the cell vector.

Modern RNNs have the disadvantage of being unable to use past meaning. However, there is no excuse not to use potential meaning in speech recognition, where full statements are compiled simultaneously. Bidirectional RNNs (BRNNs) [thirteen] do this by storing data in one direction with two separate hidden rates and then feeding it forward on the same paper sheet. For example, a BRNN evaluates the *forward* concealed sequence $\vec{h}$, the *backward* hidden sequence $\overleftarrow{h}$, and the document variablev *y* by repeating the reversible level from *t* = *T* to 1, the forward rate from *t* = 1 to *T*, and afterwards modifying the performance sheet, as shown in Figure 2 (Equations 16-18):

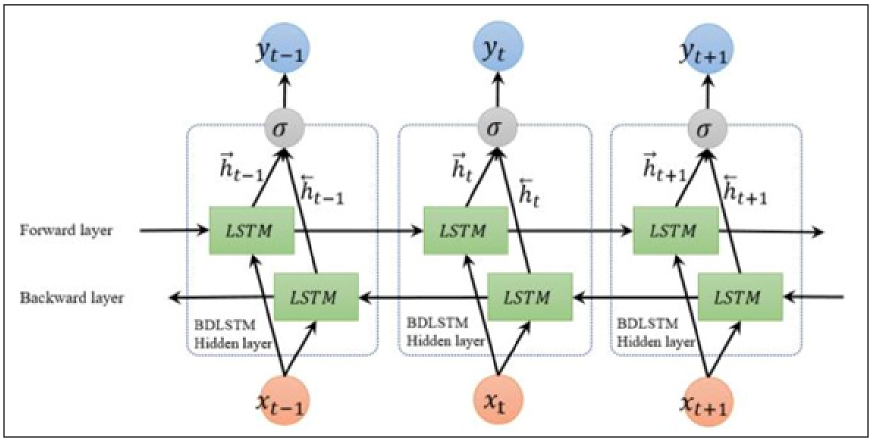


Figure 2. BILSTM layer internal architecture

$$\vec{h} = H\ (W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}) \qquad [16]$$

$$\overleftarrow{h} = H\ \left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \qquad [17]$$

$$y_t = \ W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \qquad [18]$$

When BRNNs are combined with LSTM, bidirectional LSTM is formed, using a long-range background for both entryways. The use of serious structures, capable of building gradually better level depictions of acoustic knowledge, is a key element of the most recent hybrid approach. Deep RNNs can be created by piling several RNN hidden stages on top of each other, with such a single-level text series that serves as the time step only for the arriving, as shown in Figure 3. Assume that the very same hidden state feature is used in almost all N instances in the memory, *n* = 1 to *N* and *t* = 1 to *T* were used to compute the sequence of hidden vectors (Equation 19).

$$h_t^n = H\ (\ W_{h^{n-1}h^n}h_t^{n-1} + W_{h^nh^n}h_{t-1}^n + b_h^n \qquad [19]$$

Then we calculate $h^0 = x$. The y_t machine output is in Equation 20.

$$y_t = \ W_{h^Ny}h_t^N + b_y \qquad [20]$$

Proposed Enhanced Deep Hierarchal Architecture

The dataset includes 1400 utterances delivered by ten male and ten female performers. All of them are intended to express various emotions such as neutral tone, happiness, sadness, anger, boredom, fear and disgust. The input speech data is given to data augmentation, and its value is fixed as 10 so that single data is converted into ten samples. Additionally, this evaluation value takes a long time, and its space occupancy is very high. Therefore, the feature extractions were selected among many spatial features such as Spectral Centroid, Spectral Crest, Spectral Entropy, Spectral Flatness, Spectral Flux, Spectral Skewness, Spectral Slope and Temporal Feature MFCC were utilised to extract the features from augmented data.

Then the data from feature extraction were converted into sequential data and passed into LSTM or BILSTM (where one will be selected–either LSTM or BILSTM). The layers and features were analysed, and then it is passed to the next LSTM layer with the introduction of the dropout layer, where gradient and overfitting of feature extraction can be minimised. Again, the features were analysed in LSTM or BILSTM (where one will be selected) and the layers and will be passed on to a fully connected layer with the addition of a second dropout layer, as shown in Figure 3. Thus, the fully connected layer connects all the nodes and passes the data to the soft matrix and classification layers to classify the types of emotions for the testing data.

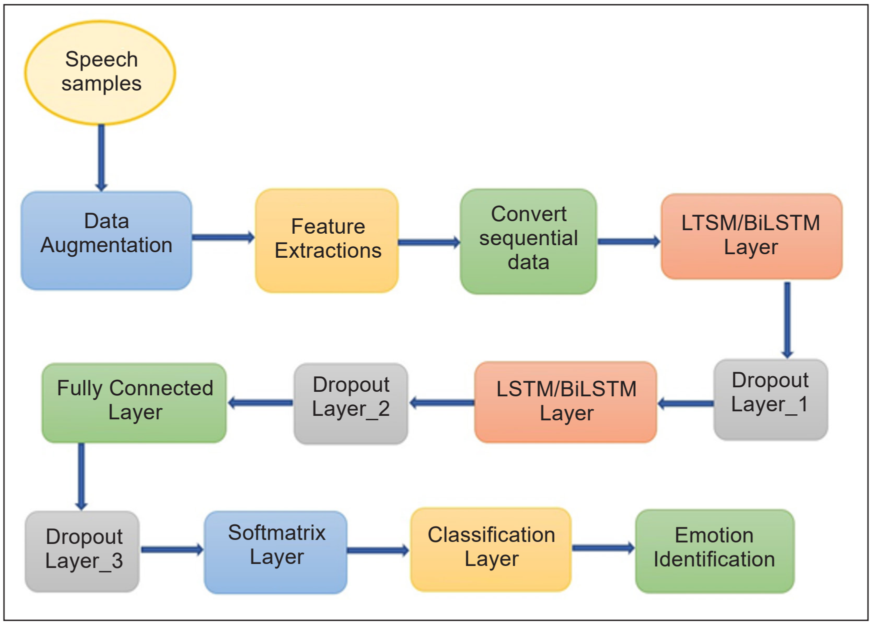


Figure 3. Proposed design flow architecture

Among other spectral features, these features were examined individually and finally concatenated by normalising the features. Then the mean and standard deviations were identified for the entire data samples according to each emotion. Even though augmentation gives a better result of overfitting, the dropout layer is also used to enhance the efficiency of emotional speech recognition. Thus, the overfitting can be reduced at the maximum level. Three dropout layers are used for the analysis dropout. Layer 1 is fixed with 0.5, Layer 2 is 0.6, and Layer 3 is 0.8. The range lies between 0 to 1, and where the standard level for dropout is 0.5, and by increasing the level to 0.8, overfitting will be reduced in the higher range.

