International Journal of Communication Networks and Information Security

ISSN: 2073-607X, 2076-0930

Volume ***14*** *Issue* ***02 Year 2022*** *Page* ***189:202***

Deep Residual Adaptive Neural Network Based Feature Extraction for Cognitive Computing with Multimodal Sentiment Sensing and Emotion Recognition Process

[1]Dr. Gopal Arora*, [2]Munish Sabharwal, [3]Ms. Pooja Kapila, [4]Divya Paikaray, [5]Dr. Vipul Vekariya, [6]Narmadha T

[1]Assistant Professor, Department of Chemistry, Sanskriti University, Mathura, Uttar Pradesh, India.
[2]Professor, School of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.
[3]Assistant Professor, Department of Computer Science Engineering, Chandigarh Engineering College, Jhanjeri, India.
[4]Assistant Professor, Department of Computer Science, ARKA JAIN University, Jamshedpur, Jharkhand, India.
[5]Professor, Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat, India.
[6]Assistant Professor, Department of Computer Science and Engineering, Jain (Deemed-to-be University), Bangalore, India.
[1] drgopalarora.chem@sanskriti.edu.in, [2] mscheckmail@yahoo.com, [3]pooja.j1004@cgc.ac.in, [4]divya.p@arkajainuniversity.ac.in, [5]vipul.vekariya18435@paruluniversity.ac.in, [6] naramadhat2021@gmail.com

Article History

Received: 24 March 2022
Revised: 28 July 2022
Accepted: 29 August 2022

Abstract

For the healthcare framework, automatic recognition of patients' emotions is considered to be a good facilitator. Feedback about the status of patients and satisfaction levels can be provided automatically to the stakeholders of the healthcare industry. Multimodal sentiment analysis of human is considered as the attractive and hot topic of research in artificial intelligence (AI) and is the much finer classification issue which differs from other classification issues. In cognitive science, as emotional processing procedure has inspired more, the abilities of both binary and multi-classification tasks are enhanced by splitting complex issues to simpler ones which can be handled more easily. This article proposes an automated audio-visual emotional recognition model for a healthcare industry. The model uses Deep Residual Adaptive Neural Network (DeepResANNet) for feature extraction where the scores are computed based on the differences between feature and class values of adjacent instances. Based on the output of feature extraction, positive and negative sub-nets are trained separately by the fusion module thereby improving accuracy. The proposed method is extensively evaluated using eNTERFACE'05, BAUM-2 and MOSI databases by comparing with three standard methods in terms of various parameters. As a result, DeepResANNet method achieves 97.9% of accuracy, 51.5% of RMSE, 42.5% of RAE and 44.9%of MAE in 78.9sec for eNTERFACE'05 dataset. For BAUM-2 dataset, this model achieves 94.5% of accuracy, 46.9% of RMSE, 42.9%of RAE and 30.2% MAE in 78.9 sec. By utilizing MOSI dataset, this model achieves 82.9% of accuracy, 51.2% of RMSE,

40.1% of RAE and 37.6% of MAE in 69.2sec. By analysing all these three databases, eNTERFACE'05 is best in terms of accuracy achieving 97.9%. BAUM-2 is best in terms of error rate as it achieved 30.2 % of MAE and 46.9% of RMSE. Finally MOSI is best in terms of RAE and minimal response time by achieving 40.1% of RAE in 69.2 sec.

Keywords- feature extraction, cognitive computing, sentimental sensing, emotion recognition, healthcare, neural network.

1. Introduction

In the living style of human, emotions drive the universal decision-making. Moreover, the type of emotion absolutely affects human attention, communication, and even personal skill of memorizing information [1]. Emotions are natural state of human. When emotional states undergo recognition and interpretation process, several severe challenges are experienced. Cognitive computing is the term which refers to the approaches used for detection, recognition and prediction of human emotions like trust, anger, joy, sadness, anticipation, surprise and so on [2]. The computing systems have the ability to exhibit empathy and even provide decision support adapted to the human emotional state. Emotional states of human are expressed through various physiological characteristics [3]. They indicate a variety of signs like muscular activities, heart rate or sweat. Several research methods are developed to gather emotions from content and the way it is communicated [4]. Basically, cognitive computing approaches are classified based on the modality of the message like it is in the form of gesture, speech, or text. Cognitive computing involves unimodal as well as multimodal analyses [5]. Videos help in recognizing facial expressions and vocal tone. The brain emotions are combined into neural network hierarchy. For fusing multiple modalities, three modules use different fusion techniques separately and a relative hierarchical structure is designed. One module accepts all inputs to perform binary polarity detection [6]. To this module, a hash layer is attached for increasing the retrieving speed and fault tolerance. In the rest of two modules, one is responsible for recognizing positive items and the other for negative items [7]. Here every module comprises of a fusion and decision layer. Various fusion and decision methods are used for binary classification. Additionally positive and negative multi-classification is also performed to ensure better performance [8]. Artificial intelligence is expected to help individuals and machines to build the society in future. In these situations, a bot can be designed to understand the complex emotions of human. Many research works were developed to interpret the complex emotions of human [9]. Deep learning techniques make the world understand the behaviours of human and fill the gap between the way a human and computer understand. With the power of AI, implementing cognitive intelligence in BOT systems and developing humanoids fulfils the needs of the public. Thus machines understanding human literally has become necessary in the future. This also supports actual sentiment analyser to understand emotions with words and actions [10]. This analyser helps in maintaining the quality of product. The contribution of this work is as follows,

- Construction of Deep Residual Adaptive Neural Network (DeepResANNet) which integrates cognitive neural science emotional processing method based on deep learning models using three modalities such as audio, visual and text file at one time
- To include adaptive window process in the pooling sub-layer to minimize the size of feature-map.

The organization of paper is as follows: In section 1 the background of emotion recognition, multimodal sentiment sensing, and the application of neural network in emotion recognition are discussed along with motivation and contribution. In section 2 the existing traditional methods for emotion recognition are discussed. Section 3 explains the proposed DeepResANNet architecture with feature extraction and fusion concept of audio, visual classification. In section 4 experimental analysis are done with three databases and graphs are obtained. Finally the paper ends with section 5 concluding this work with future enhancements.

