I.J. Modern Education and Computer Science, **2022, 4, 67-84**
Published Online August 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2022.04.06

Deep Learning Network and Renyi-entropy Based Fusion Model for Emotion Recognition Using Multimodal Signals

Jaykumar M. Vala
Ph.D. Research Scholar, Computer/IT Engineering, Gujarat Technological University, Chandkheda, Gandhinagar, Gujarat 382424, India
Email: jayvala1629@gmail.com

Udesang K. Jaliya
Ph.D. Supervisor, Gujarat Technological University, Chandkheda, Gandhinagar, Gujarat 382424, India, &
Assistant Professor, Department of Computer Engineering, BVM Engineering College, V.V. Nagar, Gujarat 388120, India
Email: udayjaliya@gmail.com

Received: 29 December 2021; Revised: 17 February 2022; Accepted: 10 April 2022; Published: 08 August 2022

Abstract: Emotion recognition is a significant research topic for interactive intelligence system with the wide range of applications in different tasks, like education, social media analysis, and customer service. It is the process of perceiving user's emotional response automatically to the multimedia information by means of implicit explanation. With initiation of speech recognition and the computer vision, research on emotion recognition with speech and facial expression modality has gained more popularity in recent decades. Due to non-linear polarity of signals, emotion recognition results a challenging task. To achieve facial emotion recognition using multimodal signals, an effective Bat Rider Optimization Algorithm (BROA)-based deep learning method is proposed in this research. However, the proposed optimization algorithm named BROA is derived by integrating Bat Algorithm (BA) with Rider Optimization Algorithm (ROA), respectively. Here, the multimodal signals include face image, EEG signals, and physiological signals such that the features extracted from these modalities are employed for the process of emotion recognition. The proposed method achieves better performance against exiting methods by acquiring maximum accuracy of 0.8794, and minimum FAR and minimum FRR of 0.1757 and 0.1806.

Index Terms: Multimodal emotion recognition, facial expression, EEG signal, physiological signal, deep learning.

1. Introduction

The process of emotion recognition depends on the voice parameters, bio-signals, as well as facial features. The engineers and the psychologists tried to examine the bio-signals, gestures, facial expressions, and vocal emotions in the attempt to recognize and classify the emotions [1]. In general, the bio-signals are being used to find the emotions and the human feelings, as it is quite simple to find emotions by the non-invasive sensor such that the reactions provoked by the emotions are low sensitive in the cultural diversity. However, there exist a burly interaction among emotional state or human emotions and bio-signal response [2]. Different studies presented various methods for emotion recognition that is depends on correlation among responses of bio-signal as well as basic emotions, like anger, fear, despair, sadness, happiness, and joy [3,4,5,6,7]. The external multimedia motivation is the important factor for recognition process such that it reveals more effective experience. Mood swing is the abstract and it shows psychological actions is a symbolic way to show multimodal physiological reactions. Due to different psycho-physiological clues degraded by the fluctuations of emotions, different researchers concentrated more in analysing the relations between multimedia features, human emotions, and EEG signal such that different techniques are explored for achieving effective computer intelligence [8,9]. Recent studies gained more benefits in the usage of physiological signals to recognize the emotions [10]. The electroencephalogram (EEG) signal is shown as the robust sole modality [11,12]. The controlling process of bio-signals is carried out by central nervous system [13]. The facial expression recognition (FER) is intended to detect the facial expressions, namely anger, happiness, fear, disgust, surprise, and sadness from the facial modalities. However, this technique makes the machine to understand emotion or intention of humans through the analysis of facial images [4]

[14,15,16,17,18,19].

When compared with uni-modal model, multi-modal scheme utilizes information from different modalities [20]. Most of the multi-modal recognition approaches used in existing works lack in accuracy and efficiency while dealing with untrimmed and raw emotional data, like speeches of subjects and raw videos [21,22]. The input data is acquired from diverse modalities, like video, EEG, and audio in multi-modal fusion such that the EEG is coherently fused together [23]. A single modality is used in the context freeway, whereas in context dependent scheme, mot than one modality is used. The information combination is categorized into different stages, namely early stage, intermediate, and the late fusion. The integration of data is made at feature or signal level in early fusion, whereas data integration is made at semantic level in the late fusion. However, the foundation in integrating the muli-modal information is the modality number [24], fusion procedure, and information derivation organization to find the fusion level of data. However, different modalities offered complementary information at fusion stage. Hence it is required to comprehend contribution of individual modality with respect to success of discrete tasks [13][25].

In past years, researchers utilized time frequency distribution model and the spectral analysis techniques, like Fourier transformation (FT), and wavelet transformation model (DWT) [12]. Due to the subjective and complex character of emotional state, it is very complex to define a scheme to analyze various emotional status. However, frequency factor changes with respect to frequency and time component information is not sufficient for classifying human emotions in the non-stationary signals. Hence, to gather the information of signal frequency in temporal and spatial domain, the continuous wavelet transform (CWT) is developed to acquire full knowledge [26]. Due to extraordinary performance of the deep learning methods, different works are emerged for emotion recognition using multiple modalities by employing deep belief network (DBN) [27,28], convolutional neural network (CNN) [13][26], deep neural network (DNN), and LSTM-RNN [29,30]. The support vector machine (SVM), CNN, deep learning [14][26], and k-nearest neighbor (KNN) are the classification methods to be commonly used in emotion recognition. The CNN, deep learning method, and SVM generate higher classifier accuracy when compared with KNN model. The major issue of using this method is that it consumes more training time [32].

This research is focused to model an effective method named BROA-based deep learning approach for recognition tasks using multimodal signals. Here, the signals used to perform the process of emotion recognition include face image, EEG signal, and physiological signal. At first, face image is captured from video and the face image is pre-processed more accurately to remove noise. Accordingly, feature extraction process is accomplished for the pre-processed image using Local Directional Ternary Optimal Oriented Pattern (LDTOOP). The process of emotion classification using face image is achieved by Deep Belief Network (DBN) classifier. On the other hand, the EEG signal and the physiological signal for the respective face image are acquired and are allowed to feature extraction phase. Accordingly, the features, such as wavelet coefficients, and spectral features that comprise power spectral density (PSD), tonal power ratio, spectral flux, spectral skewness, and the spectral centroids are acquired from EEG and physiological signals. Moreover, the features acquired from both signals are fed to Deep Recurrent Neural Network (Deep RNN) in order to generate recognition result. The deep learning classifiers named DBN and Deep RNN are trained with proposed algorithm named BROA, which is derived by the integration of BA and ROA, respectively. Finally, the Renyi-entropy based fusion model is used to fuse the classification result acquired from DBN and Deep RNN in order to generate the recognition results, as sadness, happiness, surprise, anger, fear, and disgust.

The major contribution of the research is explained as follows:

- **Proposed BROA-based deep learning:** An efficient emotion recognition approach is developed using proposed BROA using multimodal signals. The different modalities utilized for the recognition includes face image, EEG signal, and physiological signal. The features extracted from face image are processed by DBN, whereas the features captured from EEG and physiological signals are processed using Deep RNN classifier.

The paper is organized as follows: Section 2 explains the review of different emotion recognition systems, and section 3 elaborates proposed method. Section 4 presents results and discussion of proposed approach, and section 5 concludes the research.

2. Motivation

In this section, some of the conventional emotion recognition approaches along with their merits and drawbacks are explained such that it motivates the researchers to model proposed BROA-based deep learning classifier.

2.1 Literature survey

Various traditional emotion recognition schemes are presented in this section. ChaoLia, *et al.* [33] modeled a bidirectional LSTM-RNN method for emotion recognition using different modalities. Here, the physiological signal acquired from each channel was converted into the spectrogram image to capture the frequency and time information. The attention based BLSTM-RNN was considered for learning the temporal features automatically. The deep features were passed to DNN for generating the emotional output of each channel. Muhammad Adeel Asghar, *et al.* [26]

introduced a DNN approach for the recognition of emotions. The temporal-based, spatial and the bag of deep features were acquired to minimize the dimensionality of features. Here, the emotions of individual subject were represented by histogram of vocabulary set captured from raw feature. Kuan Tung, *et al.* [34] introduced an XGBoost method for predicting the emotions. Different entropy domain features were acquired from the GSR and ECG signals. This method increased the performance, but it further required to enhance the reliability. Rania M. Ghoniem, *et al.* [13] developed a hybrid fuzzy c-means-genetic algorithm- based NN scheme for the classification of unimodal data using the fitness function. The fitness measure was employed to compute optimal fuzzy cluster by minimizing classification error. A separate classifier was employed to integrate the speech signal with the EEG data. The computation time of this method was reduced, but failed to recognize the emotions by deep learning method.

