I.J. Modern Education and Computer Science, **2022, 3, 45-63**
Published Online June 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2022.03.03

Enhanced Deep Hierarchal GRU & BILSTM using Data Augmentation and Spatial Features for Tamil Emotional Speech Recognition

J. Bennilo Fernandes
Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh.
Email: bennij05@gmail.com

Kasiprasad Mannepalli
Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh.
Email: mkasiprasad@gmail.com

Received: 07 May 2021; Revised: 11 July 2021; Accepted: 28 August 2021; Published: 08 June 2022

Abstract: The Recurrent Neural Network (RNN) is well suited for emotional speech recognition because its uses constantly time shifting property. Even though RNN gives better results GRU, LSTM and BILSTM solves the gradient problem and overfitting problem joins the path to reduces the efficiency. Hence in this paper five deep learning architecture is designed in order to overcome the major issues using data augmentation and spatial feature. Five different architectures like: Enhanced Deep Hierarchal LSTM & GRU (EDHLG), EDHBG, EDHGL, EDHGB & EDHGG are developed with dropout layers. The raw data learned from LSTM will be given as the input to GRU layer for deepest learning. Thus, the gradient problem is reduced, and accuracy of each emotion was increased. Also, to enhance the accuracy level spatial features were concatenated with MFCC. Thus, in all models, the experimental evaluation with the Tamil emotional dataset yielded the best results. EDHLG has a 93.12% accuracy, EDHGL has a 92.56 percent accuracy, EDHBG has a 95.42 percent accuracy, EDHGB has a 96 percent accuracy, and EDHGG has a 94 percent accuracy. Furthermore, the average accuracy rate of a single individual LSTM layer is 74%, while BILSTM is 77%. EDHGB outperforms almost all other systems, by an optimal system of 94.27 percent and then a maximum overall accuracy of 95.99 percent. For the Tamil emotion data, emotional states such as happy, fearful, angry, sad, and neutral have a 100% prediction accuracy, while disgust has a 94 percent efficiency rate and boredom has an 82 percent accuracy rate. Also, the training time and evaluation time utilized by EDHGB is 4.43 mins and 0.42 mins which is less when compared with other models. Hence by changing the LSTM, BILSTM and GRU layers large analysis of experiment on Tamil dataset is done and EDHGB is superior to other models, and when compared with basic models LSTM and BILSTM around 26% more efficiency is gained.

Index Terms: Data Augmentation, Spatial Features, LSTM, BILSTM, GRU, Emotion Recognition.

1. Introduction

Speech emotion recognition (SER) is indeed a rapidly growing field of study that promises to be a far more effective way to engage with computers. Audio data are used in a variety of HCI systems, including medical research, media, contact centres, sports, repellents, audio monitoring, and a variety of others. However, for an effective phone system, existing SER techniques do have a range of drawbacks, including good feature selection and competent deep learning strategies. As a result, scientists are still searching for a big option in order to pick the required elements and advance artificial intelligence (AI)-based identification techniques. Similarly, noise intrusion in a native accent can be extremely useful when using the system active learning. Scientists are now using complete learning methods to solve recognition problems such as speech recognition, facial detection, hand gestures, emotion recognition, and computer vision. The primary advantage of using abundant learning techniques is the instantaneous variety of features.

This work uses a Tamil emotional collected data and a required to set method to evaluate the characteristics of GRU, LSTM, and BILSTM for emotion natural language processing. CTC employs a variety of user-defined categorization strategies to classify sets of data from beginning to end. A project has been developed by the team in emotive voice detection and computer vision. The method gives a summary of the recurrent neural network and its levels. And then there is the application of after the parameters. After that, the information collection and thus its

characteristics being described. To use the five distinct experimental databases, the analytical outputs and conclusion were documented, tried to follow by an assumption and assessment based on the correlation of the other models.

The properties of LSTM and BILSTM were investigated in this work for emotive speech using a Tamil emotional dataset with and an appropriate two structures. The work is organized as follows: initially, it introduces the GRU, LSTM, and BILSTM layers, and then it presents integrative work in the field of expressive deep learning algorithm. Then it goes into detail regarding the extracting features parameters that have been applied, as well as the database collecting. The methods and solution architecture used in this study are then maintained. The study is again carried out, and the results are presented. After that, there will be a closing and some conversation.

2. Related Work

The SER area continues to have various issues which have been overcome by a substantial and equitable method that understands both temporal and sequential emotions cues [1,2]. RNNs have been shown to be effective in a number of speeches processing activities, including voice recognition, speech enhancement, speech separation, and speech activity detection. However, disappearing and bursting vectors will make training RNNs difficult, making it difficult to discover long-term correlations [3,4,5]. While simple cutting techniques can be used to deal with collapsing gradients, the missing regression problem necessitates the use of entire context. A common technique depends on gated RNNs, which main objective is nothing but to implement a gating system to help support the sequence of data via different step of time [6,7,8,9]. The disappearing gradient problem is solved in that same group of algorithms by designing successful routes, which enable the patterns to skip a number of sequential steps. LSTMs are perhaps the most well-known gated RNNs, with noval ultimate effectiveness in a range of simple machine education methods, also joins recognition of speech [10,11]. Memory cells in LSTMs are regulated by forget, output gates, and input [12,13,14]. Regardless of how powerful theirs is, such a complex method for selecting might easily result in an unnecessarily complex design. Machine learning feasibility, but at another hand, is a significant issue [16,17,18]. Major research activities and rnns have now been committed to the development of alternative systems. Recently, studies have discovered a range of serious neural networks (DNNs) techniques to simulate the identification of feelings in continuous speech [19,20,21,23]. The essence of these designs is fundamentally different. For example, one class developed a DNN model that recognizes significant cues through raw cd recordings [24,25,26].

3. Data Augmetation and Feature Extractions

3.1. Data Augmentation

In order to get work done, harmonic noise is converted to a spectra, and neuron architecture is nourished to obtain the output. The conventional method for data supplementation is by using a specific frequency and various spectrogram-adjustment procedures [15]. It showed a spectrum analyzer, which may be seen as a graphic with the x-axis correspond to the time and the y-axis signifying the pitch. It acquires a higher training rate because it not only transmits information transition from amplitude to spectra and moreover enhances spectrum analyzer data. Voice recognition learned later SpecAugment for knowledge enhancement. There are three statistical approaches for supplementing knowledge.

Time Warping is a distinct component that would quite certainly to be detected, as well as distorting to the corners only with a length selected out of an uniform range starting at " 0 " upon this sequence' instant shift variable W.

Frequency Masking will be done with change in frequency bands at [f0, f0 + f) and f parameter can be identified using a constant segmental system by 0 on F and f0 by (0, v' f), which gives the total number of frequency segments in the signal.

Moment Masking is a with t successive phase values [t0, t0 + t), where t is chosen even from a continuous partition from neutral on the instant filter variable T, whereas t0 is chosen from [0,'t).

3.2. MFCC

This is a dynamic characterization of a voice signal that is commonly employed with speech processing, but they have also proven to be successful for other applications, including voice authentication and emotional realisation [3]. It is known for having most rich set of attributes for nearly any type of speech activity. A Mel is the result of determining the recognised pitch or tone of a general thrust. MFCCs discover a pattern characterisation that is nearer to voice perception by identifying just on Mel-scale, it can be a response upon a Hz range for amplitude only to the person's emotion of listening. They're calculated by applying a Mel scaling filtering bank to a hamming window transmitter Fourier analysis.

$$c[n] = \sum_{m=1}^{M} s[n].e^{\frac{-j2\pi nk}{N}}, 0 \leq k \leq N-1 \quad (1)$$

As a conclusion, while utilising this equation which was represented below, DCT (Discrete Cosine Transform) translates the logarithmical wavelength straight together into frequency modulation. Mel filtration institutions are made up of spanning trapezoidal screenings formed from the frequencies of an adjacent condition's central spectra. These filters were placed in a continuous path, having dispersed mid frequencies, and recovered range width across the Mel parameters. The mathematical notation gets the function in changing multiplier by allowing latest addition to be included in them.

3.3. Spectral Centroid

The spectral midpoint is still a measure used to describe a band in digital signal synthesis. It specifies the location of another spectrum's centroid. It seems to have a powerful sensory link that creates the impression of volume increase [2]. It's being used to connect the midpoint to some other metric just on spectral region, as well as the variance behind them is essentially almost the same as the actual variance seen between overall average means and standard deviations values. Due to constant audio spectrum, the inclusion from certain situations which are to show many similar behaviors, yet due to the continuous sound emission spectrum, there are widely scattered, and the two factors generally produce strong and unique outcomes. The average will become a perfect replacement than the median, as per the research. Because the stacked signal is calculated using the Transformation function, the amplitudes of respective weight training are computed as follows:

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

Hence, x(n) refers mostly to the binary variable n's normalized data rate, or perhaps amplitude, whereas f(n) refers to the binary centred pitch.