2. Related Works

In [11] a Hierarchal Attention-BiLSTM (Bidirectional Long-Short Term Memory) model was developed based on Cognitive Brain limbic system (HALCB). HALCB categorized multimodal sentiment analysis as two modules namely binary and multi-classification. The former splits the input items as two categories after recognizing their polarity and then separately forwards to the latter. In [12], Monte Carlo method fused the results and solved the problem smaller databases. Multimodal emotion recognition method and Human-robot interaction (HRI) system were integrated to achieve higher recognition rate. Further, a perceptual assessment model was developed to validate the system. In [13], multimodal effective data analysis model was designed to obtain opinion and emotions of the user from videos. Particularly, Multiple Kernel Learning (MLK) was in combining modalities like visual, audio and text. This model was better in multimodal sentiment analysis with an accuracy margin of 10–13% on polarity detection and 3–5% on emotion recognition. In [14] an emotion recognition system from facial images was designed based on edge computing. A convolutional neural network (CNN) was used to recognize emotion. This model was trained in a cloud during off time and downloaded to an edge server. During testing, an end device such as a smartphone captured a face image and did some pre-processing, which included face detection, face cropping, contrast enhancement, and image resizing. The pre-processed image was then sent to the edge server. In [15], a 2D CNN for speech and a 3D CNN for visual were employed. From speech signal, after pre-processing, PS-PA feature vector was obtained. These were blended using ELM networks which were trained with gender and emotion-specific data. In healthcare, edge computing was used prior to cloud computing. However, from the review it is known that the architectures are not integrated with fusion methods to provide a single model. Fusion method was divided into several strategies to perform tasks in a better way. Complicated problems of multimodal sentiment analysis are split to more simpler problems in various stages of classification based on the requirements. Moreover the classification is separate for audio and video files which further makes problem in the fusion step. The aforementioned shortcomings of the literature motivated the formulation of the proposed DeepResANNet method which is discussed in the forthcoming section.

3. System Model

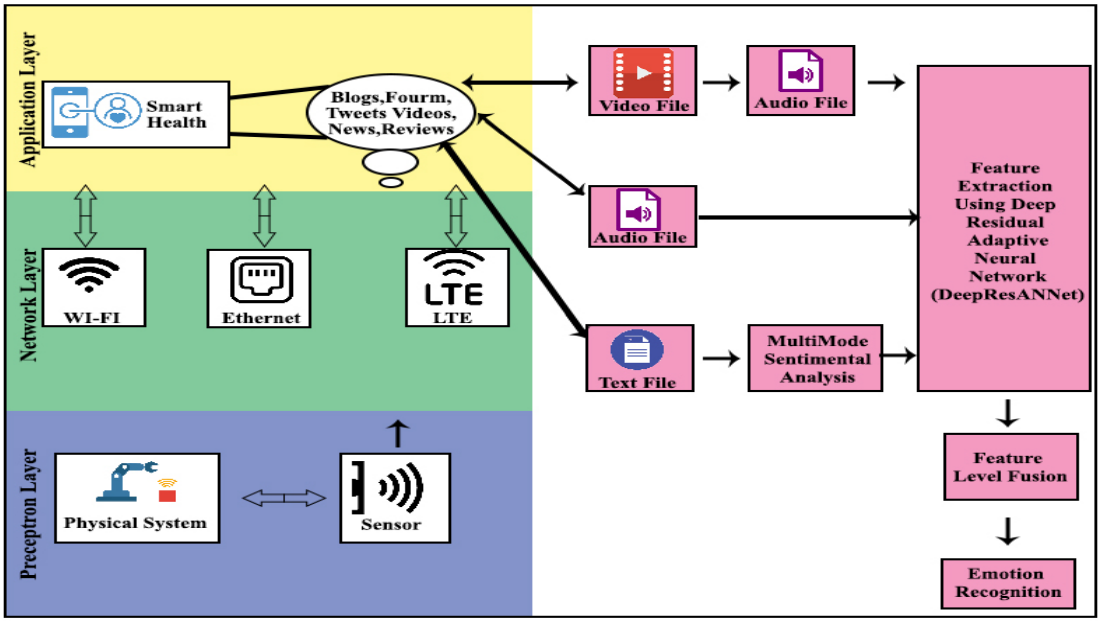

Figure-1 System architecture for emotion recognition in cognitive computing

The cloud contains a cloud manager, numerous dedicated servers, and huge storage. Requests are authenticated by cloud manager and task distribution to various servers and stakeholders are also

managed. After classifying the emotion in server, decision and few some samples are sent to the registered doctors by cloud manager followed by some prescriptions given by the doctors as reply which in turn are forwarded to the patients. Many service providers request for emotions that are classified and these requests are managed by the cloud manager. By nature, audio-visual data is large and thus it is difficult and complex to manage these data by service providers in terms of seamless transmission and storage management. Edge computing when used can provide a solution to the threat. Small cell (SC) base stations are connected to Internet of Things (IoT), and Bluetooth, RAN, or WiFi can be used to communicate between SCs and IoT. Battery-operated SCs secondary base station is as illustrated in figure-1.

Initially, text, audio and visual data are pre-processed producing computational sequences C = [Ca,Ct,Cv] and equalized the dimension of the vector as d. Here Ca, Ct and Cy represent the audio, textual and visual features respectively.