Gilsang Yoo, *et al.* [4] developed a back-propagation NN method for identifying the human emotions in the multi-dimensional segmentation using emotion parameters. The features were acquired from physiological signals and the NN classifier was employed to classify the emotions. This method was not provided reliability and stability. Tengfei Song, *et al.* [22] developed an attention-based LSTM (A-LSTM) model for emotion recognition of humans. Here, the physiological signals were recorded to describe the discrete emotions. However, the correlation among participant ratings and EEG signals were investigated. Jiamin Liu, *et al.* [30] developed an LSTM-RNN model for the emotion recognition by integrating the band and temporal attention. The EEG slice and the video were considered as input at each time stamp to generate the representation of signals. Based on fusion, the network predicted emotion label for further analysis. It needed more attention to skip redundant data automatically. Baixi Xing, *et al.* [8] developed a machine learning method for emotion recognition using video. Here, data fusion process was employed to integrate the features, such as audio visual and spectrum density features to increase the recognition result. The EEG features offered effective information for the process of recognition. It achieved higher recognition rate, but it needed to consider various modalities of features to the process of facial expression.

2.2 Challenges

Some of the issues faced by the traditional emotion recognition systems are explained as follows:

- A major issue faced in process of multi modal emotion recognition is feature fusion. Most of the existing works concentrated on the integration of features acquired from various modalities, but failed to consider conflicting information obtained from uni-modal modalities [1][13][14].
- A popular modality used to emotion recognition is facial expression, but it does not reflect intrinsic mental states of human directly, as it is the non-physiological signal [22].
- Due to the factors, like partial occlusions of head deflection, facial regions, and changes in the illumination, facial emotion recognition results a challenging task. However, such interference degrades the performance of facial detection [4] [36,37,38,39].
- To represent the EEG signal, result a complex task, as the raw EEG signal was transformed to the sequence of images over three frequency bands of the brain wave. To transform EEG data to video using spatial information simplifies the process in generating the features and to integrate the features of EEG data with video data [30].
- The traditional techniques scan the complete sequence of data and select peak data, which states that it failed to skip redundant information [30].

3. Proposed Bat Rider Optimization Algorithm-based Deep Learning for Facial Emotion Recognition Using Multimodal Signals

Emotion plays a major role in different aspects of human life, like decision making, and social communication. Multimodal emotion recognition is an important factor in different applications, like health care, education, and customer service. This research considered multimodality signals, like face image, EEG signal, and physiological signal to achieve emotion recognition process. The face image is pre-processed initially to remove the external artifacts of image and thereafter the feature extraction from pre-processed image is carried out with LDTOOP, which is derived by incorporation of LOOP and LDTP, respectively. From both EEG and physiological signal, the features like wavelet coefficient, and spectral features that comprises power spectral density (PSD), spectral flux, tonal power ratio, spectral skewness, and spectral centroids are acquired for further processing. Accordingly, features acquired from face image are processed using DBN classifier, whereas the features from both the signals are processed by Deep RNN classifier. Accordingly, training process of deep learning classifier is done using proposed BROA. At last, Renyi-entropy based fusion method is used to find the class label. Figure 1 represents schematic view of proposed method.

3.1 Acquisition of input video

Emotion recognition plays a significant role in emotional computing system. Emotion is the composite state that integrates thoughts, behavior, feelings, and the psycho-physiological reactions to the external or internal stimuli. The input used for the facial emotion recognition system is the multimodal signal that includes video, EEG signal, and the physiological signal. The inputs considered to process the recognition mechanism are acquired from DEAP dataset [40].

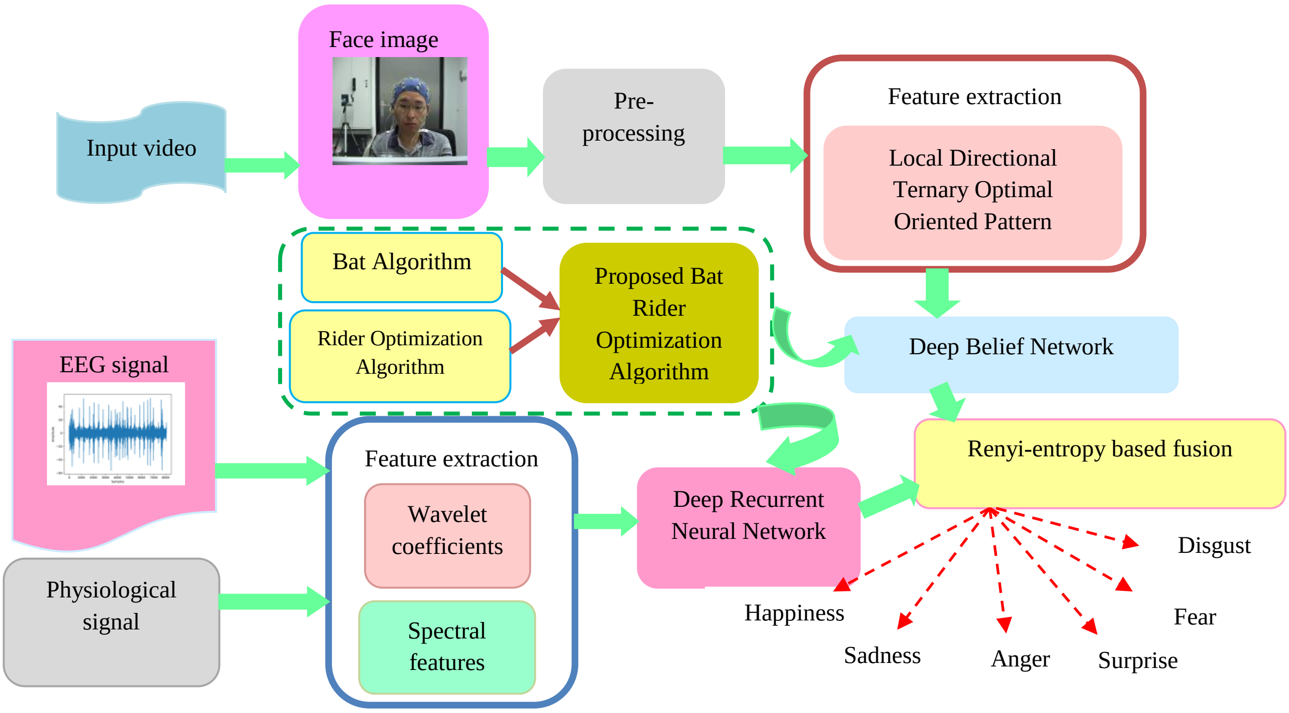


Fig.1. Schematic view of proposed BROA-based deep learning

3.1.1 Extraction of face image

Let us consider dataset as D with number of videos I is expressed as,

$$D = \{I_1, I_2, ..., I_i, ..., I_n\} \tag{1}$$

Here, D indicates dataset, and n represents number of videos. Each video contains n number of face images that are captured at different locations. The face images present in i^{th} video of dataset D is represented as,

$$I_i = \{A_1, A_2, ..., A_i, ..., A_n\} \tag{2}$$

Here, A denotes face image, and A_i represents i^{th} face image captured from i^{th} video I . Each face image corresponds to individual person such that face image of a person is captured at different locations. The face image A_i is captured from dataset and is fed to pre-processing phase to generate pre-processed result.

3.1.2 Pre-processing of face image

The input face image A_i is fed to pre-processing module, where image is pre-processed to remove noise from images. Due to noise present in images, the raw images are ineffective for analysis, hence pre-processing is essential for eliminating the noise. At this phase, the external artifacts of image are eliminated. Moreover, pre-processed face image is represented as P that is passed to feature extraction module to extract the features.

3.1.3 Feature extraction from pre-processed image using LDTOOP

It is necessary to capture the features from pre-processed result to reduce the dimension of features. Here, feature extraction process is made using developed LDTOOP feature, which is derived by integrating Local optimal oriented pattern (LOOP) [41] and Local directional ternary pattern (LDTP) [42], respectively. Let us consider the intensity of pre-processed image P at pixel (u_r, v_r) as t_r and t_c be the pixel intensity in $[3\times3]$ neighborhood of (u_r, v_r). The LOOP value for pixel (u_r, v_r) is represented as,

$$LP(u_r, v_r) = \sum_{c=1}^{7} \varpi(t_c - t_r).\, 2^{\tau_c} \tag{3}$$

Here, $\varpi(u)$is represented as,

$$\varpi(u) = \begin{cases} 1; u \geq 0 \\ 0; Otherwise \end{cases} \tag{4}$$

The LOOP descriptor is used to encode the rotation invariance to main formulation. It is specifically used to eliminate empirical assignment of parameter of Local directional pattern (LDP). The LDTP integrates two types of compass masks, namely Gaussian mask, and Frei-Chen masks. The Gaussian masks are employed to reduce noise perturbation to make the model more robust with respect to illumination variations, whereas Frei-Chen mask is used to enhance encoded structural details. The LDTP computes edge responses by eight directions based on two masks to encode the texture of image. However, the peripheral pixel is compared with central pixel to encode the contrast information. The LDTP code is designed by employing Local ternary pattern (LTP) model. The LDTP is effectively used to classify the local primitives and to extract better information, as edge reactions are low sensitive to noise and illumination than the intensity. The central pixel of $[3 \times 3]$ square neighborhood specifies texture and intensity variations of structure. The LDTP operator considers relative edge response of central pixel. Let us consider peripheral pixel as t_e $(e = 0to7)$and central pixel as t_d , respectively. However, average local gray level of $[3 \times 3]$ square neighborhood is given as,