3.4. Spectral Crest

The crest factor has become a parameter in almost any wave equation, including such oscillating voice or electricity, that shows the relationship between high quality with real values. Simple terms, the crest high visibility whether severe distinct peaks of a wave pattern are depicted. Many of the peaks, like instantaneous circuit or even an output frequency, is specified by crest position zero. Larger crest factors indicate maxima, such as stronger crested components in sound waveform. The largest intensity upon that output waveform divided via the RMS advantage of any sinusoidal waveform could be the Peak element. This would be the proportion of the L normal towards the L2 norm on sinusoidal variables:

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

3.5. Spectral Entropy

The radiation spectrum intensity components on the specified timeline series necessary for sensitivity research are part of this distributed form of Shannon's randomness. It measures that EEG signal's temporal complexity. Shannon's Entropy (ShEn) is a way to calculate a combination of linearly changing structural factors, as well as the natural logarithm of possibilities. It is indeed a measure on knowledge distribution because this is most commonly used to assess a device's variable acquisition. SEN is generally obtained by multiplying relative intensity of each frequency by the exponential from same energies, resulting in a one-to-one multiplier. Hence SEN is supplied through

$$SEN = \sum_f p_f log\left(\frac{1}{p_f}\right) \tag{4}$$

3.6. Spectral Flatness

This is a magnitude used mostly in baseband treatment to evaluate a sound frequency band and it is analysed in db. Also it gives a technique to detect how tonal wave like a voice rather than humming sound qualities. Rather than the flat spectral range connected to white noise, one importance of a tonality inside this scenario that is within the feeling of loudness in maxima although it could be a resonating structure on such a strong range. Significant visual flattening indicates that the wavelength contains a same quantity of heat throughout most wavebands, and it could appear to be white noise. As a result, the frequency band chart appears to be completely flat and elegant. Decreased spatial flattening indicates that wavelength decisions are taken in a small number of bands, and it will also normally appear as a mix of sinusoids. The wavelength will appear to have strong spikes in it. The harmonic flattening is calculated by dividing the mathematical standard of the frequency distribution by the energy spectrum's mathematical midpoint, i.e.

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

The intensity of binary characteristic n is represented by x(n). It's important to remember that cleaning out such a binary by one user (or more) would result in a flattening of nil, thus the metric is still most useful whenever enclosures aren't null.

3.7. Spectral Flux

Here the 'ith' normalised DFT component inside the 'ith' block is determined as the absolute differences between certain standardized dimensions also with wavelengths upon those two (2) subsequent quick frames. Over the next algorithm, the spectrum energy would be performed again:

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

The graphs show the average value of spectrum energy evolution of sections divided into two (2) categories: voice or musical. It can be seen that overall spectrum energy increments with voice frequency. It is expected, assuming that all localized frequency modifications utilizing audio files are much more frequent as a function of a quick auditory translation, and some of them seem to be semi cyclic. Some, here on the other hand, were connected with such a boisterous personality.

3.8. Spectral Skewness

It's the asymmetric of another channel estimation of spectral energies, that refers to the deviation of a spectral band. With a standard deviation of zero, the spectral bands energy is evenly distributed both above and below the spectral centre frequencies. Due to the negative skewness, substantially more band power will be generated above the median. The high reflect there is a lot higher spectrum energy below the centroid.

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

Where μ is the average, the variance, the 3rd centroid's 3 instant, while E is indeed the anticipated function.

3.9. Spectral Slope

This is a measure of how much the radiance varies with spectrum. Many spontaneous auditory signals have a lower power leaning during minimum and maximum frequency range, that this pitch provides and is linked towards the rhythms of auditory type of power. Another way is to measure the usage of regression analysis upon that signal Fourier spectral domain, which results in a fixed value showing the inclination on a row again from spectrometry. It is indeed a measure of how well the band of an acoustic flows from bottom to top frequencies, calculated using a regression analysis in digital logic.

$$S = \frac{R_{F1}-R_{F0}}{\lambda_1-\lambda_0} \quad (8)$$

4. RNN Networks

4.1. LSTM

The Long Short-Time Memory structure were inspired by means of a study of defect distribution in current RNNs, which was discovered that late night delays were unavailable to ensured due to backpropagated mistake that each explodes or ultimate load. An LSTM architecture is composed of system memory, which are a set of repeatedly information gain [20,23]. The blocks inside the layer may be designed in way of as a variational template of a device's memory circuits. Most have a single or multiple continuously logged in memory blocks, as well as three multiplier instruments - feedback, output, and forget about gates - that include frequent equivalents of compose, scroll via, and reboot features for those neurons. Far more specifically, the form in gate triggers the information upon this cell, the paper mostly on net is compounded by whichever of the document gates shown in Fig.1., as well as the previous cell variables were amplified mostly with forget regarding gate. Just gateways permit the network to work together along with the units. We have recently focused by analysing LSTM to live world examples pattern subsequent managing.

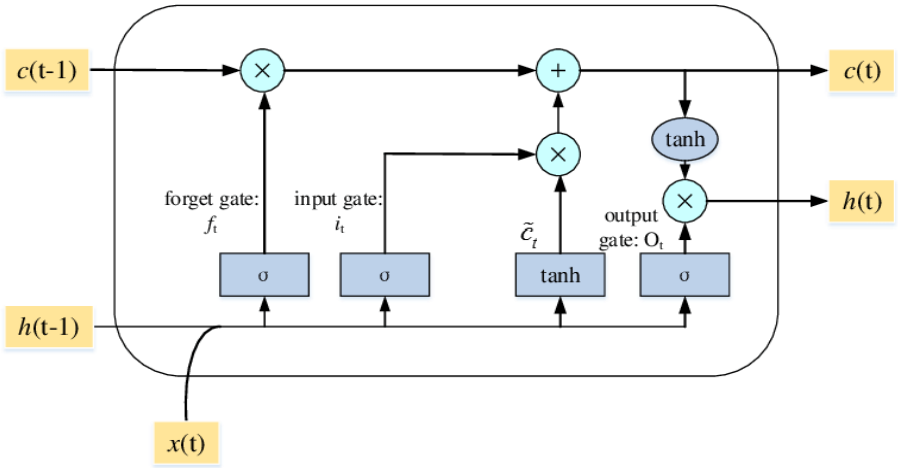

Fig.1. Sequential LSTM Layer Internal Architecture

4.2. BILSTM

Bidirectional RNNs, like dual identical RNNs, such that one travels forward and the other that travels backward, measure the combined performance from both RNNs based on a hidden layer. They just use layered idea of the LSTMs network in this document and used the two - dimensional group in our system by each both forward backward transfer. Fig.2. depicts the basic concept of the suggested fully connected bidirectional LSTM. Exterior architecture is noticed in the stated picture, in from the bidirectional learning of RNN's phase and includes all backward as well as forward transfer concealed transmit within the paper sheet [20]. Following the reporting requirement, the price and validity are calculated, and the weights and bias are modified via rear dissemination. The approach is demonstrated using 20% of the input sequence, such that it is separated from the training data, and cross entropy is used to measure the error value in the test dataset. The backward and forward passing structures are included in the complete BILSTM system, and they create a decent system for anyone to measure the contribution from the next and previous sequences with function of time since the method works across both ways.

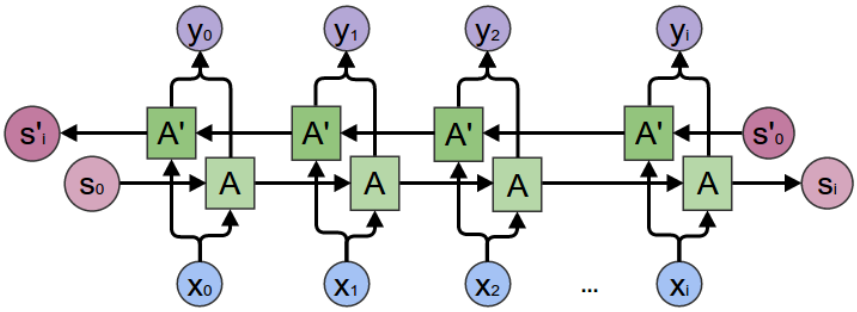

Fig.2. BILSTM Layer Internal Architecture

4.3. GRU

GRUs are indeed a form of RNN that includes thinking cells. It is also same as to the network of LSTM, but with a simplified cell construction. GRU layer also has a stacking way to ensure all information flow with cell suggestions, yet it lacks specifics and an output vector. GRU's internal theory created is depicted in Fig.3. It has made up of two gates: *r* a reset gate for finding the useful information and *z* represents an update gate [28]. Thus, upgrade level gate determines that much of the former consciousness to hold after reset gate has managed the influx of new feedback. When the GRU with LSTM is used, previous memory is a mixture of the entry and forget regarding valves, as well as the previous cell hidden state as h in Fig.3. is immediately added to this restore input gate. Additional distinction is now in the manner in which storage material is publicised. Since GRUs lack an output gate, they reveal all of their mind information, while the paper gate in LSTM handles the mind information which is used or seen by various modules in the process. In the GRU paper calculations, the following equations are used:

$$r_t = sigm\ (W_{xr}x_t + W_{hr}h_{t-1} + b_r) \tag{9}$$

$$z_t = sigm\ (W_{xz}x_t + W_{hz}h_{t-1} + b_z) \tag{10}$$

$$\tilde{h}_t = tanh\,(W_{xh}x_t + W_{hh}\,(r_t \odot h_{t-1} + b_h) \tag{11}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t)\,b_h) \odot \tilde{h}_t \tag{12}$$

x_t , h_t , r_t , z_t. signify the output vector, update gate, input vector and reset gate respectively W stands for weights vectors, and b stands for biases. The sigmoid functionality hyperbolic tangent $(tan\,h)$ and $(sigm)$ sigmoid parameters is used with same kernel function as the LSTM [27]. Because of the regulations imposed in each GRU and LSTM cell, they are suitable for long-range interactions. Despite the fact that both GRU and LSTM networks performed admirably, they were unable to determine which was preferable to another. Such studies prompted me to combine GRU and LSTM for the purpose of evaluating their efficiency throughout this survey of voice recognition analysis as demonstrated in Fig.3.

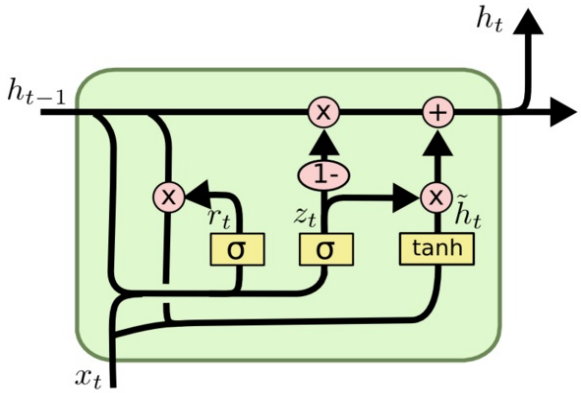

Fig.3. GRU Layer Internal Architecture

5. Database Acquisition

During study and teaching, mobile apps are applied to measure emotional speech sounds. All transmissions have a 44KHz bandwidth and seem to be mono signals. The information gathered was used to meet the forecast target. Ten separate male and female individual speakers contributed to the voice signal. So every participant had to replicate almost every emotion multiple times together in range of feelings, including normal, fear, rage, grief, happiness, disgust, and boredom. Combining female and male participants estimate a collection of 1400 cognitive and emotional speech data sources. Several slides were measured as part of the idea flowing in investigation. Examinations depending on words were performed by proficiency trainers. The samples are collected from co-working employees in the organization to measure their moods during their individual therapy and in order to achieve the study goal. A combination of 350 examples was gathered, and five databases are created at randomness, both with 50 observations and the identical 44KHz using phone applications. Therefore, these five datasets, with each 50-test data, was calculated to assess whether instructors responded regarding interacting jointly. The assessment emotional information could be detected with some more precision and effectiveness was done for Tamil emotional details, because the training information was obtained by professional actors [29,30,31].

6. Proposed Design Architecture

The database contains 1400 samples of training sets made by ten male and ten female actors, all of which are meant to represent conflicting emotions like happy, fear, normal, anger, boredom, sad and disgust. The augmentation of data is set at 10, and anything higher than that causes no progress in assessment and only time factor takes longer with more space consumption. Other features that are investigated by adjusting the settings for such acoustic sample are according continues to follow: The training pitch changing value is 0.5, the temporal changing rate is 1, the control valve of volume probability is 0.7, the amplitude boost up range is -6 to 6, the temporal lengthen rate is 0.5, the ranging from 0 to 1, and the SNR variability is -30 to 50. The most popular spectral features were utilized for the evaluation like MFCC, Spectral Skewness, Spectral Crest, Spectral Slope, Spectral centroid, Spectral Flux, Spectral Entropy and Spectral Flatness

The feature extraction for the inputs obtained from the database was transformed and processed via LSTM layer first or with BILSTM (depending on which was chosen). The different layers and characteristics have been evaluated before being transferred to another GRU layer, in which variation and underfitting of extracted features may be reduced, especially the presence of additional dropout layer. These attributes were once again evaluated in GRU or BILSTM (depending on which one was chosen), as well as the stacks were transferred onto another deep network

further with inclusion of an additional dropout layer, which is shown in Fig 3. For classifying the kinds of emotions, the deep neural network links all or most of the terminals and transmits the input towards the SoftMax layer as well as the classifier layer.

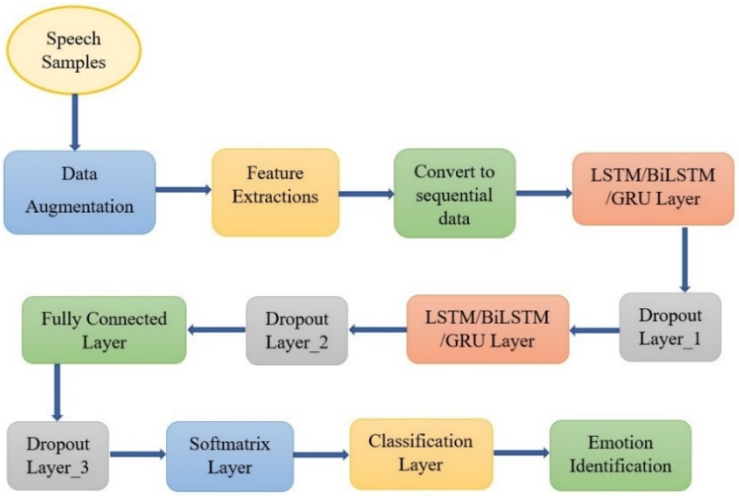

Fig.4. Proposed Design Flow Architecture

Those characteristics, as well as other features extraction, were analysed independently before being combined by standardizing the functionalities. The arithmetic mean for such complete data set will be calculated for every emotional speech data. But although augmenting improved imbalanced class performance, the dropout layer has been employed to optimize the effectiveness of emotion natural language processing, allowing performance of the classifier to be minimised towards the basic essentials, as illustrated in Fig.4. Again, for study, three dropout stages are being used: level 1 is set at 0.5, middle layer is 0.6, while access layer is 0.8. Its variation lies in the range of 0 to 1, with 0.5 being the normal threshold for dropouts or rather expanding the scope to 0.8 reducing errors in the larger level. With such an extended training level of 0.03 and 250 elements for every LSTM and GRU, a maximum of 500 units have been used for the proposed system. For enhanced performance, all selection processes have been considered while optimizing's learning. And, due to the results of the Whales Optimization Algorithm (WOA) was analyses with an algorithm to produce dimension of 250, the gradient descent dropping time is determined to 2 and also the maximum number of iterations is set to 10, indicating that WOA optimization outperforms alternative strategies. Therefore, for the collected Tamil emotion speech samples, all five-proposed architectures have been tested using these criteria.

7. Results and Discussion

7.1. Enhanced Deep Hierarchical LSTM & GRU (EDHLG) Architecture

The specifications for such raw emotional voice signal, as indicated with in design process, are created with temporal characteristics fusion and data enhancement using dropout threshold. The usefulness about the EDHLG models is investigated, and it is determined that its EDHLG architecture produces a prediction model with ten-fold confusion matrix validation set. As the 10-fold validation was obtained and it is random, every result obtained suggests a different accuracy for every data sample, an average of all five assessments had been used to calculate the test accuracy. During the evaluation step, a sample size of 250 sentimental data has been grouped into five datasets, with fifty data chosen randomly from each database and determines the design's optimal performance. Like a conclusion, the average accuracy of the built neural network architecture was based on 10-fold cross validation of 5 dataset.

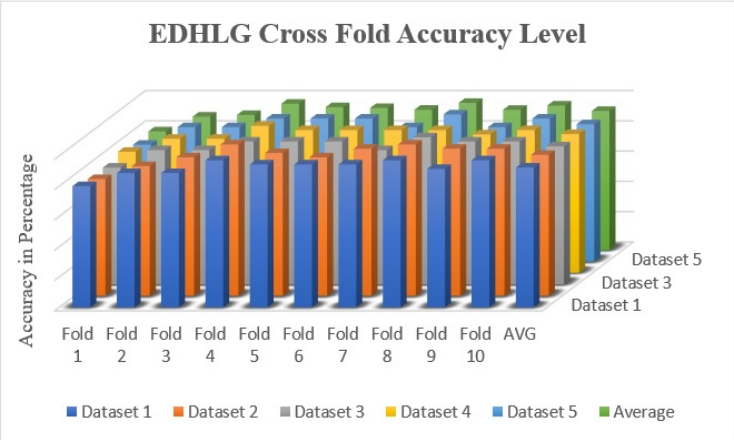

Fig.5. EDHLG Cross fold output for 5 datasets

An optimum value is determined for each cross-fold accuracy and while considering the overall accuracy from 5-dataset evaluation of with 10 folds cross evaluation. Among all folds as shown in Fig.5, folding four and eight have 97.2 percent consistency, and folding five and ten have 95.4 percent effectiveness, while most other folding have considerably significantly superior precision of approximately 88 percent, and the total mean prediction accuracy achieved for such complete five database is 91.77 percent. Thus, fourth database has a far higher recognition accuracy of 93.12 percentage when compared to the exclusive functionalities of the other five datasets.