3.1 Analysis of text files using Multi modal sentiment analysis

Once data collection is completed, two pre-processing steps are performed. Spelling errors are corrected with the help of Google Spell Check and text pre-processing which includes segmentation, stemming, and part-of-speech tagging for which Stanford Core NLP is used. A link is established between review readability and helpfulness. For higher readability, reviews helpfulness has a positive consequence while spelling error has a negative consequence. Readability related to review length were taken into consideration which also included review rating, total characters, syllables, words and sentences, average syllables and words. Reviews are rated from 1 to 5, where 1 implies highly negative review and 5 implies the highly positive review [16] as indicated in equations (1) and (2):

$$\text{Gunning Fog Index (GFI)} = \left[\left(\frac{Wordi}{sentence\ i}\right) + 100\left(\frac{complex\ i}{word\ i}\right)\right] \quad (1)$$

$$\text{Simple Measure of Gobbledygook (SMOG)} = \sqrt{complex\ i\ \times \frac{30}{sentence\ i}} \quad (2)$$

For the given sentence, sentiment polarity along with strengths and weaknesses of subjective expressions are identified using the Opinion Finder, which is an efficient tool for sentiment as well as subjectivity analysis and negation detection. Thoughts disclosed in a text like sentiment, opinion, speculation and other types can be identified by Opinion Finder [17]. Moreover, for negation detection and polarity inversion, contextual polarity recognition method was employed. The sentiment classes for a sentence si namely Positive(si) and Negative(si) are focused in this study. Sentence polarity is identified as given in equation (3)

$$\text{Sentiment(si)} = (\text{str pos(si)}\times 2 + \text{weak pos(si)}) - (\text{str neg(si)}\times 2 + \text{weak neg(si)}) \quad (3)$$

where str pos(si) and str_neg(si) denote the frequency of positive and negative sentiment class respectively with strong subjectivity. Likewise, weak_pos(si) and weak_neg(si) are used to denote the frequency of positive and negative sentimental class respectively with weak subjectivity.

3.2 Feature extraction from audio file

Time and spatial domain information can be obtained from a audio file. When a voice signal changes with time it is termed as time domain where the features of a voice signal are in spatial domain. The information of time domain and spatial domain are the time-series and spatial features respectively in the direction of frequency domain. Once the acoustic features are extracted, Deep Residual Adaptive Neural Network (DeepResANNet) is used to model these low-level features.

3.3 Construction of Deep Residual Adaptive Neural Network (DeepResANNet)

To extract the features, DeepResANNet method is employed and scores are computed based on the differences between feature and class values of adjacent instances. With the useful attribute, it is

anticipated that in the same class, the closest distances is closer throughout than other classes. Thus, for the attribute given, the weight is computed as given in equation (4)

$$W=W\text{-}diff(x_{ij},near\_hit_{ij})^2+diff(x_{ij},near\_miss_{ij})^2/m \quad (4)$$

m denotes the size of the sample selected randomly from the training subset $diff(x_{ij},near\_hit_{ij})$, $diff(x_{ij},near\_miss_{ij})$ provides the difference between attribute values of arbitrarily selected j distance and $near\_hit_{ij}$ and $near\_miss_{ij}$ which is the closest training sample of same and different class respectively. The attribute to be useful, values of$near_miss_{ij}$ has to be closer to one another. When not useful, the differences almost are distributed in the same way and an attempt is made for selecting the closest attributes related to the class labels. Simultaneously, filtering method makes an attempt to reduce the redundancy among all the attributes that are selected. Every attribute and the vector of class label vector are evaluated as a discrete coincidence for which mutual information I(x,y) is employed as defined below.The similarity between an attribute and vector of class label or two attributes is measured as given in equation (5)

$$I(x,y)=\sum_{I,J}^{n} p(x,y)\log\frac{p(xi,yj)}{p(xi)p(yj)} \quad (5)$$

p(xi) and p(yj) are functions representing marginal feature probability, and p(xi, yj) is the joint probability distribution. If two variables selected at random are entirely independent, then the value of I(x,y) is 0.

Consider X as a training set with class labels for all samples. xi ∼ xj denotes xi and xj are of the same category; xi xk are from different categories [18]. The empirical error on every triplet (xi, xj , xk), xi ∼ xj , xi xk constrains from X is defined as given in equation (6):

$$(X) = Pr(dij > dik\ |xj \sim xi, xk+ xi)= Exi\sim xj,xi\ xk\ (1d)i\ j > dik-c \quad (6)$$

where dij = (xi − xj) (xi − xj) is the distance between xi and xj , xi and xj are the features of xi and xj respectively. The margin c requires a less distance between positive pair than the negative pair, or in the loss, a penalty is added.

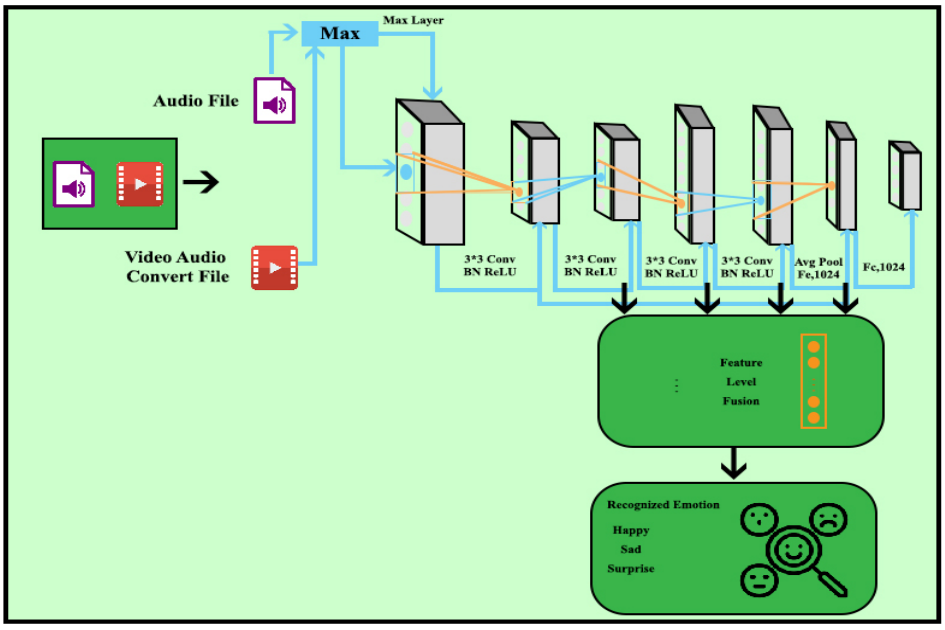

Figure-2 Architecture Of Deep Residual Adaptive Neural Network (DeepResANNet)