The number of units for each LSTM and BILSTM is 250, with a total of 500 units utilised for this design architecture with an initial learning rate of 0.03. For the training optimisation, all three techniques were taken into consideration for better efficiency. From the analysis of Whale Optimization Algorithm (WOA), ADAM, SGDM and RMSPROP with minibatch size as 250, the drop period of learning rate is fixed as two and max epochs is taken as 10, so that WOA optimisation shows better performance more than other techniques. Thus, as shown in Figure 3, all four-design architecture experimented with these parameters for the Tamil emotional dataset. The other properties which were analysed by fixing the values for the audio dataset are follows. The pitch shifting probability is given as 0.5, with the time-shifting probability as 1, the volume control probability as 0.7, the volume gain range is -6 to 6, the time stretch probability as 0.5 and the range lies between 0 to 1, with the noise add probability as 1 and SNR range between -30 to 50.

RESULTS AND DISCUSSION

Enhanced Deep Hierarchical LSTM & LSTM (EDHLL) Architecture

As stated, within the layout flow and with the parameters for the input, emotional speech signals are prepared with the inclusion of spectral features, concatenation, and data augmentation with dropout level. The functionality of the EDHLL designs is examined to attain a conclusion that EDHLL design makes a confusion matrix with ten-fold cross-validation. Since cross folding is arbitrary, each evaluation result exhibits different precision amounts for a different dataset with a mean of five assessments viewed for precision rate.

In the assessment stage, a total of 250 emotional samples were clustered into five datasets, and 50 samples for each dataset were randomly taken in order to analyse the maximum efficiency of this architecture, such that a mean of five dataset accuracy was considered as the overall efficiency for the designed deep learning architecture.

From the five datasets analysis of ten folds cross valuation, an average value was grabbed for every fold and general accuracy, as shown in Figure 4. Thus, Fold four, nine and ten show 97.2% of reliability, Fold five, seven and eight show 93.7% of reliability.

It is shown in Figure 5, where some folds likewise show much better overall accuracy performance with around 80%, and the typical total accuracy rate obtained for the entire five datasets is 92.12%. By examining the private functionality of five datasets, the fourth dataset shows a much better recognition rate of 92.9%.

By studying the time factor, among the five datasets, the time taken for evaluation and classification training were considered and shown in Figure 5. While shooting the mean worth, it is apparent that for the instruction of EDHLL, the requirement was around 0.50 seconds of evaluation time and 4.59 minutes of training time.

By thinking about the unique performance of training and evaluation time as captured in different dataset collections, the time taken in the fourth dataset is less than 0.51 minutes and 4.4 minutes for evaluation and training.

From Figure 6, it can be understood that the accuracy level of all five datasets simulation has been established. As the cross-validation folds are arbitrary, the accuracy amount changes randomly according to the fifth dataset execution. However, in each execution, it lies in the range between 91.15% to 92.9%. Among the simulation of the fifth dataset, the fourth dataset shows greater precision and efficiency rate of 92.9% more than the outcome of another dataset.

Finally, in thinking about the precision amount of each feeling as revealed in Figure 7, it is apparent that for the fourth dataset of Tamil emotional samples, the EDHLL design provides 92.9% effectiveness. However, in the confusion matrix, emotions such as anger and fear give a higher rate of 100%.

Emotions, such as happiness and neutral tone show, 98% accuracy. Also, this model lags in other emotional states. For example, emotions of boredom and disgust lag in the