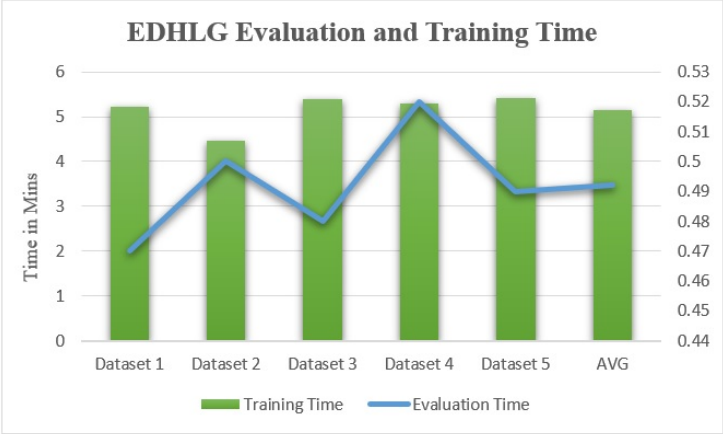

Fig.6. EDHLG Evaluation time and Training Time for 5 datasets

Moreover, by looking only at time constraint, the time being spent for evaluating and learning the classification was calculated for each of the five datasets, as shown in Fig.6. After measuring the average value, it becomes clear that process of learning EDHLG takes around 0.49 minutes of evaluation time and 5.15 minutes of training process. And, when considering the specific efficiency of training and evaluation time collected in various database collections, the time consumed in the second database becomes less as 0.47 minutes during assessment and 5.23 minutes for training.

The precision rate from all five-dataset simulations is determined by Fig.7. Because the cross-validation are randomized, the reliability value varies based on the database in each of the five runs. However, it varies within 90.58 and 93.12 percent for each iteration. The simulation outcomes of multiple datasets demonstrate that the second data source has an accuracy and performance rate of 93.12 percent, which is higher when compared to the efficiency of the other database.

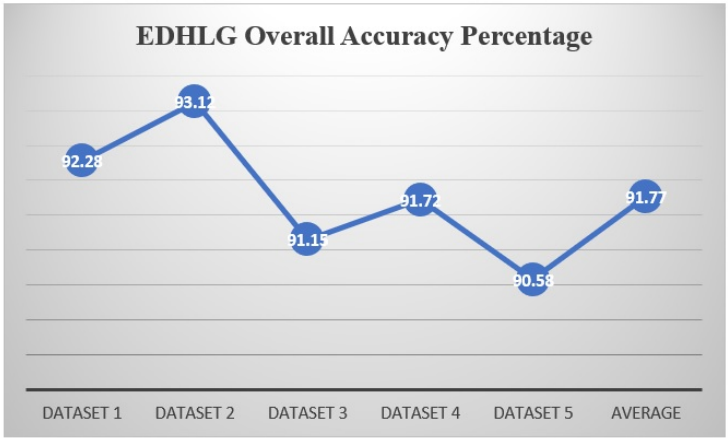

Fig.7. EDHLG overall accuracy for 5 datasets

Furthermore, considering the accuracy level of each emotion as shown in Fig.8, it is clear as using EDHLG architecture is 93.1 percent successful for the second dataset of Tamil speech samples. Fear, anger and happiness have a greater rate of 100 percent throughout the confusion matrix. Disgust and neutral emotions, for example, have a 98 percent accuracy rate. In addition, this system performs poorly in various emotional responses. In the EDHLG model, boredom and sadness emotions converge. In these conditions, just 92 percent and 66 percent precision is attained respectively, showing the least effectiveness level among all emotions. The precision rate is much higher above the DHLG system model.

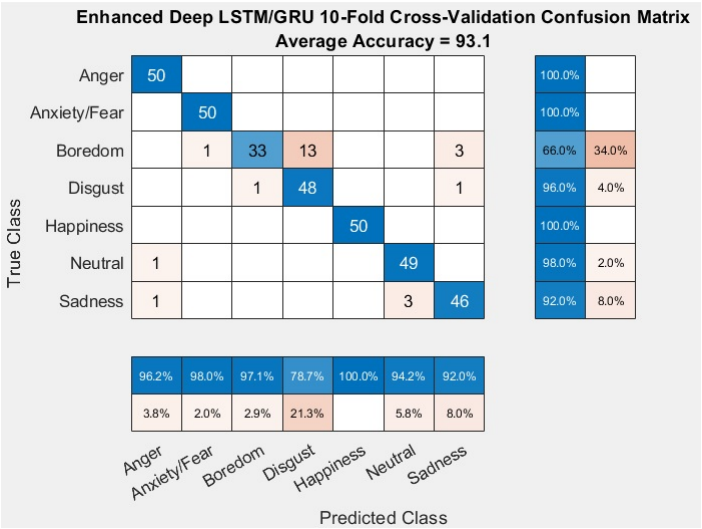

Fig.8. Cross fold confusion Matrix for EDHLG

7.2. Enhanced Deep Hierarchical BILSTM & GRU (EDHBG) Architecture

The specifications for such raw emotional voice signal, as indicated with in design process, are created with temporal characteristics fusion and data enhancement using dropout threshold. The usefulness about the EDHBG models is investigated, and it is determined that its EDHBG architecture produces a prediction model with ten-fold confusion matrix validation set. As the 10-fold validation was obtained and it is random, every result obtained suggests a different accuracy for every data sample, an average of all five assessments had been used to calculate the test accuracy. During the evaluation step, a sample size of 250 sentimental data has been grouped into five datasets, with fifty data chosen randomly from each database and determines the design's optimal performance. Like a conclusion, the average accuracy of the built neural network architecture was based on 10-fold cross validation of 5 dataset.

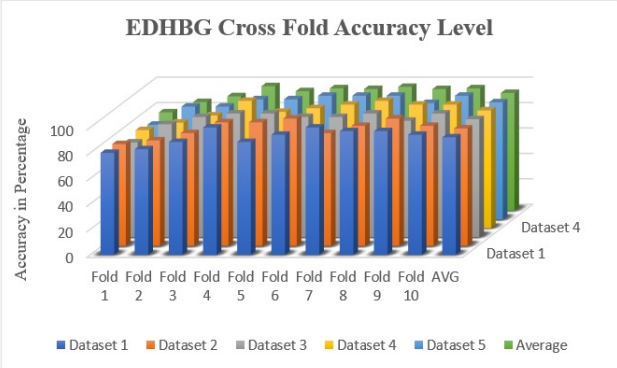

Fig.9. EDHBG Cross fold output for 5 datasets

An optimum value is determined for each cross-fold accuracy and while considering the overall accuracy from 5-dataset evaluation of with 10 folds cross evaluation. Among all folds as shown in Fig.9, folding four and eight have 97.1 percent consistency, and folding five and ten have 95.98 percent effectiveness, while most other folding have considerably significantly superior precision of approximately 80 percent, and the total mean prediction accuracy achieved for such complete five database is 93 percent. Thus, fourth database has a far higher recognition accuracy of 92.56 percentage when compared to the exclusive functionalities of the other five datasets.

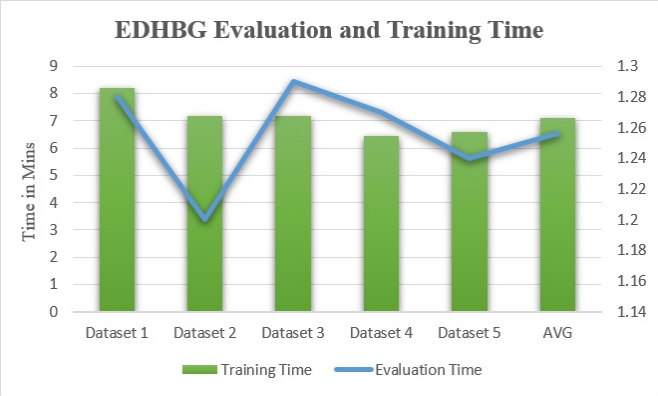

Fig.10. EDHBG Evaluation time and Training Time for 5 datasets

Moreover, by looking only at time constraint, the time being spent for evaluating and learning the classification was calculated for each of the five datasets, as shown in Fig.10. After measuring the average value, it becomes clear that process of learning EDHBG takes around 1.25 minutes of evaluation time and 7.11 minutes of training process. And, when considering the specific efficiency of training and evaluation time collected in various database collections, the time consumed in the fourth database becomes less as 1.2 minutes during assessment and 6.43 minutes for training.