As shown in Figure 2, convolutional function is applied by the DeepResANNet framework on three various sizes of window in input layer. Moreover, mean function is applied on the output of convolutional sub-layer. At last, max function is applied on mean pooling sub-layer. With different sizes of window, DeepResANNet model proves its robustness against the variations in the speech signal of a speaker, among speakers, and phoneme length. In DeepResANNet model, the input layer is mapped with convolution sub-layer as given in equation (7):

$$qj,m,nW = \sigma(\sum_{i=1}^{i(nw)} \sum_{n=1}^{f} oin + m - 1, w(i, j + wi) \quad (7)$$

here nW is the window index. The input for nW^{th} feature map is obtained from nW^{th} input window. The input features oi(i = 1,...,I(nW)) varies in dimension where I(nW) represents the nW^{th} window length. Every input feature is linked with the units of convolution sub-layer, Qj(j = 1,..., J), through weight matrices wi, j(i = 1,...,I; j = 1,..., J). oi,m denotes the m^{th} unit of input i [19]. The filter size F denotes the number of frequency bands from the input to convolutional sub-layer as given in equation (8):

$$q(j,n) = \sigma(\sum_{i=1}^{i(nw)} \sum_{n=1}^{f} oi * w(i, j) \quad (8)$$

The number of local weight matrices in the convolution maps is determined by the number of available feature-maps in the convolution sub-layer. In DeepResANNet model, three windows are considered for computation. I and F denotes the length and width of every widow. Mean-pooling function is applied on every feature-map of convolutional sub-layer and this mean obtained is provided as the output of the first pooling sub-layer. In DeepResANNet model, every mean-pooling sub-layer unit is calculated as given in equation (9):

$$T(j,m) = r* \sum_{nW=1}^{NW} qj(m, n, w) \quad (9)$$

here r represents the normalization coefficient. The max function used by the next pooling sub-layer is applied on the output of the mean sub-layer. In DeepResANNet model, every max-pooling sub-layer unit is computed as given in equation (10):

$$P(j,m) = \max [tj(m-1)*s+n] \quad (10)$$

here G denotes the size of the pooling window. The stride size s presents the overlapping of two window.

xi is considered as the contextual feature of the current time series and is calculated using RNN, the attention score βi is computed as given in equation (11):

$$\beta i = \frac{1}{1+e(xi)} \quad (11)$$

Actually sigmoid function is used to compute βi whose value ranges from 0 to 1. In audio, βi can be described as the frame's contribution score to the Final representation utterance level C of the audio as given in equation (12):

$$C=\frac{\sum_i \beta 1.xi}{\sum_i \beta i} \quad (12)$$

Where, $\sum_i \beta i.xi$ is the weighted sum of features at every position. After activating the results of utterance level, they pass through fully connected layer to determine the final result which in turn passes to softmax layer thereby obtaining posterior probability of every audio type.

3.2 Algorithm for training process

Input: audio files (Ca), emotion label Y, parameters γ and δ

Output: feature extracted audio file
compute weight (W)
W(near)= diff(x_{ij},near_hit$_{ij}$),
W(miss)= diff(x_{ij},near_miss$_{ij}$)
W← similarity (S)
For
W<S
Else
Compute the training samples (xi, xj , xk)
End for
Frequency feature for samples ← q(j,n)
window size ={1,2,3...n}
time series <= Bi
compute final weighted sum (C)

3.3 Feature level fusion

A simple heterogeneous fusion scheme is employed with the aim of enhancing the efficiency of emotion recognition from speech due to the integrated feature representation. This fusion strategy comprises of four-layer deep neural network namely one input and three hidden layers. This method captures the relationship among integrated features to recognize emotion by using DeepResANNet. From the multimodal emotion features vector V={v1,v2,v3,.....vt}, the final descriptor is created where textual (H) and acoustics (S) features are integrated as in equation (13):

V=[H,S] (13)

Then, V is given as input to a DNN comprising of three dense layers and a softmax layer. 1024, 512, and 4 are the parameters for dense layers. Softmax layer identifies the relationship among features of various traits [20] whose output is the relative probability among various emotion classes as given in equation (14):

$$P(xa) = \text{softmax}(xa) = \frac{\exp(xa)}{\sum_{a}^{A} \exp(xa')} \quad (14)$$

where a indicates emotion class and P(xa) is its probability. To get rid from overfitting, regularization is added to train multimodal features. Regularization concept adds an index which describes the model's complexity in loss function

4. Performance Analysis

The experiment is carried out and accuracy, response time, Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Mean Absolute Error (MAE) are the parameters used for analysis. These values obtained for these parameters are compared against three different databases as given below:

4.1 Datasets

eNTERFACE'05 [21]: This contains 42 subjects of 14 various nations among which 81% were men and 19% were women. Among them, 31% wore glasses, 17% had beard. Recording was made in two weeks. Every subject heard six successive English short stories and elicited some specific emotion. Then, these subjects reacted according to the situation which was judged by two experts. They identified the expressed reaction whether it was an emotion and if so the sample was stored in the database or else the sample was discarded.

BAUM-2 [22]: contains annotated audio-visual clips of face which includes different head poses, variations in illumination conditions, temporary occlusions, accessories, and moreover, subjects were from various age groups. These clips simulate the conditions of real-world.
MOSI [23]: is extensively used for sentiment analysis. This dataset contains 93 video-blogs or vlogs which were selected randomly from YouTube. The annotation range of sentiment intensity ranges from −3 to +3 on utterance level.
The proposed model compares with three standard methods such as Hierarchal Attention-BiLSTM (Bidirectional Long-Short Term Memory) model based on Cognitive Brain limbic system (HALCB), Multiple Kernel Learning (MLK), Convolutional Neural Network (CNN) which is capable of achieving better classification performance. The description and formula for parameters used are as follows:
True positive (TP) and true negative (TN) estimates the ability of the classifier. False positive (FP) and false negative (FN) represents the false predictions produced by the model.