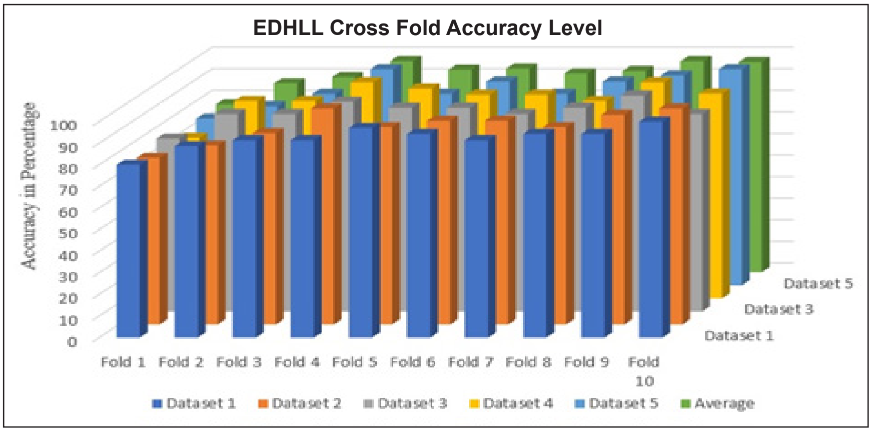


Figure 4. EDHLL cross fold output for five datasets

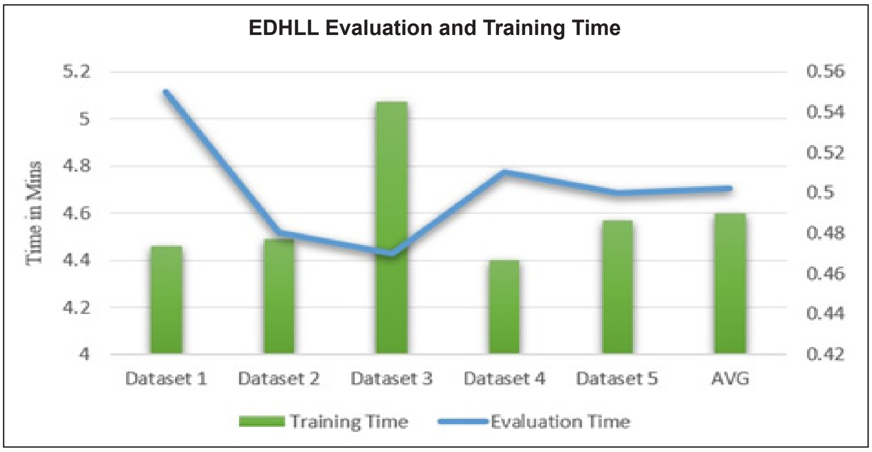


Figure 5. EDHLL evaluation time and training time for five datasets

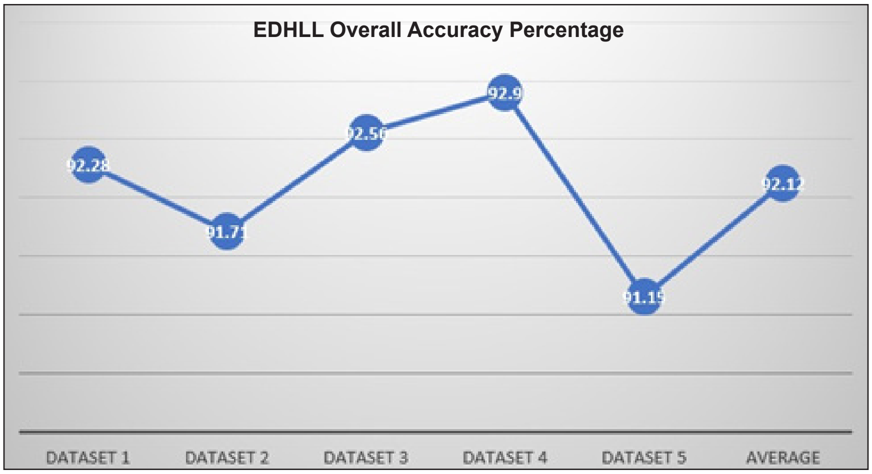
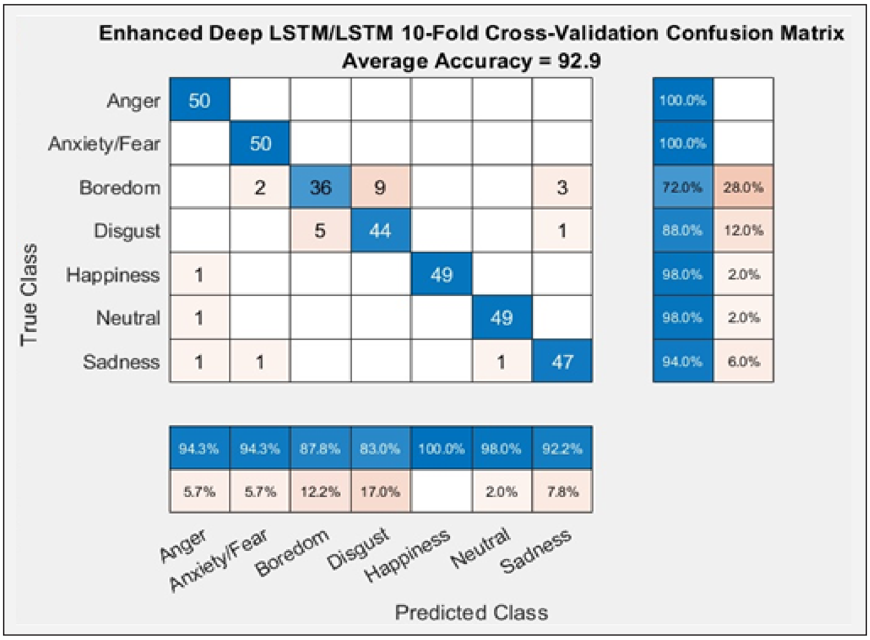


Figure 6. EDHLL overall accuracy rate for five datasets



True Class	Anger	Anxiety/Fear	Boredom	Disgust	Happiness	Neutral	Sadness		
Anger	50							100.0%	
Anxiety/Fear		50						100.0%	
Boredom		2	36	9			3	72.0%	28.0%
Disgust			5	44			1	88.0%	12.0%
Happiness	1				49			98.0%	2.0%
Neutral	1					49		98.0%	2.0%
Sadness	1	1				1	47	94.0%	6.0%
	94.3%	94.3%	87.8%	83.0%	100.0%	98.0%	92.2%		
	5.7%	5.7%	12.2%	17.0%		2.0%	7.8%		

Figure 7. Cross fold confusion matrix for EDHLL

EDHLL model. About 72% and 88% of accuracy has been obtained in both states, and it shows the lowest efficiency rate on all emotions. Overall, the accuracy rate is better than DHLL design architecture.

Enhanced Deep Hierarchical LSTM & BILSTM (EDHLB) Architecture

The functionality of the EDHLB designs was examined to attain a conclusion that EDHLB design makes a confusion matrix with ten-fold cross-validation. Since cross folding is arbitrary and each evaluation result exhibits a different precision amount for a different dataset with a mean of five assessments, this was viewed for precision rate. Again, the same five datasets used in the earlier model are used to evaluate the EDHLB design architecture in the assessment stage. Fifty samples for each dataset are randomly taken to analyse the maximum efficiency of this architecture. A mean of five dataset accuracies was considered as the overall efficiency for the designed deep learning architecture.

From the 5-dataset analysis of ten folds cross valuation, an average value was grabbed for every fold and general accuracy. Thus, fold ten shows 98.2% of reliability and also fold five, eight and nine shows 96.56% of reliability as shown in Figure 8, where other folds likewise show a better overall performance of accuracy, which stands around 80% to 93% and also the typical total accuracy rate obtained for the entire five dataset is 92.44%. By examining the private functionality of the fifth datasets, the fourth dataset shows a better recognition rate of 93.13%.