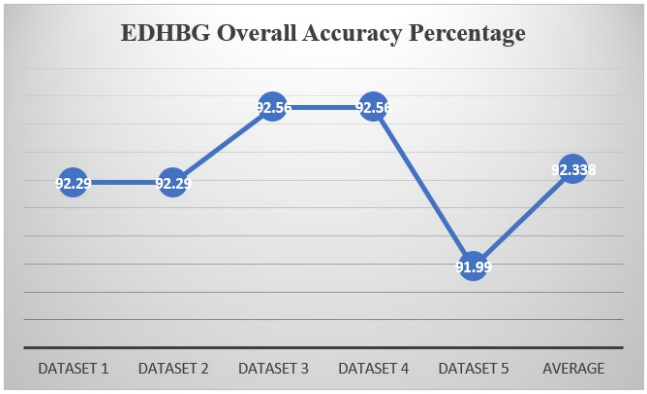

Fig.11. EDHBG overall accuracy for 5 datasets

The precision rate from all five-dataset simulations is determined by Fig.11. Because the cross-validation are randomized, the reliability value varies based on the database in each of the five runs. However, it varies within 91.99 to 92.56 percent for each iteration. The simulation outcomes of multiple datasets demonstrate that the third and fourth data source has an accuracy and performance rate of 92.6 percent, which is higher when compared to the efficiency of the other database.

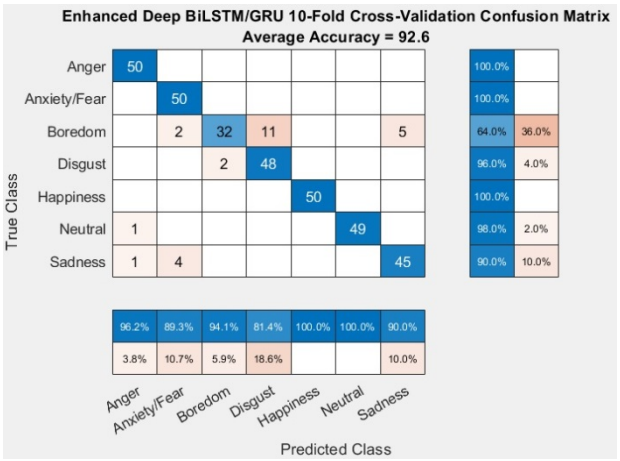

Fig.12. Cross fold confusion Matrix for EDHLBG

Furthermore, considering the accuracy level of each emotion as shown in Fig.12, it is clear as using EDHBG architecture is 92.6 percent successful for the fourth dataset of Tamil speech samples. Fear, anger and happiness have a greater rate of 100 percent throughout the confusion matrix. Disgust and neutral emotions, for example, have a 98 percent accuracy rate. In addition, this system performs poorly in various emotional responses. In the EDHBG model, boredom and sadness emotions converge. In these conditions, just 90 percent and 64 percent precision is attained respectively, showing the least effectiveness level among all emotions. The precision rate is much higher above the DHBG system model.

7.3. Enhanced Deep Hierarchical GRU & LSTM (EDHGL) Architecture

The specifications for such raw emotional voice signal, as indicated with in design process, are created with temporal characteristics fusion and data enhancement using dropout threshold. The usefulness about the EDHGL models is investigated, and it is determined that its EDHGL architecture produces a prediction model with ten-fold confusion matrix validation set. As the 10-fold validation was obtained and it is random, every result obtained suggests a different accuracy for every data sample, an average of all five assessments had been used to calculate the test accuracy. During the evaluation step, a sample size of 250 sentimental data has been grouped into five datasets, with fifty data chosen randomly from each database and determines the design's optimal performance. Like a conclusion, the average accuracy of the built neural network architecture was based on 10-fold cross validation of 5 dataset

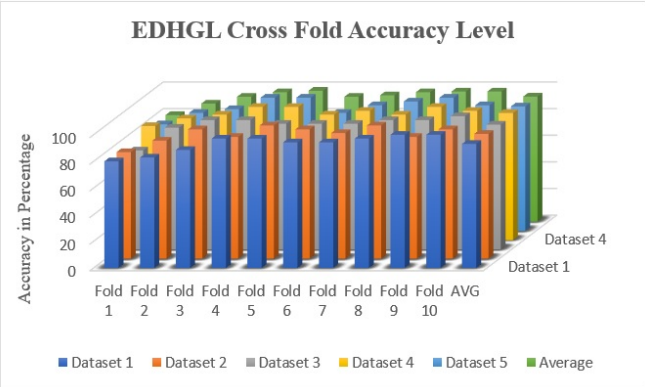

Fig.13. EDHGL Cross fold output for 5 datasets

An optimum value is determined for each cross-fold accuracy and while considering the overall accuracy from 5-dataset evaluation of with 10 folds cross evaluation. Among all folds as shown in Fig.13, folding four and eight have 97.7 percent consistency, and folding five and ten have 94.86 percent effectiveness, while most other folding have considerably significantly superior precision of approximately 80 percent, and the total mean prediction accuracy

achieved for such complete five database is 93.87 percent. Thus, fourth database has a far higher recognition accuracy of 95.2 percentage when compared to the exclusive functionalities of the other five datasets.

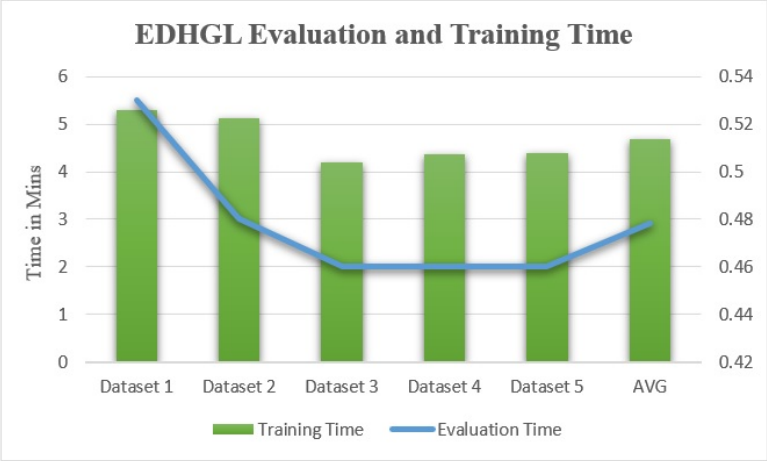

Fig.14. EDHGL Evaluation time and Training Time for 5 datasets

Moreover, by looking only at time constraint, the time being spent for evaluating and learning the classification was calculated for each of the five datasets, as shown in Fig.14. After measuring the average value, it becomes clear that process of learning EDHGL takes around 0.48 minutes of evaluation time and 4.67 minutes of training process. And, when considering the specific efficiency of training and evaluation time collected in various database collections, the time consumed in the third database becomes less as 0.46 minutes during assessment and 4.20 minutes for training.

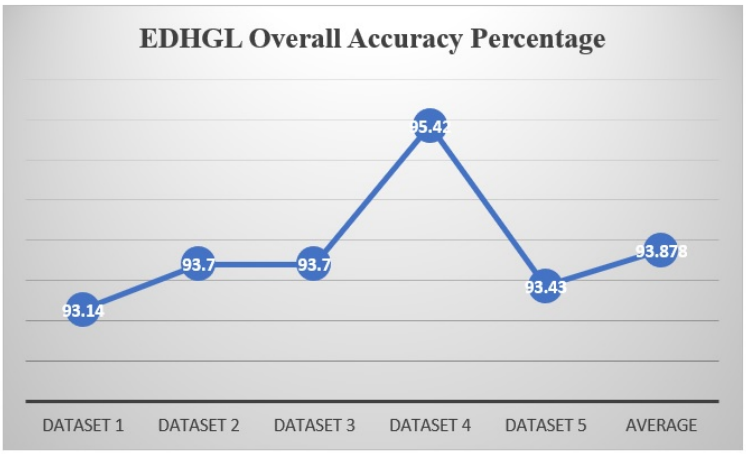

Fig.15. EDHGL overall accuracy for 5 datasets

The precision rate from all five-dataset simulations is determined by Fig.15. Because the cross-validation are randomized, the reliability value varies based on the database in each of the five runs. However, it varies within 93.14 to 95.42 percent for each iteration. The simulation outcomes of multiple datasets demonstrate that the fourth data source has an accuracy and performance rate of 95.42 percent, which is higher when compared to the efficiency of the other database.