a. Accuracy
It is the ability of prediction of the proposed technique which is as given in equation (15):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

b. Response time
This is the average time taken by the algorithm which is as given in equation (16):

$$\text{RT}=\frac{1}{n}\sum_{vi,yi}\mu(I)vi,yi \tag{16}$$

c. Root Mean Square Error (RMSE)
This measure gives the difference between values predicted by a model and values observed as given in equation (17):

$$\text{RMSE} = \sqrt{\sum_t^T (y'(t) - y(t)} \tag{17}$$

d. Relative Absolute Error (RAE)
This is the probability of mean error to the errors caused by the model as given in equation (18):

$$\text{RAE} = \frac{\sum_{i=1}^{n}(pi-Ai)^2}{\sum_{i=1}^{n} Ai} \tag{18}$$

e. Mean Absolute Error (MAE)
This is the error of observations expressed by the same process as given in equation (19):

$$\text{MAE}= \sum_{i=1}^{n}(yi - xi) \tag{19}$$

Configuration of DeepResANNet with the state and size of each layers in indicated in table-1

Table-1 Configuration of DeepResANNet

	State	**size**
Preprocessing	Image	125×125
	Image division	231×4×4
max layer	CONV1	245×4×4×56
	CONV1	245×2×2×71
	CONV1	245×4×6×68
	CONV1	245×1×1×31
pooling layer	MULTI	245×4×67
	MAX POOLING	9×9×50

	MULTI	18×18×39
Stride layer	FC6	437
	SOFTMAX	6

Table-2 shows the analysis of various parameters such as accuracy, response time, RMSE, RAE and MAE for three databases namely, eNTERFACE'05, BAUM-2, and MOSI.

Table-2 Comparison of various parameters

Datasets	**Method**	**accuracy**	**response time**	**RMSE**	**RAE**	**MAE**
eNTERFACE'05 [21]	**HALCB [11]**	96.5	72.9	54.1	44.5	47.5
	MLK [13]	97.3	74.2	53.5	43.9	47.3
	CNN [14]	97.5	75.4	52.9	43.5	47.1
	DeepResANNet [proposed]	97.9	78.9	51.5	42.5	44.9
BAUM-2 [22]	**HALCB [11]**	93.5	73	52.5	43.5	45
	MLK [13]	93.9	75.1	49.6	43.5	44.9
	CNN [14]	94.3	75.6	49.6	43.1	37.6
	DeepResANNet [proposed]	94.5	78.9	46.9	42.9	30.2
MOSI [23]	**HALCB [11]**	76.1	80	53.5	44.9	39.5
	MLK [13]	80	76.5	52.9	44.7	38.9
	CNN [14]	82.3	73.1	51.6	43.2	38.1
	DeepResANNet [proposed]	82.9	69.2	51.2	40.1	37.6

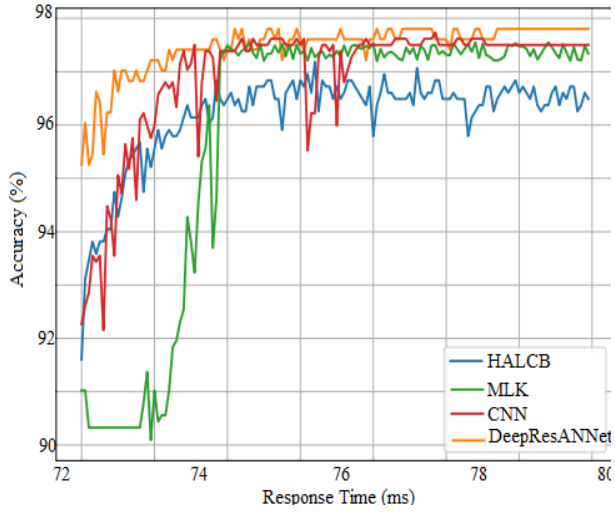

Figure 3(a)

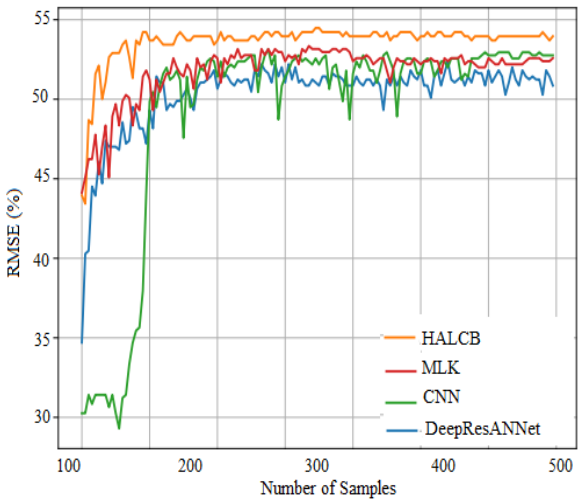

Figure 3(b)

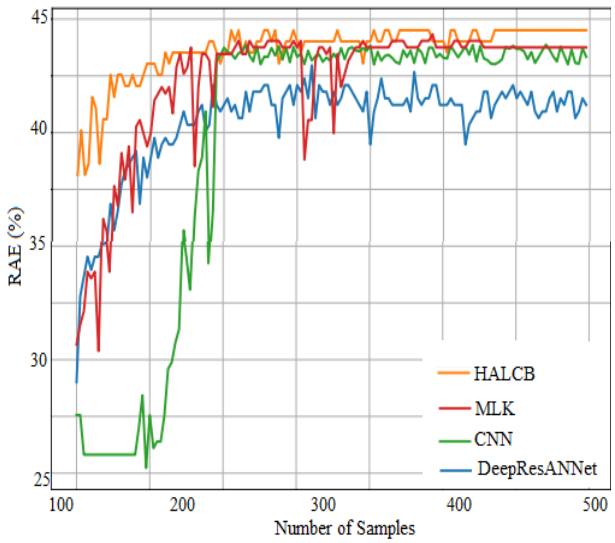

Figure 3(c)

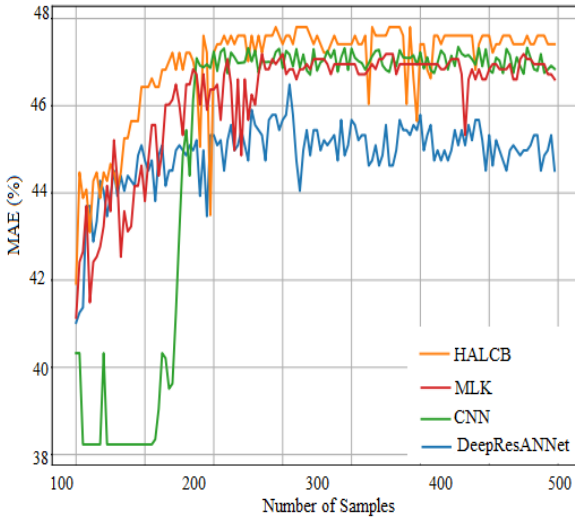

Figure 3(d)