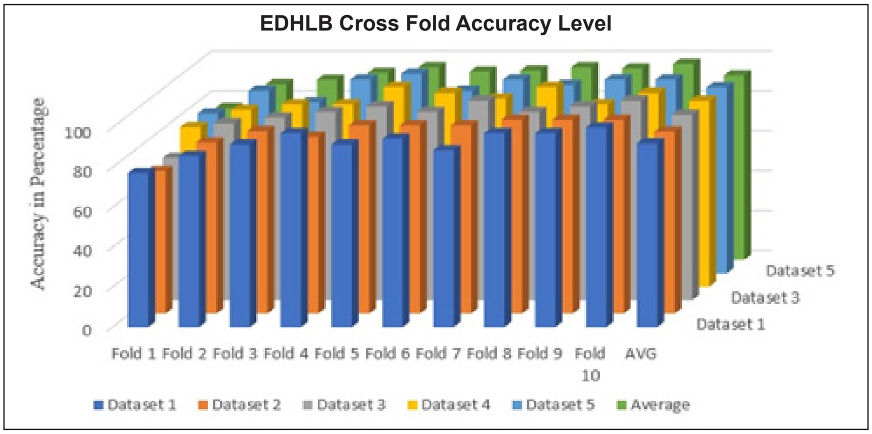


Figure 8. EDHLB cross fold output for five datasets

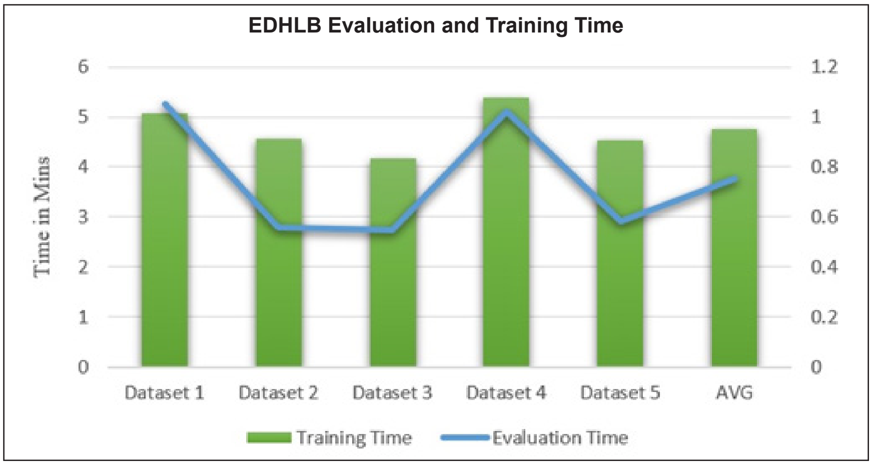


Figure 9. EDHLB evaluation time and training time for five datasets

By studying the time factor, among the five datasets, the time taken for evaluation and classification training were considered and shown in Figure 9. While shooting, the mean worth it is apparent that the instruction of EDHLB requires around 1.05 minutes of evaluation time and 5.14 minutes of training time.

Thinking of the unique performance of training and evaluation time captured in different dataset collection, the time taken for the third dataset is less than 0.55 minutes with 4.18 minutes for evaluation and training.

Figure 10 indicates that the accuracy level of all five-dataset simulations has been established. As the cross-validation folds are arbitrary, the accuracy amount changes randomly according to the dataset in five executions, but it lies between 91.13% to 93.13% in each execution. Thus, among the simulation of the fifth dataset, the fourth dataset shows greater precision and efficiency rate of 93.13% and more than the outcome of another dataset.

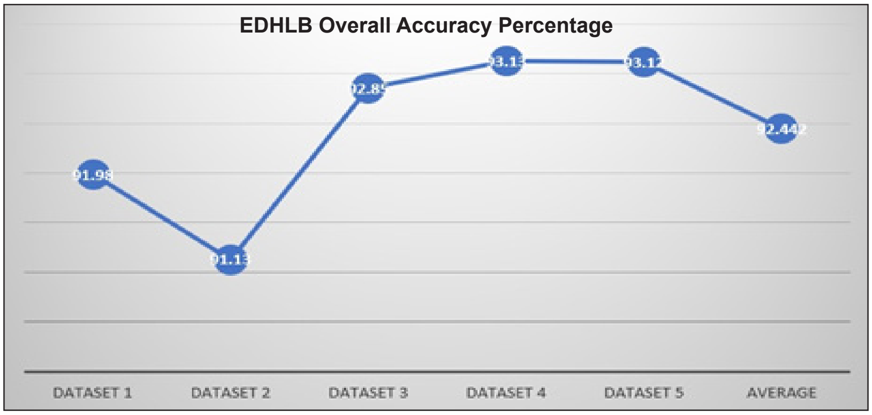


Figure 10. EDHLB overall accuracy rate for five datasets

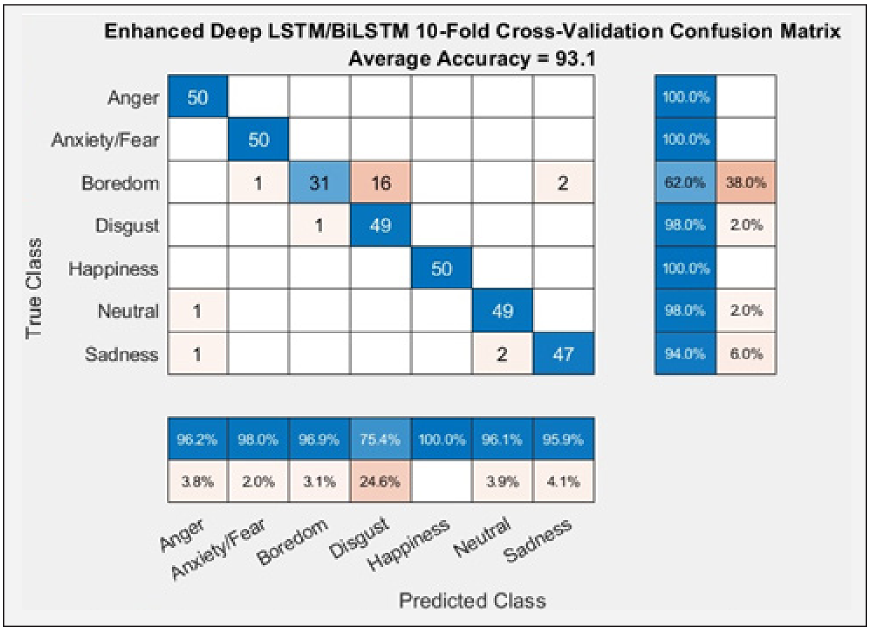


Figure 11. Cross fold confusion matrix for EDHLB

Finally, by thinking about the precision amount of each feeling as revealed in Figure 11, it is apparent that for the fourth dataset of Tamil emotional samples, the EDHLB design provides 93.1% effectiveness. In the confusion matrix, emotions such as happiness, anger and fear give a higher rate of 100%, and emotions such as disgust and neutral tone show 98% accuracy.

This model too lags with regard to other emotional states. The emotions of boredom and sadness lag in the EDHLB model, and only 62% and 94% of accuracy have been obtained in both states, and it shows the lowest efficiency rate of all emotions. Overall, the accuracy rate is better than the DHLB design architecture.

Enhanced Deep Hierarchical BILSTM & LSTM (EDHBL) Architecture

The functionality of the EDHBL designs is examined to attain a conclusion that EDHBL design makes a confusion matrix with ten-fold cross-validation. Since cross folding is arbitrary, each evaluation result exhibits different precision amounts for different datasets with a mean of five assessments, which were viewed for precision rate. Again, the same five datasets used in previous models are used to evaluate EDHBL design architecture in the assessment stage. Fifty samples for each dataset were randomly taken to analyse the maximum efficiency of this architecture. A mean of five dataset accuracy was considered the overall efficiency for the designed deep learning architecture.