Furthermore, considering the accuracy level of each emotion as shown in Fig.16, it is clear as using EDHGL architecture is 95.42 percent successful for the fourth dataset of Tamil speech samples. Neutral, fear and anger have a greater rate of 100 percent throughout the confusion matrix. Happiness and sadness emotions, for example, have a 98 percent accuracy rate. In addition, this system performs poorly in various emotional responses. In the EDHGL model, boredom and disgust emotions converge. In these conditions, just 90 percent and 86 percent precision is attained respectively, showing the least effectiveness level among all emotions. The precision rate is much higher above the DHGL system model.

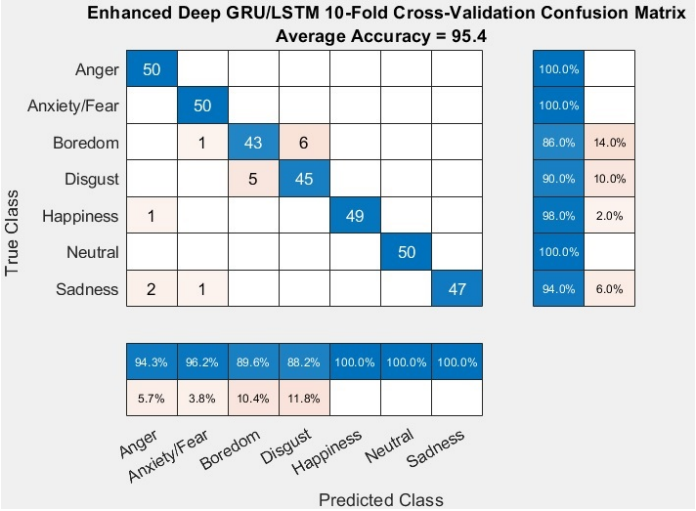

Fig.16. Cross fold confusion Matrix for EDHGL

7.4. Enhanced Deep Hierarchical GRU & BiLSTM (EDHGB) Architecture

The specifications for such raw emotional voice signal, as indicated with in design process, are created with temporal characteristics fusion and data enhancement using dropout threshold. The usefulness about the EDHGB models is investigated, and it is determined that its EDHGB architecture produces a prediction model with ten-fold confusion matrix validation set. As the 10-fold validation was obtained and it is random, every result obtained suggests a different accuracy for every data sample, an average of all five assessments had been used to calculate the test accuracy. During the evaluation step, a sample size of 250 sentimental data has been grouped into five datasets, with fifty data chosen randomly from each database and determines the design's optimal performance. Like a conclusion, the average accuracy of the built neural network architecture was based on 10-fold cross validation of 5 dataset.

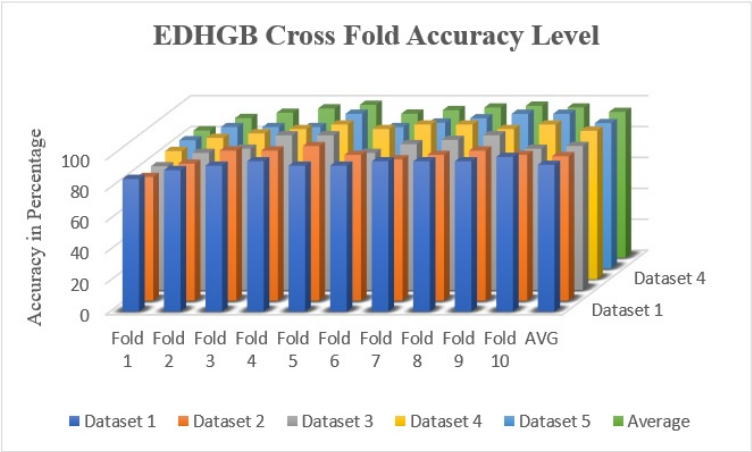

Fig.17. EDHGB Cross fold output for 5 datasets

An optimum value is determined for each cross-fold accuracy and while considering the overall accuracy from 5-dataset evaluation of with 10 folds cross evaluation. Among all folds as shown in Fig.5, folding four and eight have 98.2 percent consistency, and folding five and ten have 96.54 percent effectiveness, while most other folding have considerably significantly superior precision of approximately 82 percent, and the total mean prediction accuracy achieved for such complete five database is 94.27 percent. Thus, fourth database has a far higher recognition accuracy of 96 percentage when compared to the exclusive functionalities of the other five datasets.

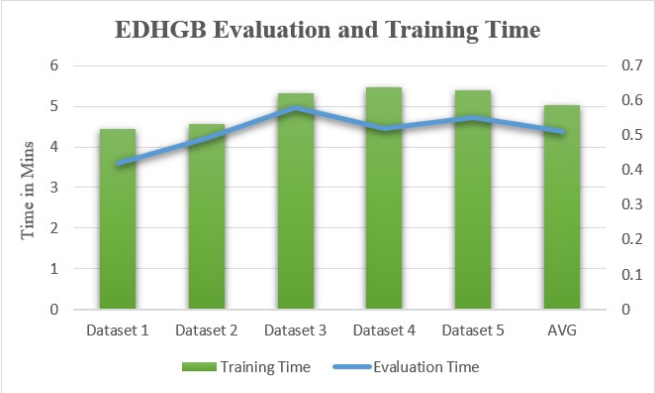

Fig.18. EDHGB Evaluation time and Training Time for 5 datasets

Moreover, by looking only at time constraint, the time being spent for evaluating and learning the classification was calculated for each of the five datasets, as shown in Fig.18. After measuring the average value, it becomes clear that process of learning EDHGB takes around 0.512 minutes of evaluation time and 5.03 minutes of training process. And, when considering the specific efficiency of training and evaluation time collected in various database collections, the time consumed in the first database becomes less as 0.42 minutes during assessment and 4.43 minutes for training.

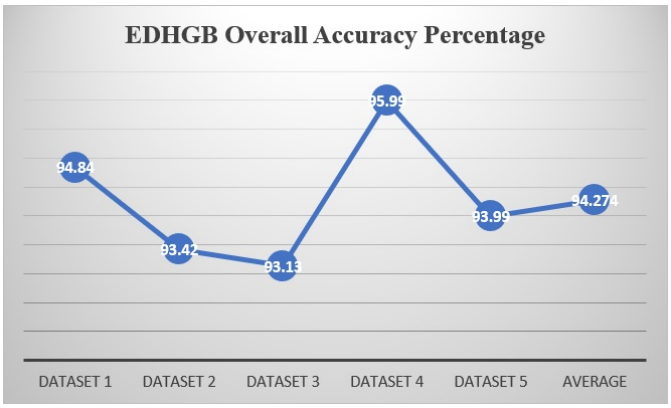

Fig.19. EDHGB overall accuracy for 5 datasets

The precision rate from all five-dataset simulations is determined by Fig.19. Because the cross-validation are randomized, the reliability value varies based on the database in each of the five runs. However, it varies within 93.13 to 95.99 percent for each iteration. The simulation outcomes of multiple datasets demonstrate that the fourth data source has an accuracy and performance rate of 96 percent, which is higher when compared to the efficiency of the other database.

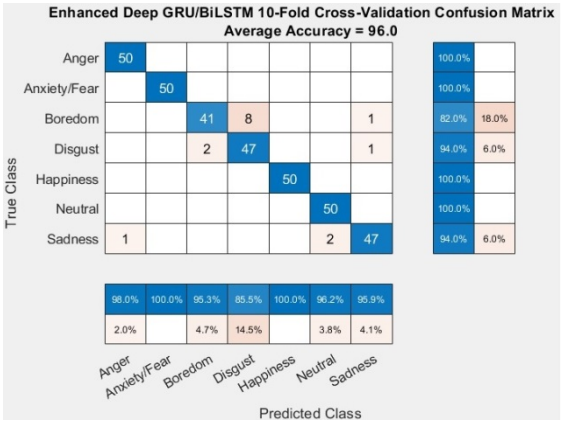

Fig.20. Cross fold confusion Matrix for EDHGB

Furthermore, considering the accuracy level of each emotion as shown in Fig.20, it is clear as using EDHGB architecture is 96 percent successful for the fourth dataset of Tamil speech samples. Neutral, happiness, anger and fear have a greater rate of 100 percent throughout the confusion matrix. Disgust and sadness emotions, for example, have a 94 percent accuracy rate. In addition, this system performs poorly in various emotional responses. In the EDHGB model, boredom emotions converge. In these conditions, just 82 percent precision is attained respectively, showing the least effectiveness level among all emotions. The precision rate is much higher above the DHGB system model.

7.5. Enhanced Deep Hierarchical GRU & GRU (EDHGG) Architecture

As the 10-fold validation was obtained and it is random, every result obtained suggests a different accuracy for every data sample, an average of all five assessments had been used to calculate the test accuracy. During the evaluation step, a sample size of 250 sentimental data has been grouped into five datasets, with fifty data chosen randomly from each database and determines the design's optimal performance. Like a conclusion, the average accuracy of the built neural network architecture was based on 10-fold cross validation of 5 dataset.