$$\lambda = \frac{1}{9}\left[t_d + \sum_{e=0}^{7} t_e\right] \tag{5}$$

Moreover, the concept of LTP is employed to generate LDTP code. With LTP, each ternary pattern is partitioned into namely positive as well as negative part for generating LDTP code. Here, kernel function specifies lower and upper LDTP that is given as,

$$LDTP_{lower} = 2^8 \rho(\lambda - t_d, ED_d) + \sum_{e=0}^{7} 2^e \rho(t_e - t_d, ED_e) \tag{6}$$

$$LDTP_{upper} = 2^8 \hat{\rho}(\lambda - t_d, ED_d) + \sum_{e=0}^{7} 2^e \hat{\rho}(t_e - t_d, ED_e) \tag{7}$$

where, ρ and $\hat{\rho}$ is given as,

$$\rho(u, v) = \begin{cases} 1; u \geq 0 and v \geq 0 \\ 0; Otherwise \end{cases} \tag{8}$$

$$\hat{\rho}(u, v) = \begin{cases} 1; u \leq 0 and v \leq 0 \\ 0; Otherwise \end{cases} \tag{9}$$

The histogram from upper and lower LDTP scale analysis are fused together to generate the feature and is incorporated with the LOOP descriptor to build the feature f with the dimension of $[1 \times 10]$, respectively.

3.1.4 Facial emotion recognition using DBN

The DBN classifier is used to achieve emotion recognition classification based on features acquired from face image. The feature f is employed as input to DBN [43] classifier to accomplish process of facial emotion recognition. The structure of DBN contains two RBM layers, and a single MLP layer. Here, connection exists among visible and hidden neurons. Moreover, output generated by the previous layer is fed as input to the next layer. The advantage of using DBN classifier is that it requires small sized dataset, and it consumes less training time. Figure 2 portrays structure of DBN.

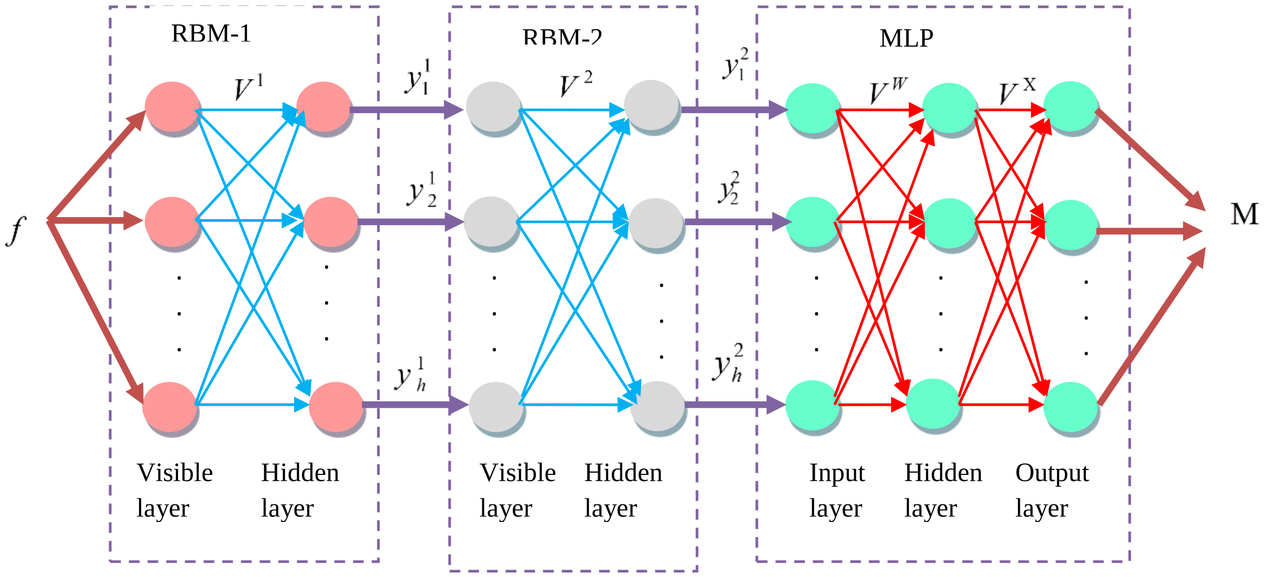


Fig.2. Architecture of DBN

The feature f is subjected as input to visible layer such that output of hidden layer of RBM layer-1 is given as,

$$y^1 = \{y_1^1, y_2^1, \dots, y_l^1, \dots, y_h^1\}; 1 \le j \le h \tag{10}$$

where, y_j^1 indicates j^{th} hidden neuron and h indicates total count of hidden neurons. Here, individual neuron associated with visible and hidden layer holds a bias. Let us consider bias of the visible and the hidden layer as,

$$Q^1 = \{Q_1^1, Q_2^1, \dots, Q_U^1, \dots, Q_\beta^1\} \tag{11}$$

$$T^1 = \{T_1^1, T_2^1, \dots, T_j^1, \dots, T_h^1\} \tag{12}$$

where, Q_U^1 signifies the bias correspond to U^{th} visible neuron and T_j^1 indicates bias of j^{th} hidden neuron. Moreover, weight initialized to RBM layer-1 is indicated as,

$$V^1 = \{V_{Uj}^1\}; 1 \le U \le \beta; 1 \le j \le h \tag{13}$$

where, V_{Uj}^1 implies weight among U^{th} visible neuron and j^{th} hidden neuron.

The output compute at RBM layer-1 is indicated as,

$$y_j^1 = \delta\left[T_j^1 + \sum_U f_U^1 V_{Uj}^1\right] \tag{14}$$

where, δ indicates activation function. The count of visible neurons at layer-2 is equal to count of hidden neurons of layer-1. However, the bias associated with the visible and hidden layer of RBM layer-2 is signified as Q^2, and T^2, respectively.

Accordingly, weight factor connected with RBM layer-2 is given as,

$$V^2 = \{V_{jj}^2\}; 1 \le j \le h \tag{15}$$

where, V_{jj}^2 signifies weight amongst j^{th} visible neuron and j^{th} hidden neuron. Accordingly, output computed at RBM layer-2 is represented as,

$$y_j^2 = \delta\left[T_j^2 + \sum_U f_U^2 V_{jj}^2\right] \forall f_U^2 = y_j^1 \tag{16}$$

where, T_j^2 signifies the bias connected with j^{th} hidden neuron. However, input of MLP is given as,

$$z = \{z_1, z_2, \dots, z_j, \dots, z_h\} = \{y_j^2\}; 1 \leq j \leq h \tag{17}$$

where, h indicates total count of neurons at input layer. Let the weight connected amongst input and hidden layer is given as,

$$V^W = \{V_{jY}^W\}; 1 \leq j \leq h; 1 \leq Y \leq \gamma \tag{18}$$

where, V_{jY}^W specifies weight among j^{th} input neuron and Y^{th} hidden neuron. By considering the weight and bias, output of hidden layer is computed as,

$$O_Y = \left[\sum_{j=1}^{h} V_{jY}^W * z_j\right]\eta_Y \forall z_j = y_j^2 \tag{19}$$

where, η_Y indicates bias of hidden neuron. Accordingly, output generated at RBM layer-2 is given as input to the MLP layer. Accordingly, weight connected among hidden and the output layer is specified as V^X and is shown as,

$$V^X = \{V_{Yo}^X\}; 1 \leq Y \leq \gamma; 1 \leq o \leq \Upsilon \tag{20}$$

With the weight and the output of the hidden layer, the output vector is expressed as,

$$M_o = \sum_{Y=1}^{\gamma} V_{Yo}^X * O_Y \tag{21}$$

where, V_{Yo}^X indicates weight among Y^{th} hidden neuron and o^{th} output neuron and O_Y signifies output of hidden layer.

3.2 Acquisition of EEG signal and physiological signal

EEG signal is created by central nervous system and quickly respond to the emotional variations. The EEG signals offered significant features for the emotion recognition. For each person, the EEG and the physiological signal are collected for processing emotion recognition system. However, the captured EEG signal that correspond to i^{th} person is represented as E. Similarly, the physiological signal collected from i^{th} person of video I_i is specified as H, respectively.

3.2.1 Feature extraction from EEG and physiological signal

The input EEG signal E and physiological signal H are allowed to acquire features from signal to perform process of emotion recognition. Some of the features extracted from the signal E and H includes wavelet coefficients, spectral feature that comprises power spectral density (PSD), tonal power ratio, spectral flux, spectral skewness, and spectral centroid.

i) Wavelet coefficients: It is connected in small neighborhood. Accordingly, large wavelet coefficient encompasses high coefficients at the neighbors such that the extracted wavelet coefficients is specified as g_1 with the dimension of $[1 \times 50]$.

ii) Spectral features: Some of the spectral features acquired from EEG signal and physiological signal are PSD, tonal power ratio, spectral flux, spectral skewness, and spectral centroid.