From the 5-dataset analysis of ten folds cross valuation, an average value was grabbed for every fold for general accuracy. Thus, fold five, eight, nine and ten shows 98.2% of reliability and folds two, three, four and six show 94.26% of reliability as in Figure 12, where some folds likewise show a much better overall performance of accuracy around 80% to 93% and also the typical total mean accuracy rate obtained for entire five datasets is 94.15%. By examining the private functionality of the fifth dataset, the fourth datasets show a much better recognition rate of 95.7%.

In studying the time factor, among the five datasets, the time taken for evaluation and classification training was considered in Figure 13. While shooting, the mean worth it becomes apparent that the instruction of EDHBL requires around 1.38 minutes of evaluation time and 4.32 minutes of training time.

In thinking about the unique performance of training and evaluation time captured in different dataset collections, the time taken for the third dataset is as little as 1.12 minutes and 4.25 minutes for evaluation and training.

From Figure 14, there is evidence that the accuracy level of all five-dataset simulations has been established. As the cross-validation folds are arbitrary, the accuracy amount changes randomly according to the dataset in five executions. In each execution, the range is between 93.41% to 95.7%. For example, among the simulation of the fifth dataset, the fourth dataset shows a greater precision and efficiency rate of 95.7% than the outcome of the other dataset.

Finally, considering the precision amount of each feeling as revealed in Figure 15, it becomes apparent that for the fourth dataset of Tamil emotional samples, EDHBL design provides 93.1% of effectiveness. In the confusion matrix, emotions such as happiness, anger and neutral tone give a higher rate of 100%. Emotions such as fear and sadness show 98% and 94% of accuracy.

Also, this model lags in other emotional states and emotions such as boredom and disgust lag in the EDHBL model, and only 90% & 88% of accuracy is obtained in both states. It shows the lowest efficiency rate of all emotions. Overall, the accuracy rate is better than DHBL design architecture.

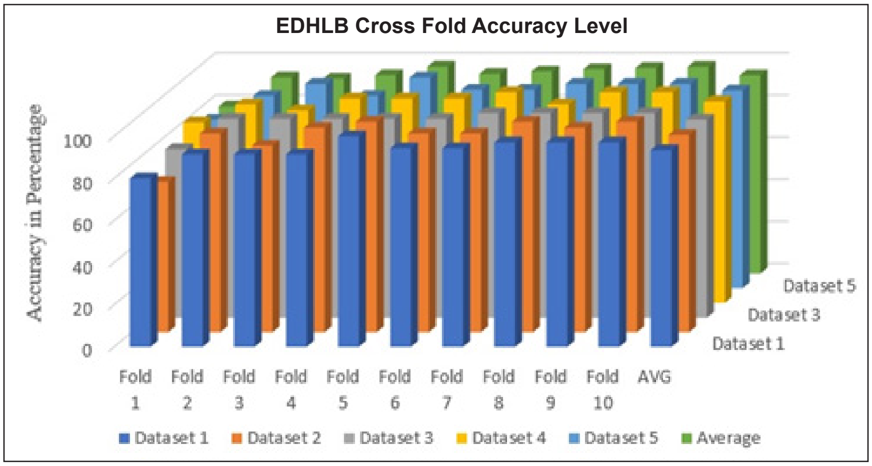


Figure 12. EDHLB cross fold output for five datasets

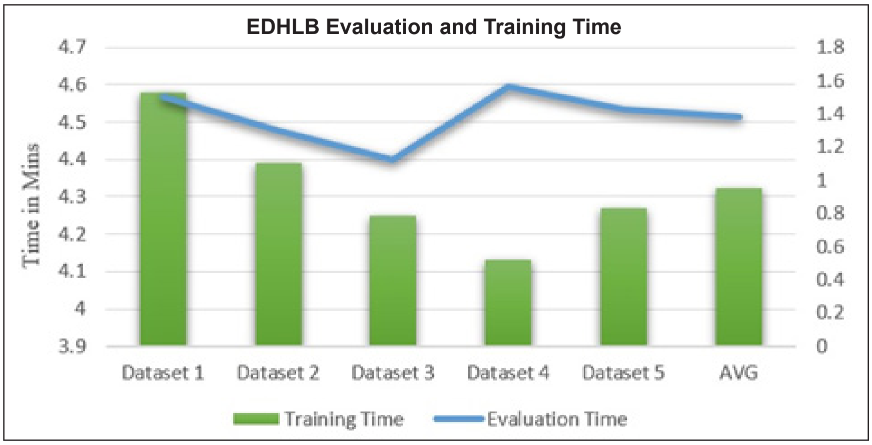


Figure 13. EDHBL evaluation time and training time for five datasets

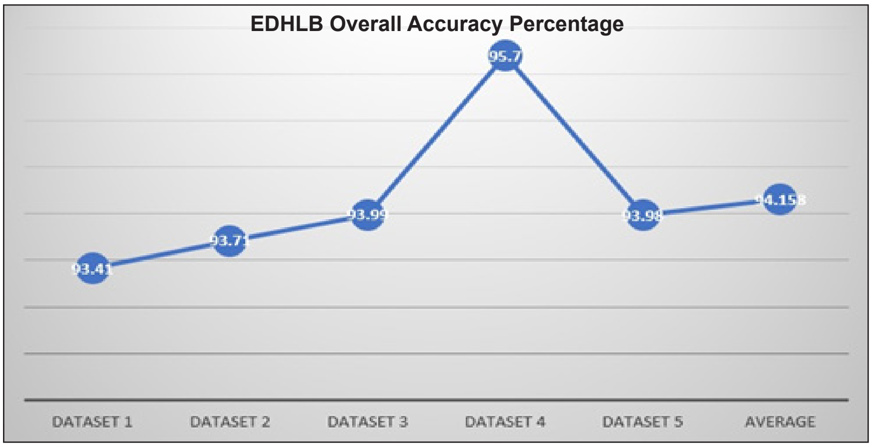
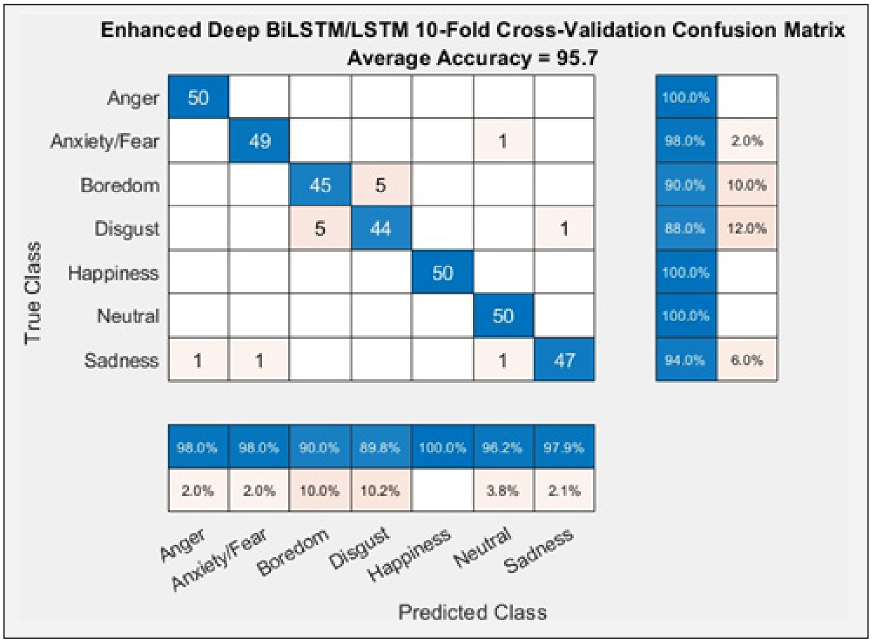