Figure 3 Comparison of various parameters on eNTERFACE'05 database

Figure 3 Comparison of various parameters on eNTERFACE'05 database Figure 3(a) indicates the analysis between accuracy and response time and it shows that HALCB achieves 96.5% of accuracy in 72.9sec, MLK achieves 97.3% of accuracy in 74.2sec, CNN achieves 97.5% of accuracy in 75.4sec when comparing with proposed DeepResANNet method achieves 97.9% of accuracy in 78.9sec. Figure 3(b) indicates the analysis of RMSE for database eNTERFACE'05and it shows that HALCB achieves 54.1% of RMSE, MLK achieves 53.5% RMSE, CNN achieves 52.9% of RMSE when comparing with proposed DeepResANNet method achieves 51.5% of RMSE. Figure 3(c) indicates the analysis of RAE for database eNTERFACE'05and it shows that HALCB achieves 44.5% of RAE, MLK achieves 43.9% RAE, CNN achieves 43.5% of RAE when comparing with proposed DeepResANNet method achieves 42.5% of RAE. Figure 3(d) indicates the analysis of MAE for database eNTERFACE'05and it shows that HALCB achieves 47.5% of MAE, MLK achieves 47.3% MAE, CNN achieves 47.1% of MAE when comparing with proposed DeepResANNet method achieves 44.9% of MAE.

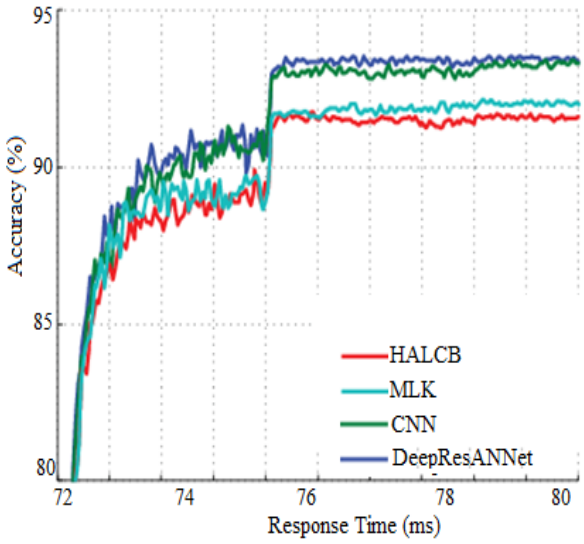

Figure 4(a)

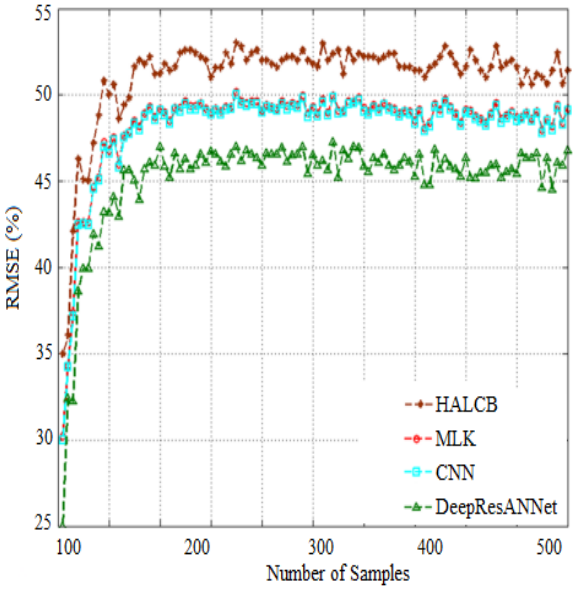

Figure 4(b)

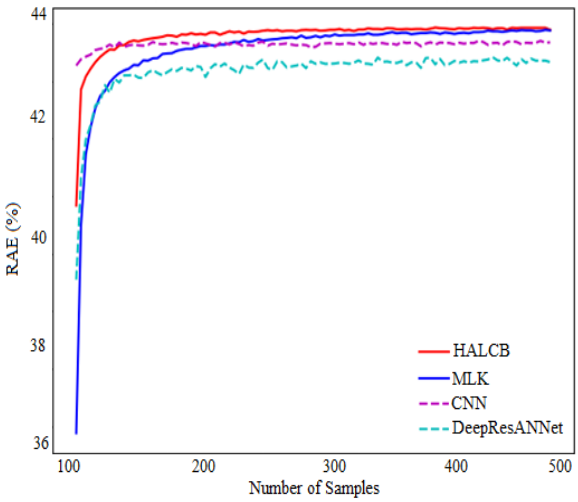

Figure 4(c)

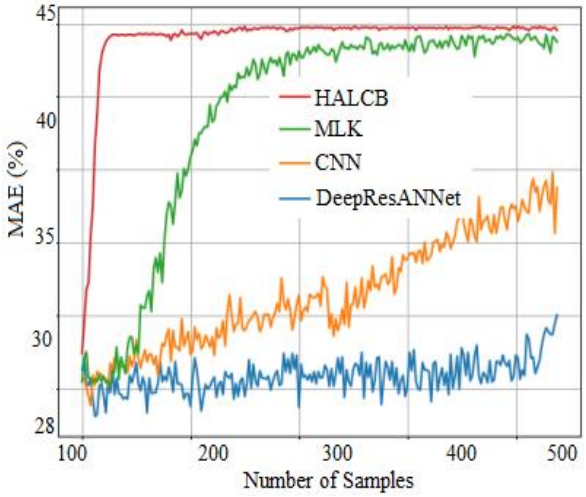

Figure 4(d)

Figure 4 Comparison of various parameters on BAUM-2 database

Figure 4 Comparison of various parameters on BAUM-2 database. Figure 4(a) indicates the analysis between accuracy and response time and it shows that HALCB achieves 93.5% of accuracy in 73 sec, MLK achieves 93.9% of accuracy in 75.1sec , CNN achieves 94.3% of accuracy in 75.6sec when comparing with proposed DeepResANNet method achieves 94.5% of accuracy in 78.9sec.