PSD: It is used to compute power distribution of input signal based on certain frequency range and is represented as g_2 with the dimension of $[1 \times 50]$.

Tonal power ratio: It is an important feature used to analyze speech signal. It is the factor utilized to compute spectrum of signal and it is termed as the ratio of tonal power of spectrum factors to overall power. This feature is specified as g_3 with the size of $[1 \times 1]$.

Spectral flux: This feature is termed as spectral correlation among adjacent windows. However, it specifies degree of change of spectrum among windows and is represented as g_4 with the dimension of $[1 \times 1]$, respectively.

Spectral skewness: It is the factor of asymmetry of probability distribution that is close to the mean value. This feature is denoted as g_5 with the size of $[1 \times 1]$.

Spectral centroid: It signifies where center of mass of spectrum and also it computes the brightness of sound. However, this feature is represented as g_6 with the size of [1 × 1], respectively.

The feature vector employed to Deep RNN classifier is represented as,

$$g = \{g_1, \ldots, g_6\} \tag{22}$$

Here, g represents the feature vector with the dimension of [1 × 104]that contains six different features acquired from EEG and physiological signal.

The feature vector g is subjected as input to Deep RNN classifier to accomplish emotion recognition classification.

3.2.2 Facial emotion recognition using Deep RNN

The input takes the Deep RNN classifier is feature vector y extracted from both the EEG signal and physiological signal. Deep RNN [44] is network structure, which comprises different recurrent hidden layers. The major concept of this classifier is the presence of recurrent connection at hidden layer. It operates under varying input length with respect to sequence of information. Accordingly, the previous state knowledge is fed as input to next state to generate classification result. The recurrent structure makes the classifier to work in an effective way based on the features. Deep RNN is more effective in processing the features acquired from signals. Figure 3 portrays the structure of Deep RNN.

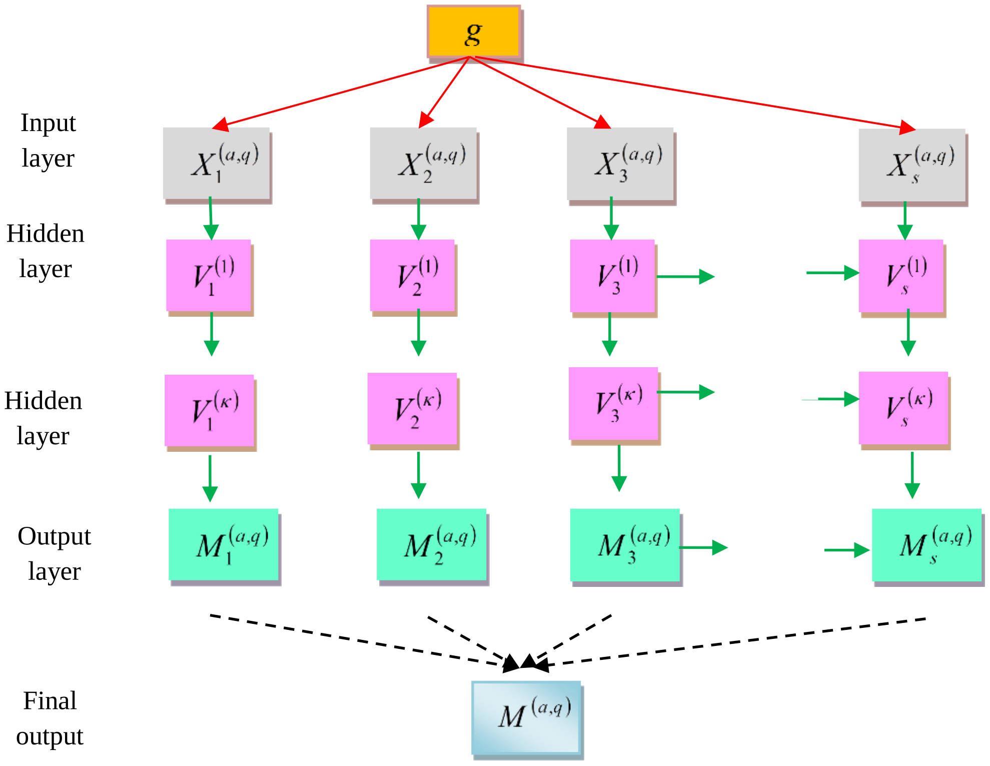


Fig.3. Architecture of Deep RNN classifier

The architecture of Deep RNN is illustrated with input of a^{th} layer at q^{th} time as $X^{(a,q)} = \left\{X_1^{(a,q)}, X_2^{(a,q)}, \ldots X_p^{(a,q)}, \ldots X_s^{(a,q)}\right\}$ and output of a^{th} layer at q^{th} time as $M^{(a,q)} = \left\{M_1^{(a,q)}, M_2^{(a,q)}, \ldots M_p^{(a,q)}, \ldots M_s^{(a,q)}\right\}$, respectively. Here, input as well output pairs are called as unit. Moreover, p indicates arbitrary unit count, and s specifies number of units of a^{th} layer. Moreover, arbitrary unit count of $(a-1)^{th}$ layer is represented as p, and number of units of $(a-1)^{th}$ layer is denoted as U. Accordingly, input propagation weight from $(a-1)^{th}$ layer to a^{th} layer is represented as, $K^{(a)} \in C^{s \times U}$, respectively. Where, C shows set of weights. Accordingly, components of input vector is given as,

$$X_p^{(a,q)} = \sum_{b=1}^{U} w_{pb}^{(a)} M_b^{(a-1,q)} + \sum_{p'}^{s} v_{pp'}^{(a)} M_{p'}^{(a,q-1)} \tag{23}$$

where, $w_{pb}^{(a)}$ and $v_{pp'}^{(a)}$ specifies the elements of $K^{(a)}$ and $k^{(a)}$. Here, p' signifies arbitrary unit count of a^{th} layer. Output vector elements of a^{th} layer is given as,

$$M_p^{(a,q)} = \chi^{(a)}\left(X_p^{(a,q)}\right) \tag{24}$$

where, $\chi^{(a)}$ signifies activation function. Moreover, activation functions employed here are sigmoid function as $\chi(X) = tanh(X)$, rectified linear unit function (ReLU) as$\chi(X) = max(X, \ell)$, and the logistic sigmoid function as $\chi(X) = \frac{1}{(1+e^{-X})}$. To simplify the process, let us introduce ℓ^{th} weight as $w_{p\ell}^{(a)}$ and ℓ^{th} unit as $M_{\ell}^{(a-1,q)}$ and hence, the bias is specified as,

$$M^{(a,q)} = \chi^{(a)}.\left(K^{(a)}M^{(a-1,q)} + k^{(a)}.M^{(a,q-1)}\right) \tag{25}$$

Here, $M^{(a,q)}$ indicates output of classifier.

3.3 Proposed Bat Rider Optimization Algorithm

The training process of both the deep learning classifier named DBN and Deep RNN are carried out with proposed optimization algorithm named BROA, which is the integration of ROA [45] and BA [46], respectively. BA is operated using echolocation characteristics of bats. It uses the echolocation for sensing the distance and it recognizes the distance among background barriers and the food in a magical way. The bats randomly fly with the velocity to the position by considering some fixed frequency, loudness and varying wavelength for searching the prey. It adjusts the wavelength automatically of the emitted pulses and alters the pulse emission rate based on proximity of target. ROA considers four rider groups, namely bypass rider, follower, attacker and overtaker. Each rider group follows some plans to move towards target to win the race. With success rate, leading rider is selected at each time instant. It employs the concept of fictional computing to solve the optimization problem by considering the thought and imaginary ideas. By incorporating the parametric features of both optimizations, the performance of classification result is improved. The steps involved in proposed BROA are explained as follows:

i) Initialization: The first step is to define four groups of riders as R such that their positions are randomly initialized as given below,

$$G_x = \{G_x(m,k)\}; 1 \leq m \leq \omega; 1 \leq k \leq \sigma \tag{26}$$

Here, ω indicates number of riders and is equivalent to R, σ represent number of coordinates, and $G_x(m,k)$ specifies location of m^{th} rider at the time interval x. Let us define rider parameters, steering angle as SA, gear as GR, accelerator as AL, and brake as BK, respectively.

ii) Compute fitness measure: The fitness function is utilized to find optimal solution for the optimization problem. The fitness with minimum error value is declared as best solution such that fitness measure is represented as,

$$F = \frac{1}{N}\sum_{\varepsilon=1}^{N}[O_\varepsilon - L_\varepsilon]^2 \tag{27}$$

Here, N indicates total count of samples, O_ε signifies target output, and L_ε indicates output of deep learning classifier such that $L \in \{\mathrm{M}, M\}$.