Figure 14. EDHBL overall accuracy rate for five datasets



	Anger	Anxiety/Fear	Boredom	Disgust	Happiness	Neutral	Sadness		
Anger	50							100.0%	
Anxiety/Fear		49				1		98.0%	2.0%
Boredom			45	5				90.0%	10.0%
Disgust			5	44			1	88.0%	12.0%
Happiness					50			100.0%	
Neutral						50		100.0%	
Sadness	1	1				1	47	94.0%	6.0%
	98.0%	98.0%	90.0%	89.8%	100.0%	96.2%	97.9%		
	2.0%	2.0%	10.0%	10.2%		3.8%	2.1%		

Figure 15. Cross fold confusion matrix for EDHBL

Enhanced Deep Hierarchical BILSTM & BILSTM (EDHBB) Architecture

The functionality of EDHBB designs was examined to conclude that EDHBB designs make a confusion matrix with ten-fold cross-validation. Since cross folding is arbitrary, each evaluation result exhibits different precision amounts for different datasets, a mean of five assessments was viewed for precision rate. Again, the same five datasets were used as in the previous model to evaluate EDHBB design architecture in the assessment stage. Fifty samples for each dataset were randomly taken in order to analyse the maximum efficiency of this architecture. A mean of five dataset accuracy was considered as the overall efficiency for the designed deep learning architecture.

From the 5-dataset analysis of ten folds cross valuation, an average value was grabbed for every fold and general accuracy. Thus, fold five and nine show 99.42% reliability and also folds four, seven, eight and ten show 95.4% reliability as shown in Figure 16, where some folds likewise show a better overall performance of accuracy of around 80% to 93% and also the typical total mean accuracy rate obtained for the entire five datasets is 93.19%. By examining the private functionality of the fifth dataset, the fourth dataset shows a better recognition rate of 94%.

In studying the time factor, among the five datasets, the time taken for evaluation and classification training was considered in Figure 17. While shooting, the mean worth, it is apparent that the instruction of EDHBB requires around 2.06 minutes of evaluation time and 5.17 minutes of training time.

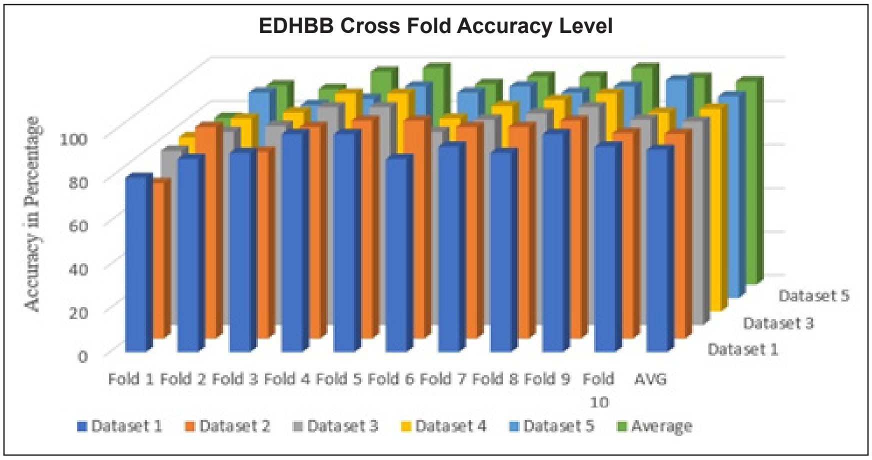


Figure 16. EDHBB cross fold output for five datasets

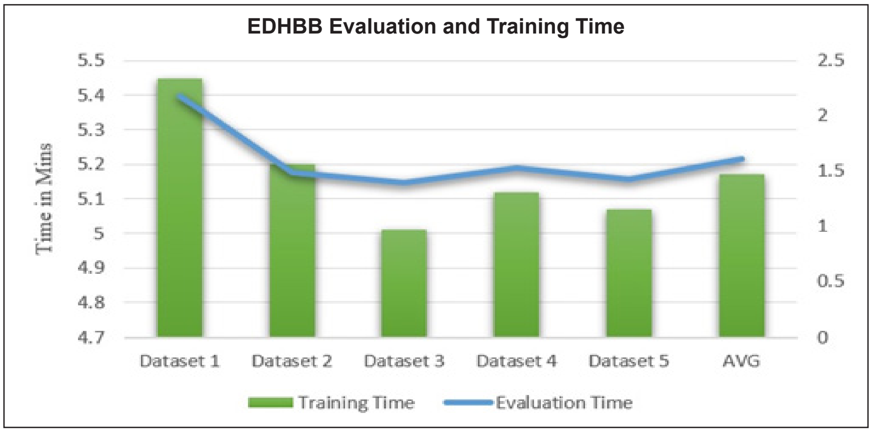


Figure 17. EDHBB evaluation time and training time for five datasets

In thinking about the unique performance of training and evaluation, the time captured in different dataset collections, the time is taken for the third dataset is as little as 1.04 minutes and 5.01 minutes for evaluation and training.

Figure 18 indicates that the accuracy level of all five-dataset simulations has been established. As the cross-validation, the folds are arbitrary; the accuracy amount changes randomly according to the dataset in 5 executions. In each execution, it lies in the range of 92.56% to 93.98%. Among the simulation of five datasets, the second dataset shows greater precision and efficiency rate of 93.98% more than the outcome of the other datasets.