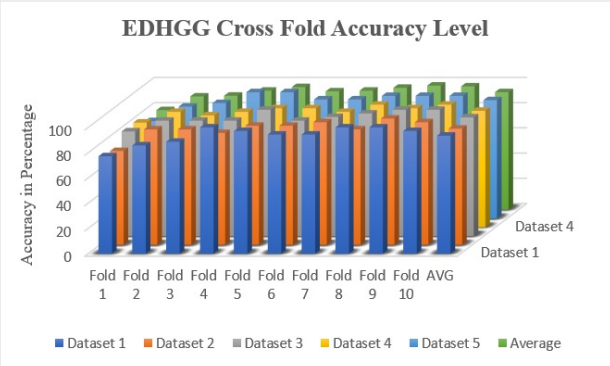

Fig.21. EDHGG Cross fold output for 5 datasets

An optimum value is determined for each cross-fold accuracy and while considering the overall accuracy from 5-dataset evaluation of with 10 folds cross evaluation. Among all folds as shown in Fig.5, folding four and eight have 98.2 percent consistency, and folding five and ten have 96.56 percent effectiveness, while most other folding have considerably significantly superior precision of approximately 78 percent, and the total mean prediction accuracy achieved for such complete five database is 93.07 percent. Thus, fourth database has a far higher recognition accuracy of 93.99 percentage when compared to the exclusive functionalities of the other five datasets.

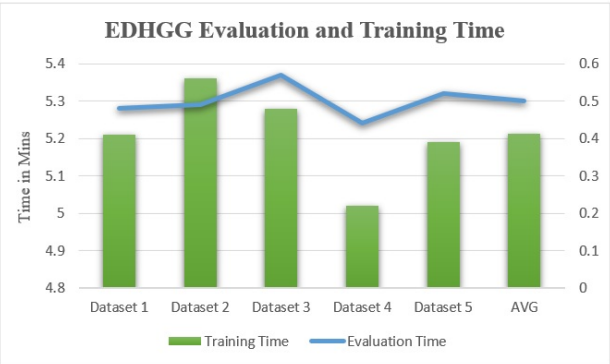

Fig.22. EDHGG Evaluation time and Training Time for 5 datasets

Moreover, by looking only at time constraint, the time being spent for evaluating and learning the classification was calculated for each of the five datasets, as shown in Fig.22. After measuring the average value, it becomes clear that process of learning EDHGG takes around 0.5 minutes of evaluation time and 5.22 minutes of training process. And, when considering the specific efficiency of training and evaluation time collected in various database collections, the time consumed in the third database becomes less as 0.50 minutes during assessment and 5.02 minutes for training.

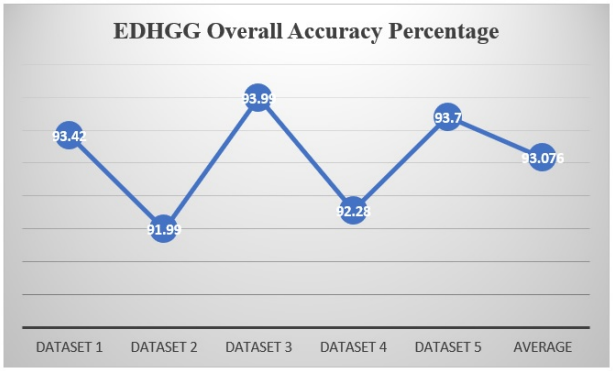

Fig.23. EDHGG overall accuracy for 5 datasets

The precision rate from all five-dataset simulations is determined by Fig.23. Because the cross-validation are randomized, the reliability value varies based on the database in each of the five runs. However, it varies within 93.13 to 95.99 percent for each iteration. The simulation outcomes of multiple datasets demonstrate that the third data source has an accuracy and performance rate of 93.99 percent, which is higher when compared to the efficiency of the other database.

Furthermore, considering the accuracy level of each emotion as shown in Fig.24, it is clear as using EDHGG architecture is 94 percent successful for the third dataset of Tamil speech samples. Happiness, fear and anger have a greater rate of 100 percent throughout the confusion matrix. Disgust, sadness, and neutral emotions, for example, have a 96 percent accuracy rate. In addition, this system performs poorly in various emotional responses. In the EDHGG model, boredom emotions converge. In these conditions, just 68 percent precision is attained respectively, showing the least effectiveness level among all emotions. The precision rate is much higher above the DHGG system model.

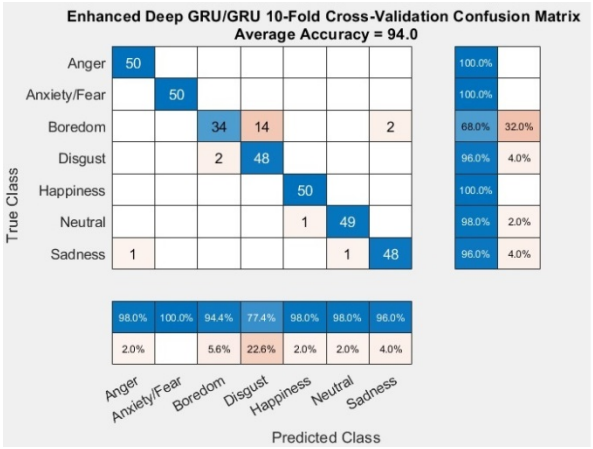

Fig.24. Cross fold confusion Matrix for EDHLG

As a result, the cumulative efficiency for all the models is shown in tables 1 and 2. When compared to all other systems, EDHGB outperforms them by 95.99 percent. EDHGL and EDHGG also behave similarly to EDHGB. In comparison to the simple LSTM and BILSTM algorithms, these systems have an overall precision of 95.4 percent and 94 percent on its gathered Tamil emotion dataset.

The evaluation is now done to detect the distinct emotional categorization with its information of 50 data by analyzing the duration for training. The learning and assessment times vary with each database, although just secs of variance could be recognized. EDHGB generally has a shorter time for training over some comparable designs in evaluation, taking only 4.43 min to finish the learning. Other systems, on the other hand, contain a temporal delay in the training phase. The overall training methodology for the EDHGB model is 5.03 minutes, whereas other systems require somewhat longer to finish the process.

Table 1. Cross Fold Accuracy of EDH LG/BG/GL/GB/GG Layers

Fold Accuracy/Methodology	LSTM	BILSTM	EDHLG	EDHBG	EDHGL	EDHGB	EDHGG
Fold 1	69.2	71.4	77.1	77.1	85.7	82.9	82.9
Fold 2	72.1	74.9	85.7	82.9	91.4	91.4	91.4
Fold 3	72.1	74.7	91.4	88.6	94.3	94.3	91.4
Fold 4	75.5	73.8	100	100	100	97.1	91.4
Fold 5	73.9	79.6	94.3	91.4	100	100	100
Fold 6	72.1	74.9	91.4	94.3	94.3	97.1	91.4
Fold 7	73.9	79.6	97.1	97.1	97.1	100	94.3
Fold 8	75.5	79.2	100	100	94.3	100	97.1
Fold 9	75.5	76.1	97.1	97.1	100	97.1	100
Fold 10	76.8	79.6	97.1	97.1	97.1	100	100

When comparing EDHGB to certain other systems during training and measuring the assessment time including all systems, EDHGB comes out on top. Even although EDHLG and EDHGG function as well against EDHGB, reviewing this data lasts forever. When we use a larger sample in EDHBG, we save more than half of the time. The databases was evaluated in roughly 0.42 minutes in the EDHGB method, using an average length of 0.51 minutes, while EDHGL requires approximately 0.48 minutes to perform the assessment and EDHGG requires 0.5 minutes. As a result, evaluating all of the EDHGB devices requires lesser time to finish the assessment.

Table 2. Overall Performance of EDH LG/BG/GL/GB/GG Models.

Overall Performance (5 Datasets)	LSTM	BILSTM	EDHLG	EDHBG	EDHGL	EDHGB	EDHGG
Best Accuracy	74	77	93.12	92.56	95.42	95.99	93.99
Average accuracy	73	76	91.77	92.33	93.87	94.27	93.07
Best Evaluation Time	0.45	0.51	0.47	1.2	0.43	0.42	0.44
Average Evaluation Time	0.56	0.59	0.49	1.25	0.48	0.51	0.5
Best Training Time	4.5	5.11	5.23	6.43	4.2	4.43	5.02
Average Training Time	5.08	5.27	5.156	7.11	5.07	5.03	5.22

Enhanced Deep Hierarchical GRU and BILSTM provide the greatest result in RNN, following by EDHGL and EDHGG, which are somewhat more productive. When analyzing the crossing folds from this list, folds 5, 7, 8, and 10 in EDHGB and EDHGL produce a 100 percent average accuracy, whereas folds 5, 9 and 10 in EDHGG likewise provide a 100 percent prediction performance. The EDHGB method obtains the greatest accuracy rate, training time, and evaluation time, whereas other strategies require just few seconds longer to finish the assessment. This study efficiently generates and presents the findings gained using various designs. As a result, more study into architectural layering may improve and maximize the system using a deal of computing and information.