iii) Update solution of riders: The rider updates their position to identify leading rider based on the characteristic features.

a) Update equation of bypass rider: The bypass rider commonly bypasses rider path and it does not follow path of leading rider in such a way that position update equation is given as,

$$G_{x+1}^{by}(m,k) = \alpha\left[G_x(\mu,k) * R_1(k) + G_x(R_2,k) * [1 - R_1(k)]\right] \tag{28}$$

where, α and R_1 implies random number that lies in the interval of $[0,1]$, μ and R_2 indicates random number with the range of $[1 to\, \omega]$, respectively.

b) Update equation of follower: The follower follows the location of leading rider to update its position in order to reach the target. Moreover, position update equation is represented as,

$$G_{x+1}^{fl}(m,l) = G^{LR}(LR,l) + \left[cos\left(SA_{m,l}^{x}\right) * G^{LR}(LR,l) * S_{m}^{x}\right] \tag{29}$$

$$G_{x+1}^{fl}(m,l) = G^{LR}(LR,l)\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) \tag{30}$$

The standard equation of BA is expressed as,

$$G_{x+1}^{fl}(m,l) = G_{x}^{fl}(m,l) + B_{x}(m,l) \tag{31}$$

$$G_{x+1}^{fl}(m,l) = G_{x}^{fl}(m,l) + B_{x+1}(m,l) - \left(G_{x}^{fl}(m,l) - G^{*}\right)J(m,l) \tag{32}$$

$$G_{x+1}^{fl}(m,l) = G_{x}^{fl}(m,l) + B_{x+1}(m,l) - G_{x}^{fl}(m,l)J(m,l) + G^{*}J(m,l) \tag{33}$$

$$G^{*}J(m,l) = G_{x+1}^{fl}(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l) + G_{x}^{fl}(m,l)J(m,l) \tag{34}$$

$$G^{*} = \frac{G_{x+1}^{fl}(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l) + G_{x}^{fl}(m,l)J(m,l)}{J(m,l)} \tag{35}$$

By substituting the above Eq. (35) in Eq. (30) in place of leading rider position is expressed as,

$$G_{x+1}^{fl}(m,l) = \frac{G_{x+1}^{fl}(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l) + G_{x}^{fl}(m,l)J(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) \tag{36}$$

$$G_{x+1}^{fl}(m,l) = \frac{G_{x+1}^{fl}(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) - \frac{G_{x}^{fl}(m,l) + B_{x+1}(m,l) - G_{x}^{fl}(m,l)J(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) \tag{37}$$

$$G_{x+1}^{fl}(m,l) - \frac{G_{x+1}^{fl}(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) = \frac{G_{x}^{fl}(m,l)J(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) \tag{38}$$

$$G_{x+1}^{fl}(m,l)\left[1 - \frac{1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}}{J(m,l)}\right] = \frac{G_{x}^{fl}(m,l)J(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) \tag{39}$$

$$G_{x+1}^{fl}(m,l)\left[\frac{J(m,l) - 1 - cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}}{J(m,l)}\right] = \frac{G_{x}^{fl}(m,l)J(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right) \tag{40}$$

$$G_{x+1}^{fl}(m,l) = \frac{J(m,l)}{J(m,l) - 1 - cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}}\left[\frac{G_{x}^{fl}(m,l)J(m,l) - G_{x}^{fl}(m,l) - B_{x+1}(m,l)}{J(m,l)}\left(1 + cos\left(SA_{m,l}^{x}\right) * S_{m}^{x}\right)\right] \tag{41}$$

where, $J(m,l) = J(Jmin_{max}R_{3})_{min}$, R_{3} indicates random number that ranges from 0 to 1, l denotes coordinate selector, G^{LR} represents the location of leading rider, LR represents index of leading rider, $SA_{m,l}^{x}$ specifies steering angle of m^{th} rider in l^{th} coordinate, and S_{m}^{x} implies distance to be travelled by m^{th} rider.

c) Update solution of overtaker: The overtaker updates the position with the factors, like fitness rate, coordinate selector, and direction indicator. Hence, position update equation of overtaker is given as,

$$G_{x+1}^{ot}(m,l) = G_{x}(m,l) + [Z_{x}(m) * G^{LR}(LR,l)] \tag{42}$$

where, $G_{x}(m,l)$represents location of m^{th} rider, and Z represents direction indicator of m^{th} rider, respectively.

d) Update equation of attacker: The attacker tries to take the location of leader and is given as,

$$G_{x}^{att}(m,k) = G^{LR}(LR,k) + \left[cos\left(SA_{m,k}^{x}\right) * G^{LR}(LR,k) * S_{m}^{x}\right] \tag{43}$$

where, $G^{LR}(LR,k)$indicates the location of leading rider, $SA_{m,k}^{x}$ represents steering angle of m^{th} rider in k^{th} coordinate, and S_{m}^{x} denotes distance to be travelled by m^{th} rider.

iv) Evaluating feasibility: The fitness measure is measured for each solution to generate the best solution in such a way that the fitness function with minimal error value is declared as optimal solution.

v) Termination: the above steps are repeated until best solution is generated. Algorithm 1 portrays the pseudo code of proposed BROA-based deep learning model.

Algorithm 1. Pseudo code of proposed BROA-based deep learning

Sl. No	Pseudo code of proposed BROA-based deep learning
1	**Input:** G_x
2	**Output:** Leading rider
3	Begin
4	Initialize the population
5	Define the parameters, steering angle SA, gear GR, accelerator AL, and brake BK
6	Compute fitness function
7	while $\left(x < x_{off}\right)$; x_{off} --off time
8	for $(m = 1 to \, \omega)$
9	Update solution of bypass rider using Eq. (28)
10	Update the position of follower using Eq. (41)
11	Update the equation of overtaker using Eq. (42)
12	Update solution of attacker using Eq. (43)
13	Rank the rider
14	Select leading rider
15	Update SA, GR, AL, and BK
16	Return best solution
17	$x = x + 1$
18	end for
19	end while
20	end

3.4 Renyi-entropy based fusion model

The final step of proposed method is fusion process, where the output computed by DBN classifier and the output generated by the Deep RNN classifier are fused together to generate the emotion recognition results, as sadness, happiness, anger, disgust, fear, and surprise using Renyi-entropy based fusion. The Renyi entropy of L of order ℓ is given as,

$$H = \frac{1}{1-\ell} ln(L) \tag{44}$$

Here, $L \in \{M, M\}$, and H indicates Renyi entropy.

4. Results and Discussion

This section explains the discussion of results of proposed BROA-based deep learning method with respect to performance measures.

4.1 Experimental setup

The implementation of proposed scheme is done in PYTHON tool with windows 10 OS, and 4 GB RAM using DEAP dataset [40].

4.2 Dataset description

This dataset contains two parts, namely ratings from online self-assessment, and participant ratings. It is the publicly available dataset used to emotion classification. This dataset contains different physiological recordings, EEG signals, and face video. It contains 40 music videos. The EEG and the physiological signals are recorded such that each participant rated the videos. The face video is recorded for 22 participants.

4.3 Evaluation metrics

The performance of proposed model is analyzed with the measures, like accuracy, FAR, and FRR.

Accuracy: It is the factor that shows the computation of accurate closeness of measurements to total count of observations. It is represented as,

$$AY = \frac{x_p + x_n}{x_p + x_n + y_p + y_n} \tag{45}$$

where, x_p represents true positive rate, x_n specifies true negative, y_p indicates false positive, and y_n represents false negative, respectively.

False acceptance rate (FAR): It is the factor that specifies the ratio of false acceptances to total number of identification results.

False rejection rate (FRR): It is defined as ratio of total count of false recognition with the total number of recognized results.

4.4 Experimental results

Figure 4 portrays experimental results of developed method. Figure 4 a) and 4 b) depicts input face image-1, and face image-2. Figure 4 c) and 4 d) portrays EEG signal of image-1, and image-2. The PSD feature extracted from image-1 and image-2 is portrayed in figure 4 e) and 4 f). Figure 4 g) and 4 h) represents the wavelet coefficients extracted from image-1 and image-2, respectively.