Finally, thinking about the precision amount of each feeling, as revealed in Figure 19, it is apparent that for the second dataset of Tamil emotional samples, the EDHBB design provides 93.98% effectiveness. In the confusion matrix, emotions such as happiness, anger and fear give a higher rate of 100%. Moreover, emotions such as neutral tone and disgust show 96% and 94% accuracy.

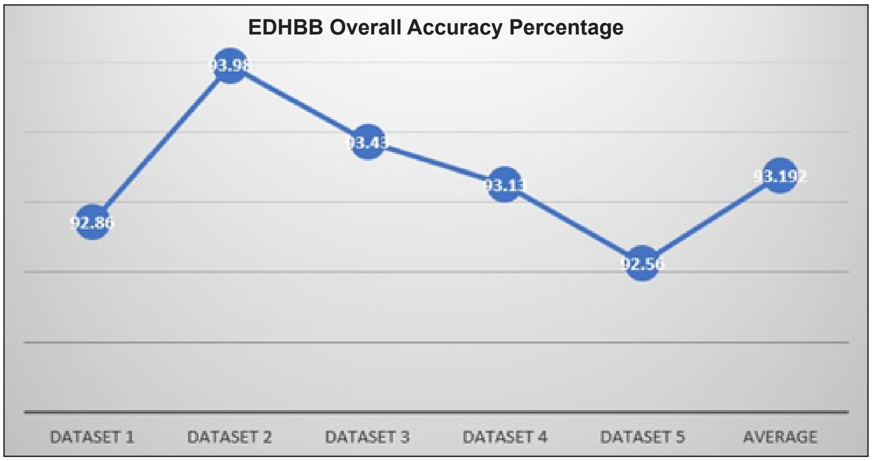


Figure 18. EDHBB overall accuracy rate for five datasets

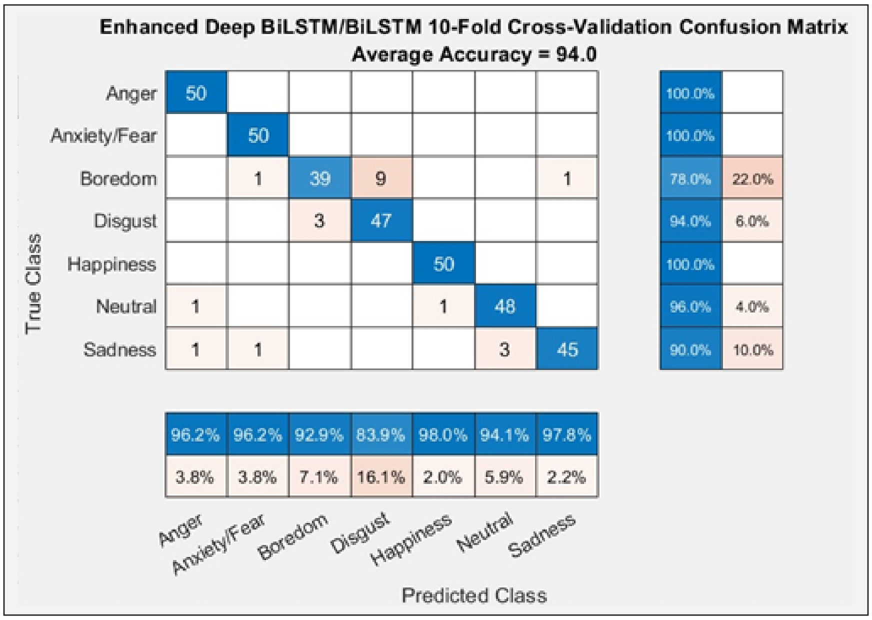


Figure 19. Cross fold confusion matrix for EDHBB

Also, this model lags in other emotional states. Emotions such as boredom and sadness lag in the EDHBL model. Only 78% and 90% of accuracy is obtained in both states and shows the lowest efficiency of all emotions. Overall, the accuracy rate is better than the DHBB design architecture.

Tables 1 and 2 show the overall performance of the complete designs. When comparing the base models of LSTM and BILSTM, EDHBL shows better performance than the other models. The EDHBB achieves comparable performance to EDHBL, and both models give average accuracy of 95.7% and 94% for the collected Tamil emotional database.

Table 1
Cross fold accuracy of EDH LL/LB/BL/BB Layers

Fold Accuracy/ Methodology	**LSTM**	**BILSTM**	**EDHLL**	**EDHLB**	**EDHBL**	**EDHBB**
Fold 1	69.2	71.4	74.3	80	85.7	**71.4**
Fold 2	72.1	74.9	91.4	88.6	94.3	**97.1**
Fold 3	72.1	74.7	91.4	91.4	91.4	**85.7**
Fold 4	75.5	73.8	100	91.4	97.1	**97.1**
Fold 5	73.9	79.6	97.1	100	97.1	**100**
Fold 6	72.1	74.9	94.3	97.1	97.1	**100**
Fold 7	73.9	79.6	94.3	94.3	100	**97.1**
Fold 8	75.5	79.2	91.4	100	94.3	**97.1**
Fold 9	75.5	76.1	100	91.4	100	**100**
Fold 10	76.8	79.6	94.8	97.1	100	**94.3**

Table 2
Overall performance of EDH LL/LB/BL/BB Layers

Overall Performance (Among 5 dataset)	**LSTM**	**BILSTM**	**EDHLL**	**EDHLB**	**EDHBL**	**EDHBB**
Best Accuracy	74	77	92.9	93.13	95.7	**94**
Average accuracy	73	76	92.12	92.44	94.1	**93.19**
Best Evaluation Time	0.45	0.51	0.47	0.55	1.12	**1.4**
Average Evaluation Time	0.56	0.59	0.5	1.05	1.38	**2.06**
Best Training Time	4.5	5.11	4.4	4.18	4.13	**5.01**
Average Training Time	5.08	5.27	4.59	5.04	4.32	**5.17**

When comparing the training time to identify the different emotional classifications for the input of 50 samples, it was learnt that the training time and the evaluation time also varied in each dataset. However, only seconds of variation can be identified. EDHBL shows a lower training time than the other models in the testing phase, and it takes only 4.13 minutes to complete the training. The other models have a lag in time for the training process. The average training time taken by the EDHBL model is 4.32 minutes, where other models take more than 30 seconds slightly to complete the training.