8. Conclusion

The reverse relations cannot be handled by a plain feedforward neural network. They are unable to store historical data in order to make decisions in the present, which would be crucial for activities such as speech recognition. RNNs is created specifically for this. However, due to the constraints of RNNs, including the exploding gradient phenomenon and the inability to operate on long-term temporal dependencies inside the results, LSTMs, BILSTMs, and GRU are added, which overcome the drawbacks using memory cells in their design. Because of certain drawbacks, such as gradient fading, long-term dependencies, and overfitting, an augmented data solution for SER has been developed to increase classification precision while also reducing total model cost computing and processing time. This work developed five new designs to select the most successful Tamil emotive voice series, addressing issues such as

stochastic fading, long-term dependency, and underfitting. An technique for SER that uses augmented input and spectral characteristics to increase detection performance and lower overall measurement cost computation and process time. Enhanced Deep Hierarchal LSTM & GRU (EDHLG), Enhanced Deep Hierarchal BILSTM & GRU (EDHBG), EDHGL, EDHGB, and EDHGG are five novel designs designed in this study to pick a much more effective schedule for Tamil emotional utterance. Data augmentation, gradient explosion, and long-term correlations were decreased at the desired rate by strengthening the DHLL, DHLB, DHBL, & DHBB models using convolution operation and fusion of CNN & Spectral Features. Thus, for the experimental analysis properties like average accuracy rate, training time and evaluation time were taken for consideration. And from the analysis EDHGB shows best performance than other modes. EDHGB design achieved the optimum prediction reliability of around 96 percent with a training time of 4.43 minutes and an assessment time of 0.42 minutes. As a result, whereas emotions like angry, sad, neutral, joyful, and fear have a 100% prediction performance in the Tamil emotional dataset, feelings like disgust and boredom have a lower precision. In addition, the EDHGL and EDHGG designs had superior results of 95.4 percent and 94 percent, respectively. As a result, EDHGB models outperform than other models for the gathered Tamil emotional dataset, showing an overall average accuracy of 94.27 percent, assessment time of 0.51 minutes, and training time of 5.03 minutes.

References

[1] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958. 2014.

[2] Yin Win Chit, Win Ei Hlaing, Myo Myo Khaing, " Myanmar Continuous Speech Recognition System Using Convolutional Neural Network ", International Journal of Image, Graphics and Signal Processing, Vol.13, No.2, pp. 44-52, 2021.

[3] K. Mannepalli, P. N Sastry, and M. Suman, “MFCC-GMM based accent recognition system for Telugu speech signals,” *International Journal of Speech Technology*, Vol: 19, Issue: 1, pp: 87 – 93. 2016.

[4] S. Ioffe, and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *Proc. of ICML*, pp. 448–456. 2015.

[5] Ahmed Iqbal, Shabib Aftab, " A Classification Framework for Software Defect Prediction Using Multi-filter Feature Selection Technique and MLP ", International Journal of Modern Education and Computer Science, Vol.12, No.1, pp. 18-25, 2020.

[6] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545. 2014.

[7] A.S.C.S. Sastry, P.V.V. Kishore, C. Raghava Prasad, and M.V.D. Prasad, “Denoising ultrasound medical images: A block based hard and soft thresholding in wavelet domain,” *Medical Imaging: Concepts, Methodologies, Tools, and Applications*, Vol: Issue: pp: 761 – 775. 2016.

[8] Moner N. M. Arafa, Reda Elbarougy, A. A. Ewees, G. M. Behery," A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation", International Journal of Image, Graphics and Signal Processing, Vol.10, No.4, pp. 31-38, 2018.

[9] Vidyashree Kanabur, Sunil S Harakannanavar, Dattaprasad Torse, "An Extensive Review of Feature Extraction Techniques, Challenges and Trends in Automatic Speech Recognition", International Journal of Image, Graphics and Signal Processing, Vol.11, No.5, pp. 1-12, 2019.

[10] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J.R. Glass, “Highway long short-term memory RNNS for distant speech recognition,” *Proc. of ICASSP*, pp. 5755–5759. 2016.

[11] Prashengit Dhar, Sunanda Guha," A System to Predict Emotion from Bengali Speech ", International Journal of Mathematical Sciences and Computing, Vol.7, No.1, pp. 26-35, 2021.

[12] K.V.V. Kumar, P.V.V. Kishore, and D. Anil Kumar, “Indian Classical Dance Classification with Adaboost Multiclass Classifier on Multi feature Fusion,” *Mathematical Problems in Engineering*, Vol:20, issue: 5, pp: 126 - 139, 2017.

[13] K. Mannepalli, P.N. Sastry, and M. Suman, “FDBN: Design and development of Fractional Deep Belief Networks for speaker emotion recognition,” *International Journal of Speech Technology*, Vol: 19, Issue: 4, pp: 779 – 790. 2016.

[14] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 745–777. 2014.

[15] G.A. Rao, K. Syamala, P.V.V. Kishore, and A.S.C.S. Sastry, “Deep convolutional neural networks for sign language recognition,” *International Journal of Engineering and Technology (UAE)*, Vol: 7, Issue: 1.5, Special Issue 5, pp: 62 to 70, 2018.

[16] Yogesh Kumar, Navdeep Singh,"Automatic Spontaneous Speech Recognition for Punjabi Language Interview Speech Corpus", International Journal of Education and Management Engineering, Vol.6, No.6, pp.64-73, 2016.

[17] G.A. Rao, and P.V.V. Kishore, “Sign language recognition system simulated for video captured with smart phone front camera,” *International Journal of Electrical and Computer Engineering*, Vol: 6, Issue: 5, pp: 2176 – 2187, 2016.

[18] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, “Spatial Diffuseness Features for DNN-Based Speech Recognition in Noisy and Reverberant Environments,” *Proc. of ICASSP*, pp. 4380–4384, 2015.

[19] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, “Batch-normalized joint training for dnn-based distant speech recognition,” *Proc. of SLT*, pp. 28–34. 2016.

[20] Hajer Rahali, Zied Hajaiej, Noureddine Ellouze,"Robust Features for Speech Recognition using Temporal Filtering Technique in the Presence of Impulsive Noise", International Journal of Image, Graphics and Signal Processing, vol.6, no.11, pp.17-24, 2014.

[21] P.V.V. Kishore, and M.V.D. Prasad, "Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural network," *International Journal of Software Engineering and its Applications*, Vol: 10, Issue: 2, pp: 149 – 170. 2016.
[22] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "A network of deep neural networks for distant speech recognition". *Proc. of ICASSP*, pp. 4880–4884. 2017.
[23] S. Hochreiter. And J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780. 1997.
[24] G. Zhou, J. Wu, C. Zhang, and Z. Zhou, "Minimal gated unit for recurrent neural networks," *International Journal of Automation and Computing*. 2016.
[25] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks". *Proc. of Interspeech*, pp. 3274–3278. 2015.
[26] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. W. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," *Proc. of LVA/ICA*, Vol: 9237, Pages 91–99, 2015.
[27] H. Erdogan, J.R. Hershey, S. Watanabe, and J. L Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *Proc. of ICASSP*, pp. 708–712. 2015.
[28] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies," *IEEE International Conference on Acoustics*, pp. 708-712, 2013.
[29] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," *Proc. of ICML*, pp. 2342–2350. 2015.
[30] J. Chung, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Proc. Of NIPS*, Vol: 37, Pages 2342–2350, 2014.
[31] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," *Proc. of ICASSP*, pp. 2657–2661, 2016.

Authors' Profiles

J. Bennilo Fernandes received the Bachelor of Electronics and Communication Engg (ECE) degree from LIT, Anna University and Master of Technology in Embedded Systems (ES) degrees from HIT, Tamil Nadu, India. He is currently working as Assistant Professor in Dept of ECE, Koneru Lakshmaiah Education Foundation (K L University, Vijayawada), Andhra Pradesh. His research interest includes speech recognition, image processing, applications of machine learning and embedded systems. He has published 5 papers in International Journals and 3 papers in International / National Conferences. He has 6 years of teaching experience.

Kasiprasad Mannepalli is working as Assoc. Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (K L University, Vijayawada), Andhra Pradesh. His research interest includes Speech Processing, Image Processing, machine intelligence. He has published 10 papers in International Journals and 6 Papers in International / National Conferences. He has 12 years of teaching experience.

How to cite this paper: J. Bennilo Fernandes, Kasiprasad Mannepalli, " Enhanced Deep Hierarchal GRU & BILSTM using Data Augmentation and Spatial Features for Tamil Emotional Speech Recognition", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.3, pp. 45-63, 2022.DOI: 10.5815/ijmecs.2022.03.03