Figure 4(b) indicates the analysis of RMSE and it shows that HALCB achieves 52.5% of RMSE, MLK achieves 49.6% RMSE, CNN achieves 49.6% of RMSE when comparing with proposed DeepResANNet method achieves 46.9% of RMSE. Figure 4(c) indicates the analysis of RAE and it shows that HALCB achieves 43.5% of RAE, MLK achieves 43.5% RAE, CNN achieves 43.1% of RAE when comparing with proposed DeepResANNet method achieves 42.9% of RAE.Figure 4(d) indicates the analysis *of MAE and it shows that HALCB achieves 45% of MAE, MLK achieves 44.9% MAE, CNN achieves 37.6% of MAE when comparing with proposed DeepResANNet method achieves 30.2% of MAE.*

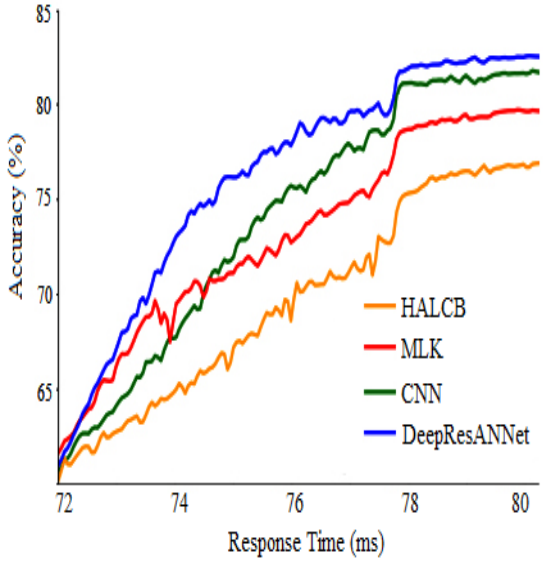

Figure 5(a)

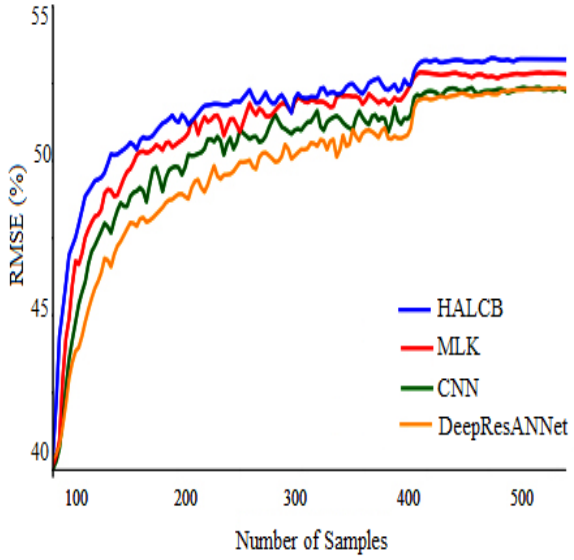

Figure 5(b)

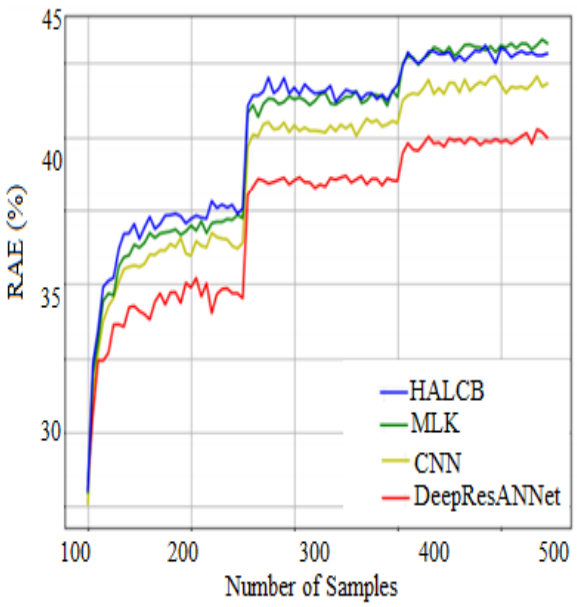

Figure 5(c)

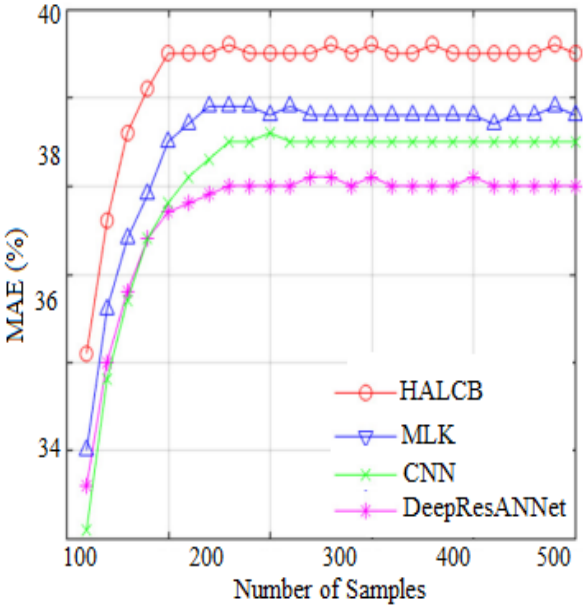

Figure 5(d)

Figure 5 Comparison of various parameters on MOSI database

Figure 5 Comparison of various parameters on MOSI database. Figure 5(a) indicates the analysis between accuracy and response time and it shows that HALCB achieves 76.1% of accuracy in 80sec, MLK achieves 80% of accuracy in 76.5sec, CNN achieves 82.3% of accuracy in 73.1sec when comparing with proposed DeepResANNet method achieves 82.9% of accuracy in 69.2sec. Figure 5(b) indicates the analysis of RMSE and it shows that HALCB achieves 53.5% of RMSE, MLK achieves 52.9% RMSE, CNN achieves 51.6% of RMSE when comparing with proposed DeepResANNet method achieves 51.2% of RMSE.Figure 5(c) indicates the analysis of RAE and it shows that HALCB achieves 44.9% of RAE, MLK achieves 44.7% RAE, CNN achieves 43.2% of RAE when comparing with proposed DeepResANNet method achieves 40.1% of RAE. Figure 5(d) indicates the analysis of MAE and it shows that HALCB achieves 39.5% of MAE, MLK achieves 38.9% MAE, CNN achieves 38.1% of MAE when comparing with proposed DeepResANNet method achieves 37.6% of MAE.