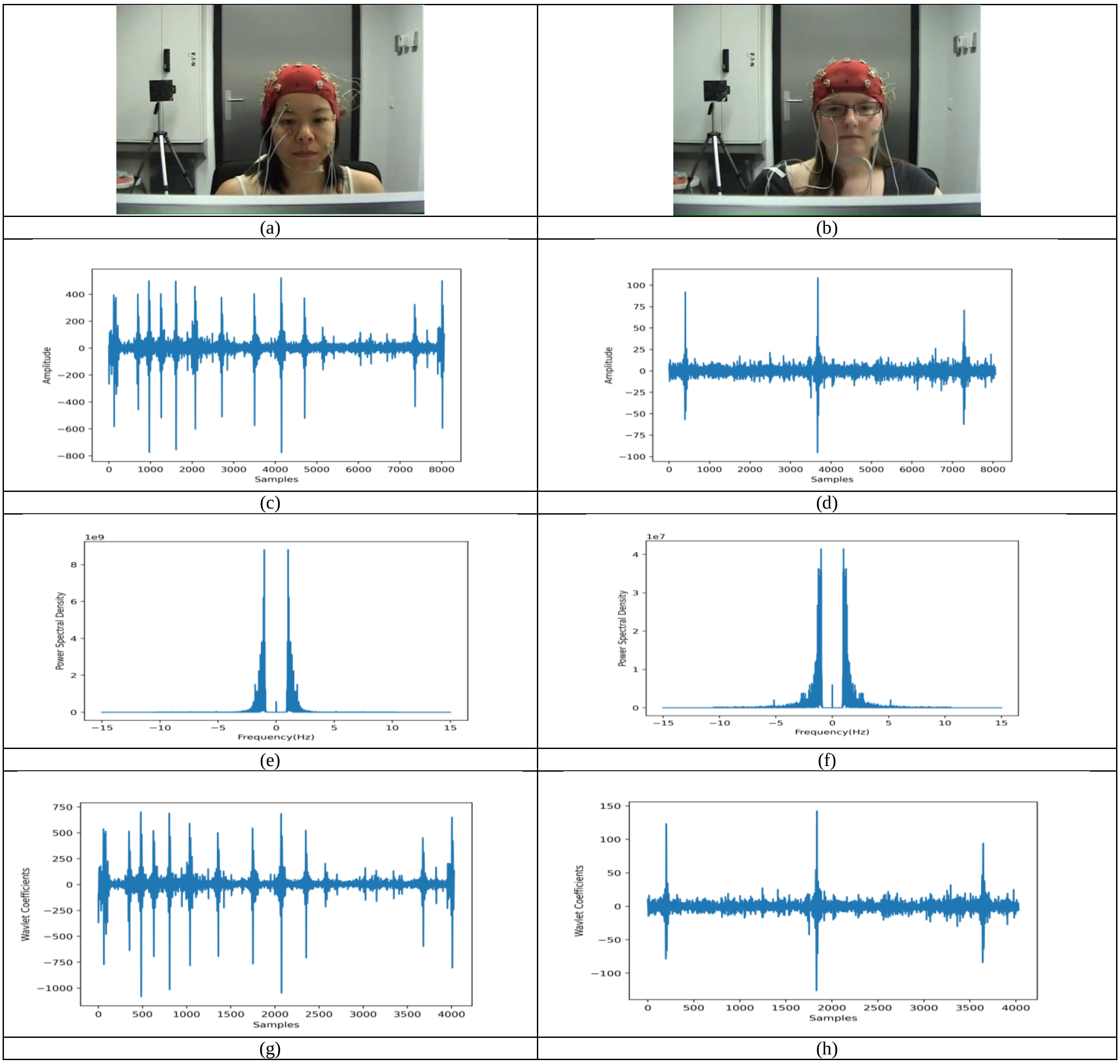


Fig.4. Experimental results, a) input face image-1, b) input face image-2, c) EEG signal of image-1, d) EEG signal of image-2, e) extracted PSD feature of image-1, f) extracted PSD features from image-2, g) wavelet coefficients of image-1, h) wavelet coefficients of image-2

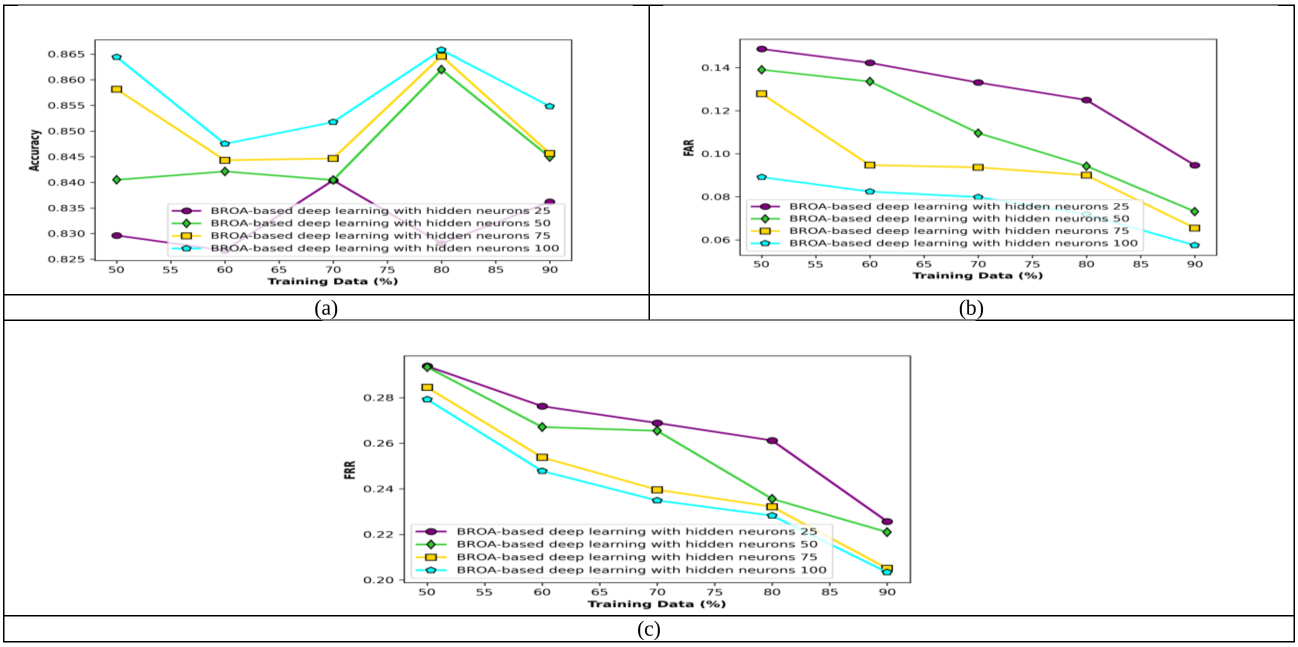
4.5 Performance analysis

This section presents the performance analysis of developed method based on training data by varying number of neurons, learning rate, and features.

a) Analysis based on hidden neurons

Figure 5 depicts the analysis by varying hidden neurons. Figure 5 a) represents analysis with accuracy. At training data is assumed as 60%, the accuracy of proposed BROA-based deep learning by considering hidden neurons 25 is 0.8267, hidden neurons 50 is 0.8421, hidden neurons 75 is 0.8443, and hidden neurons 100 is 0.8475. At 70% of training value, accuracy measured by proposed method with hidden neurons 25 is 0.8404, hidden neurons 50 is 0.8404, hidden neurons 75 is 0.8447, and hidden neurons 100 is 0.8517. At 80% of training value, accuracy measured by proposed method with hidden neurons 25 is 0.8281, hidden neurons 50 is 0.8620, hidden neurons 75 is 0.8646, and hidden neurons 100 is 0.8658.

The analysis made by FAR is represented in figure 5 b). At 70% of training data, FAR computed by proposed method by considering the hidden neurons 25 is 0.133, hidden neurons 50 is 0.110, hidden neurons 75 is 0.094, and hidden neurons 100 is 0.080. By considering 80%, the FAR measured by the proposed BROA-based deep learning by considering hidden neurons 25 is 0.1249, hidden neurons 50 is 0.0942, hidden neurons 75 is 0.09, and with hidden neurons 100 is 0.0718. At 90% of training data, FAR computed by proposed method by considering the hidden neurons 25 is 0.0946, hidden neurons 50 is 0.0731, hidden neurons 75 is 0.0654, and hidden neurons 100 is 0.0574.