After training, the evaluation time for all models EDHBL lag when compared with the other models. Even though EDHBB shows comparable performance towards EDHLL and EDHBB, it takes additional to evaluate the testing database. The time is reduced by 50% in EDHLL. Around 1.12 minutes were taken to evaluate the database in the EDHBL model with an average time of 1.38 minutes. In contrast, EDHBB took around 1.4 minutes to complete the evaluation, and EDHLB took 0.55 minutes for evaluating the dataset. Though it lags in training time, it shows better results in training time.

Most efficiently, Enhanced Deep Hierarchical BILSTM and LSTM give better performance than the basic models of LSTM and BILSTM. For example, when comparing the cross folds from the above table in EDHBL folds 7, 9 and 10, it gives an accuracy rate of 100%, and in EDHBB also folds 5, 6 and 9 yield an accuracy of 100%. However, the EDHBL model takes slightly more time for training and is best in evaluation, whereas other techniques take a few more seconds to complete the evaluation. The results obtained from different models have been generated and presented effectively in this paper. Thus, further design layers can enhance this model and optimise its use with added computation and data.

CONCLUSION

Since the ordinary feedforward neural networks cannot deal with speech information for the maximum accuracy rate, the RNNs were exposed to grab the temporal dependencies of speech information were taken into account. RNNs cannot take care of the extended dependencies of gradient vanishing issues and also with regard to overfitting. Therefore, LSTMs and BILSTM were introduced to overcome the shortcomings of RNNs. In this study, the limitations of gradient vanishing, long term dependencies and overfitting problems have been reduced with an augmented data approach for SER. It improves the recognition accuracy and reduces the limitations when the overall model cost computation and processing time are considered.

In this paper, four new architecture designs were developed to select an efficient sequence for Tamil emotional speech: Enhanced Deep Hierarchal LSTM & BILSTM (EDHLB), Enhanced Deep Hierarchal BILSTM & LSTM (EDHBL), EDHLL and EDHBB. In enhancing the DHLL, DHLB, DHBL, and DHBB models with data augmentation and concatenation of MFCC & Spectral Feature problems such as overfitting, gradient exploding, and long-term dependencies were reduced to the maximum rate. Furthermore, training time and evaluation time were considered for experimental analysis properties such as the average accuracy rate. From the analysis, EDHBL shows better performance when compared to all other modes. The best accuracy rate is approximately 95.7%, with a minimal training time of 4.13 minutes, and the evaluation time of 1.12 minutes was obtained from EDHBL architecture. Therefore, for the collected Tamil emotional database, emotions such as anger, sadness, neutral tone, happiness, and fear show an efficiency rate of 100%. On the other hand, motions such as disgust and boredom still lag in the accuracy rate. Also, the EDHBB model indicates a result of 94%. Hence for the collected Tamil emotional dataset, the EDHBL model was considered to perform better when compared with other models with an average computation of accuracy rate of 94.1%, evaluation time of 1.38 minutes and training time of 4.32 minutes.

ACKNOWLEDGEMENT

The author thanks KL University, AP, India, for supporting this work and the faculty members who assisted in providing the data for this research.

REFERENCES

Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences, 6*(5), Article 143. https://doi.org/10.3390/app6050143.

Chen, Z., Watanabe, S., Erdogan, H., & Hershey, J. (2015). Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association* (pp. 3274-3278). IEEE Publishing. https://doi.org/10.1109/SLT.2016.7846281

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). ACM Publishing. https://doi.org/10.1145/3123266.3123371

Erdogan, H., Hershey, J. R., Watanabe, S., & Roux, J. L. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 708-712). IEEE Publishing. https://doi.org/10.1109/ICASSP.2015.7178061.

Eyben, F., Weninger, F., Squartini, S., & Schuller, B. (2013). Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 483-487). IEEE Publishing. https://doi.org/10.1109/ICASSP.2015.7178061.

Graves, A., Jaitly, N., & Mohamed. A. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 273-278). IEEE Publishing. https://doi.org/10.1109/ASRU.2013.6707742.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

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

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (pp. 448-456). MLResearchPress. https://doi.org/10.5555/3045118.3045167.

Karim, F., Majumdar, S., & Darabi, H. (2019). Insights into LSTM fully convolutional networks for time series classification. *IEEE Access, 7*, 67718-67725. https://doi.org/10.1109/ACCESS.2019.2916828

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.

Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 22*(4), 745-777. https://doi.org/10.1109/TASLP.2014.2304637

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

Liu, Y., Zhang, P., & Hain, T. (2014). Using neural network front-ends on far field multiple microphones based speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5542-5546). IEEE. https://doi.org/10.1109/ICASSP.2014.6854663.

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

Mannepalli, K., Sastry, P. N., & Suman, M. (2016b). 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

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

Park, D. S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E. D., & Le, Q.V. (2019). *SpecAugment: A simple data augmentation method for automatic speech recognition*. ArXiv Publishing.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning* (pp. 1310-1318). MLResearchPress. https://doi.org/10.5555/3042817.3043083.

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

Ravanelli, M., Brakel, P., Omologo, M., & Bengio, Y. (2016). Batch-normalized joint training for dnn-based distant speech recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 28-34). IEEE Publishing. https://doi.org/10.1109/SLT.2016.7846241.

Ravanelli, M., Brakel, P., Omologo, M., & Bengio, Y. (2017). A network of deep neural networks for distant speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4880-4884). IEEE Publishing. https://doi.org/10.1109/ICASSP.2017.7953084.

Sak, H., Senior, A. W., & Beaufays, F. (2014, September 14-18). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association* (pp. 338-342). Singapore.

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. *Medical Imaging: Concepts, Methodologies, Tools, and Applications*, 761-775. https://doi.org/10.1016/j.procs.2015.08.040

Schwarz, A., Huemmer, C., Maas, R., & Kellermann, W. (2015). Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4380-4384). IEEE Publishing. https://doi.org/10.1109/ICASSP.2015.7178798.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929-1958. https://doi.org/10.5555/2627435.2670313.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J. L., Hershey, J. R., & Schuller, B. W. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International conference on latent variable analysis and signal separation* (pp. 91-99). Springer. https://doi.org/10.1007/978-3-319-22482-4_11

Zhang, Y., Chen, G., Yu, D., Yao, K., Khudanpur, S., & Glass, J. R. (2016). Highway long short-term memory RNNS for distant speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5755-5759). IEEE Publishing. https://doi.org/10.1109/ICASSP.2016.7472780

Zhou, G., Wu, J., Zhang, C., & Zhou, Z. (2016). Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing, 13*(3), 226-234. https://doi.org/10.1007/s11633-016-1006-2.