5. Conclusion

Cognitive computing permits inferring in emotional states either individually or collectively from text information and thus an anthropomorphic route is offered for to make decisions. Deep learning approaches has provided remarkable performance is different tasks. But in natural language processing, the process of feature extraction and fusion are challenging while recognizing emotions. The proposed Deep Residual Adaptive Neural Network (DeepResANNet) method helps to overcome the issue by evaluate the scores on various feature and class values of adjacent instances. The effectiveness of this proposed approach is validated with standard methods and was observed that the proposed DeepResANNet method achieves 97.9% of accuracy, 51.5% of RMSE, 42.5% of RAE and 44.9%of MAE in 78.9sec for eNTERFACE'05 dataset. By utilizing BAUM-2 dataset, the proposed DeepResANNet achieves 94.5% of accuracy, 46.9% of RMSE, 42.9%of RAE and 30.2% MAE in 78.9 sec. By utilizing MOSI dataset, the proposed DeepResANNet achieves 82.9% of accuracy, 51.2% of RMSE, 40.1% of RAE and 37.6% of MAEin 69.2sec. By analyzing all these three databases, eNTERFACE'05 is best interms of accuracy, since it achieves 97.9% of accuracy. BAUM-2 is best interms of MAE since it achieves 30.2% of MAE. Finally MOSI is best by achieving 51.2% RMSE, 40.1 of MAE in 69.2 sec. The future work may be concentrated on analyzing more datasets by using various ensemble deep neural network approaches to increase the accuracy rate with minimized response time.

References

[1] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Harnessing the cloud of patient experience: Using social media to detect poor quality healthcare. BMJ Quality & Safety, 22 , 251{255.

[2] Izard, C. E. (2009). Emotion theory and research: Highlights, unanswered questions, and emerging issues. Annual Review of Psychology, 60 , 1{25.

[3] Mahmoudi, N., Docherty, P., & Moscato, P. (2018). Deep neural networks understand investors better. Decision Support Systems, 112 , 23{34.

[4] Meisheri, H., Saha, R., Sinha, P., & Dey, L. (2017). Textmining at EmoInt 2017: A deep learning approach to sentiment intensity scoring of english tweets. In Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 193{199).

[5] Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems, 53 , 730{741.

[6] Ramesh, S., Nirmalraj, S., Murugan, S., Manikandan, R., & Al-Turjman, F. (2021). Optimization of Energy and Security in Mobile Sensor Network Using Classification Based Signal Processing in Heterogeneous Network. Journal of Signal Processing Systems, 1-8.

[7] J. Zhao, X. Mao, and L. Chen, ``Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomed. Signal Process. Control, vol. 47, pp. 312_323, Jan. 2019.

[8] C. Wang, ``Interpreting neural network hate speech classi_ers," in Proc. 2nd Workshop Abusive Lang. Online (ALW2), 2018, pp. 86_92.

[9] M. Shamin Hossain- Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, Ghulam Muhammad - Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, "An Emotion Recognition System for Mobile Applications", IEEE Special Section of emotion-aware mobile computing(2017)

[10] Jian Guo, Zhen Lei, Jun Wan, Eglis Avots – Dept. of Electrical and Electronic Engineering, Hasan Kalyponeu University, Turkey, "Dominant and Complementary Emotion Recognition from Still Images of Faces", IEEE Special Section on Visual Surveillance And Biometrics (2018

[11] Li, Y., Zhang, K., Wang, J., & Gao, X. (2021). A cognitive brain model for multimodal sentiment analysis based on attention neural networks. Neurocomputing, 430, 159-173.

[12] Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C. F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. Biomedical Signal Processing and Control, 70, 103029.

[13] Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing, 261, 217-230.

[14] Muhammad, G., & Hossain, M. S. (2021). Emotion Recognition for Cognitive Edge Computing Using Deep Learning. IEEE Internet of Things Journal.

[15] Hossain, M. S., & Muhammad, G. (2019). An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework. IEEE Wireless Communications, 26(3), 62-68.

[16] Wilson, T., Wiebe, J., & Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of human language Technology conference and conference on empirical methods in natural language processing.

[17] Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews – a text summarization approach. Expert Systems with Applications, 36(2), 2107–2115

[18] Robinson, A. E., Hammon, P. S., and de Sa, V. R. (2007). Explaining brightness illusions using spatial filtering and local response normalization. Vis. Res. 47, 1631–1644. doi: 10.1016/j.visres.2007.02.017

[19] C. Wei, L.l. Chen, Z.-z. Song, X.-g. Lou, D.-d. Li, EEG-based emotion recognition using simple recurrent units network and ensemble learning, Biomed. Signal Process. Control 58 (2020) 101756

[20] T. Song, W. Zheng, P. Song, Z. Cui, "EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks," IEEE Transactions on Affective Computing, 2018

[21] O. Martin et al., "The Enterface'05 Audiovisual Emotion Database," IEEE Wksp. Multimedia Database Management, 2006.

[22] C. E. Erdem, C. Turan, and Z. Aydn, "BAUM-2: A Multilingual Audio-Visual Affective Database," Multimedia Tools and Applications, vol. 74, no. 18, Sept. 2015, pp. 7429–59; http://baum2.bahcesehir.edu.tr. DOI: 10.1007/s11042-014- 1986-2

[23] Dhablia, A. (2021). Integrated Sentimental Analysis with Machine Learning Model to Evaluate the Review of Viewers. Machine Learning Applications in Engineering Education and Management, 1(2), 07–12.

[24] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259, 2016.