(a) (b)

(c)

Fig.5. Analysis with hidden neurons, a) accuracy, b) FAR, c) FRR

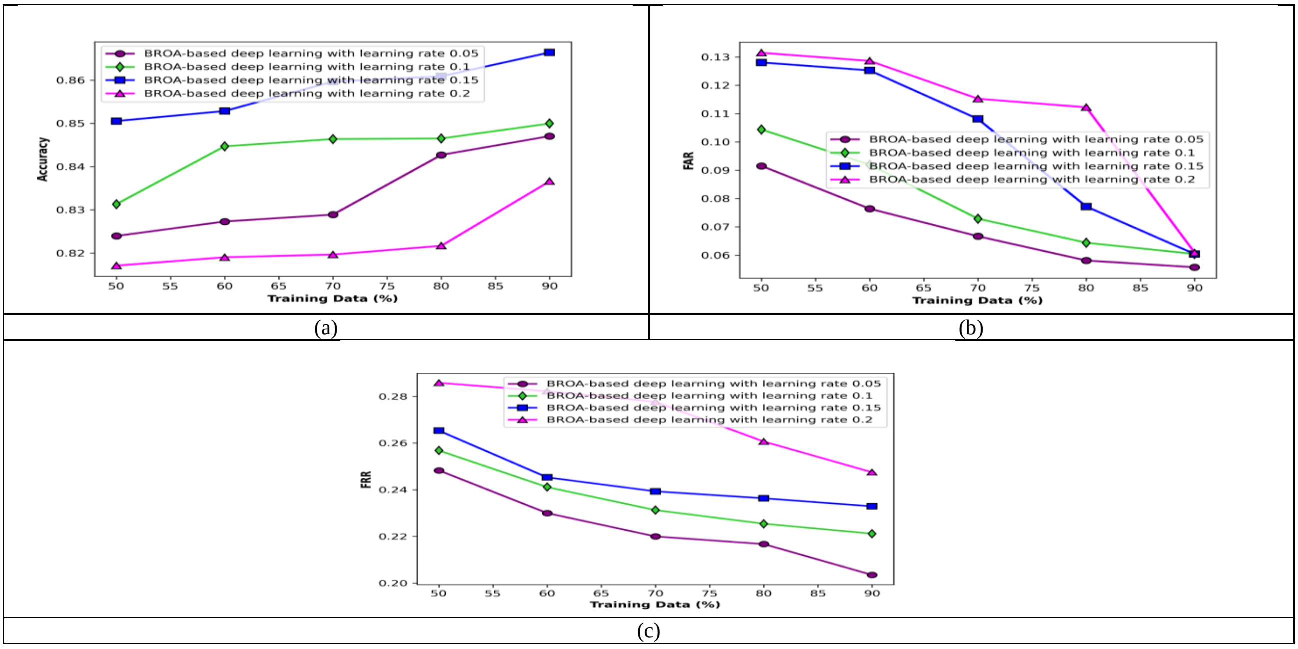
Figure 5 c) portrays analysis of FRR. By considering 70%, the FRR measured by the proposed BROA-based deep learning by considering hidden neurons 25 is 0.2689, hidden neurons 50 is 0.2655, hidden neurons 75 is 0.2396, and hidden neurons 100 is 0.2349. At 80% of training value, FRR obtained by BROA-based deep learning method with hidden neurons 25 is 0.2612, hidden neurons 50 is 0.2356, hidden neurons 75 is 0.2322, and hidden neurons 100 is 0.2283. At 90% of training value, FRR obtained by BROA-based deep learning method with hidden neurons 25 is 0.2257, hidden neurons 50 is 0.2210, hidden neurons 75 is 0.2051, and hidden neurons 100 is 0.2034.

b) Analysis with learning rate

Figure 6 represents analysis of proposed method with the learning rate. Figure 6 a) portrays the analysis of accuracy. At 70% training value, accuracy of BROA-based deep learning by considering learning rate 0.05 is 0.8289, learning rate 0.1 is 0.8464, learning rate 0.15 is 0.8596, and with learning rate 0.2 is 0.8196. When considering the training data as 80%, accuracy of proposed BROA-based deep learning with learning rate 0.05 is 0.8427, learning rate 0.1 is 0.8465, learning rate 0.15 is 0.8609, and with learning rate 0.2 is 0.8217.

The analysis made with the FAR measure is portrayed in figure 6 b). By considering 80% training value, accuracy attained by proposed BROA-based deep learning with learning rate 0.05 is 0.0581, learning rate 0.1 is 0.0644, learning rate 0.15 is 0.0772, and with learning rate 0.2 is 0.1122. When considering the training data as 90%, accuracy of proposed BROA-based deep learning with learning rate 0.05 is 0.0557, learning rate 0.1 is 0.0604, learning rate 0.15 is 0.0604, and with learning rate 0.2 is 0.0609.

Figure 6 c) depicts the analysis carried out using FRR measure. When training data is considered as 70%, accuracy of proposed BROA-based deep learning with learning rate 0.05 is 0.2200, learning rate 0.1 is 0.2313, learning rate 0.15 is 0.2393, and with learning rate 0.2 is 0.2777. When considering the training value as 80%, accuracy measured by the proposed BROA-based deep learning with learning rate 0.05 is 0.2167, learning rate 0.1 is 0.2254, learning rate 0.15 is 0.2363, and with learning rate 0.2 is 0.2607.



(a)

(b)

(c)

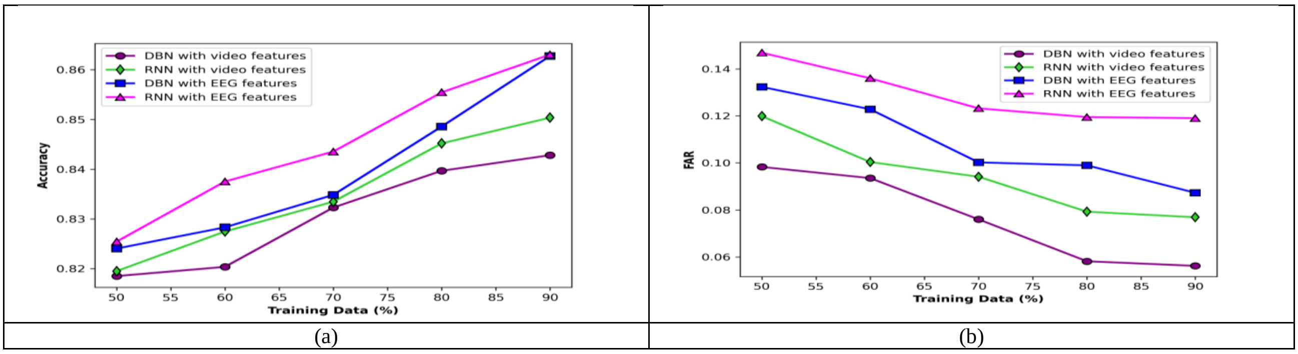
Fig.6. Performance analysis with learning rate, a) accuracy, b) FAR, c) FRR

c) Analysis based on features

Figure 7 represents the analysis of developed scheme by varying features. Figure 7 a) depicts the analysis carried out using accuracy measure. By considering 80% of training value, accuracy measured by DBN with video features is 0.8397, RNN with video features is 0.8452, DBN with EEG features is 0.8486, and RNN with EEG features is 0.8555. At 90% of training value, accuracy achieved by DBN with video features is 0.8428, RNN with video features is 0.8504, DBN with EEG features is 0.8628, and RNN with EEG features is 0.8631.

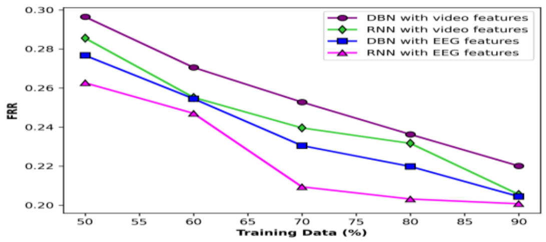
The analysis made by FAR metric is represented in figure 7 b). At 80% of training value, FAR achieved by DBN with video features is 0.0581, RNN with video features is 0.0792, DBN with EEG features is 0.0990, and RNN with EEG features is 0.1195. At 90% of training value, FAR achieved by DBN with video features is 0.0562, RNN with video features is 0.0769, DBN with EEG features is 0.0873, and RNN with EEG features is 0.1191.

Figure 7 c) depicts the analysis made using FRR measure. By considering training value as 80%, FRR obtained by DBN with video features is 0.2363, RNN with video features is 0.2317, DBN with EEG features is 0.2198, and RNN with EEG features is 0.2031. With the 90% of training value, FRR achieved by DBN with video features is 0.2201, RNN with video features is 0.2055, DBN with EEG features is 0.2044, and RNN with EEG features is 0.2007.



(a)

(b)



(c)

Fig.7. Analysis based on features, a) accuracy, b) FAR, c) FRR

4.6 Comparative methods

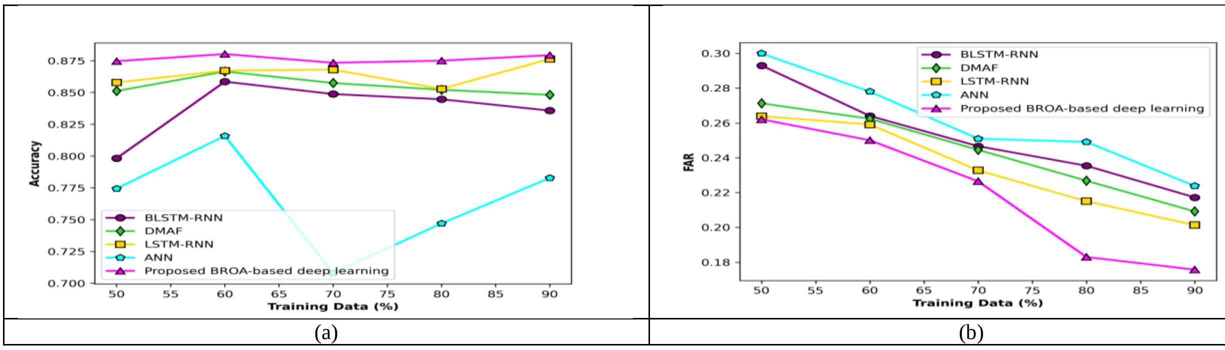
The performance of proposed method is analyzed by considering the existing methods, like Attention-based Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN) [33], Neural Network (NN) [4], LSTM-RNN [30], and Deep Multimodal Attentive Fusion (DMAF) [1][14], respectively.

4.7 Comparative analysis

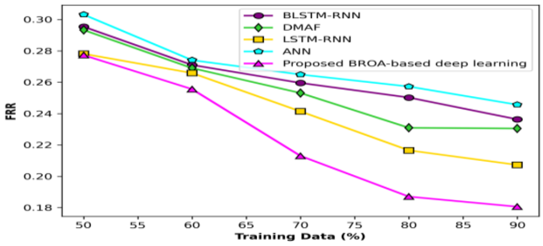
Figure 8 portrays comparative analysis of proposed BROA-based deep learning framework. The analysis carried out by considering the accuracy measure is represented in figure 8 a). When the training value is considered as 60%, accuracy computed by BLSTM-RNN, DMAF, LSTM-RNN, ANN, and proposed BROA-based deep learning is 0.8586, 0.8665, 0.8672, 0.8159, and 0.8803 that results the performance of improvement while comparing the developed scheme with the traditional BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 2%, 2%, 1%, and 7%, respectively. The accuracy measured by proposed BROA-based deep learning method at 80% of training data is 0.8447, whereas the existing BLSTM-RNN, DMAF, LSTM-RNN, and ANN computed the accuracy of 0.8522, 0.8529, 0.7471, and 0.8751. By assuming training data to 90%, accuracy measured by BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 0.8358, 0.8482, 0.8765, and 0.7829, whereas proposed BROA-based deep learning achieved the accuracy of 0.8794 which outcomes the performance improvement with that of BLSTM-RNN, DMAF, and ANN is 5%, 4%, and 11%.

Figure 8 b) portrays the analysis of FAR metric. By considering 60% of training data, FAR computed by BLSTM-RNN, DMAF, LSTM-RNN, ANN, and proposed BROA-based deep learning is 0.2641, 0.2624, 0.2591, 0.278, and 0.2501 such that it reports the performance enhancement with that of BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 5%, 5%, 3%, and 10%, respectively. When increasing the training data to 80%, FAR measured by existing methods, such as BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 0.2354, 0.2268, 0.2150, and 0.2491 , whereas the proposed BROA-based deep learning method computed the FAR of 0.1830 in such a way that it outcomes the performance improvement while comparing the proposed with traditional BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 22%, 19%, 15%, and 27%.

The analysis made using the FRR measure is depicted in figure 8 c). At 70% of training data, FRR obtained by BLSTM-RNN, DMAF, LSTM-RNN, ANN, and proposed BROA-based deep learning is 0.2595, 0.2531, 0.2414, 0.2649, and 0.2130 that shows the performance enhancement with BLSTM-RNN is 18%, DMAF is 16%, LSTM-RNN is 12%, and ANN is 20%, respectively. When training data is increased to 80%, FAR measured by conventional techniques, such as BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 0.2502, 0.2309, 0.2165, and 0.2572, whereas the proposed BROA-based deep learning method attained the FRR of 0.1870 in such a way that it outcomes the performance improvement while comparing the developed method with that of BLSTM-RNN, DMAF, LSTM-RNN, and ANN is 25%, 19%, 14%, and 27%. With the deep learning classifier, the proposed method achieves higher performance based on the features acquired from different modalities.



(a) (b)



(c)

Fig.8. Comparative analysis, a) accuracy, b) FAR, c) FRR

4.8 Comparative discussion

Table 1 portrays the comparative discussion of proposed model. The below table depicts the performance values computed by the existing and proposed method by considering the training value of 90%. The accuracy computed by BLSTM-RNN, DMAF, LSTM-RNN, ANN, and proposed BROA-based deep learning is 0.8358, 0.8482, 0.8765, 0.7829, and 0.8794, respectively. However, the FAR obtained by existing BLSTM-RNN, DMAF, LSTM-RNN, ANN, and proposed BROA-based deep learning is 0.2172, 0.2092, 0.2014, 0.2238, and 0.1757. Moreover, the FRR achieved by BLSTM-RNN, DMAF, LSTM-RNN, ANN, and proposed BROA-based deep learning is 0.2363, 0.2305, 0.2072, 0.2456, and 0.1806.

Table 1. Comparative discussion

Metrics/Methods	**BLSTM-RNN**	**DMAF**	**LSTM-RNN**	**ANN**	**Proposed BROA-based deep learning**
Accuracy	0.8358	0.8482	0.8765	0.7829	0.8794
FAR	0.2172	0.2092	0.2014	0.2238	0.1757
FRR	0.2363	0.2305	0.2072	0.2456	0.1806

5. Conclusion

In this research, an efficient method named BROA-based deep learning is developed for facial emotion recognition using multimodal signals. However, the proposed BROA is developed by the integration of BA and ROA, respectively. Here, different modalities considered for the process of emotion recognition includes face image, EEG signal, and physiological signal. The face image is pre-processed more effectively to eliminate the noise exist in images. Thereafter, the feature extraction is done using LDTOOP such that the features acquired from face image are processed by DBN classifier. Similarly, the EEG and the physiological signals for the respective face image of the person is allowed to extract the features such that the extracted features include wavelet coefficients, and spectral feature that comprises PSD, tonal power ratio, spectral skewness, spectral flux, and spectral centroid, respectively. The deep learning classifier is employed to achieve emotion recognition process more effectively. The proposed method achieves higher performance by obtaining maximum accuracy of 0.8794, and minimum FAR and minimum FRR of 0.1757 and 0.1806, respectively. In future, the research work can include some other optimization algorithm for training the deep learning classifier.

References

[1] Wang F, Sahli H, Gao J, Jiang D, Verhelst W, “Relevance units machine based dimensional and continuous speech emotion prediction”, Multimedia Tools and Applications, vol.74, no.22, pp.9983–10000, 2015.
[2] Drummond PD, Quah SH, “The effect of expressing anger on cardiovascular reactivity and facial blood flow in Chinese and Caucasians”, Psychophysiology, vol.38, pp.190–196, 2001.
[3] Richard Jiang, Anthony T.S.Ho, Ismahane Cheheb, NoorAl-Maadeed, SomayaAl-Maadeed, and Ahmed Bouridane, "Emotion recognition from scrambled facial images via many graph embedding", Pattern Recognition, vol.67, pp.245-251, July 2017.
[4] Gilsang Yoo, Sanghyun SeoSungdae Hong Hyeoncheol Kim , "Emotion extraction based on multi bio-signal using back-propagation neural network ", Multimedia Tools and Applications, vol.77, no.4, pp.4925–4937, February 2018.
[5] Pantic M, Caridakis G, Andre E, Kim J, Karpouzis K, Kollias S, “Multimodal emotion recognition from low-level cues”, In Emotion Oriented Systems, 2011.
[6] Dhall A, Goecke R, Lucey S, Gedeon T, “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark”, In Proceedings of the International Conference on Computer Vision Workshops, pp 2106–2112, 2011.
[7] Eleftheriadis S, Rudovic O, Pantic M, “Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition”, IEEE Transactions on Image Processing, vol.24, no.1, pp.189–204, 2015.
[8] Baixi Xing , Hui Zhang, Hui Zhang, Lekai Zhang, Xinda Wu, Xiaoying Shi, Shanghai Yu, and Sanyuan Zhang, "Exploiting EEG Signals and Audiovisual Feature Fusion for Video Emotion Recognition", IEEE Access, vol.7, pp.59844 - 59861, 03 May 2019.

[9] R. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[10] Kolodyazhniy, V.; Kreibig, S.D.; Gross, J.J.; Roth, W.T.; Wilhelm, F.H, “An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions”, Psychophysiology, vol.48, 908–922, 2011.

[11] Liu, Y.-J.; Yu, M.; Zhao, G.; Song, J.; Ge, Y.; Shi, Y, “Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals”, IEEE Transactions on Affective Computing, vol.9, no.4, pp.550–562, 2017.

[12] Menezes, M.L.R.; Samara, A.; Galway, L.; Sant’Anna, A.; Verikas, A.; Alonso-Fernandez, F.; Wang, H.; Bond, R, “Towards emotion recognition for virtual environments: an evaluation of EEG features on benchmark dataset”, Personal and Ubiquitous Computing, vol. 21, pp.1003–1013, 2017.

[13] Rania M. Ghoniem , Abeer D. Algarni, and Khaled Shaalan, " Multi-Modal Emotion Aware System Based on Fusion of Speech and Brain Information", Information vol.10, no.7, 2019.

[14] Jianzhu Guo ; Zhen Lei ; Jun Wan ; Egils Avots ; Noushin Hajarolasvadi ; Boris Knyazev ; Artem Kuharenko, Julio C. Silveira Jacques , Xavier Bar, Hasan Demirel, Sergio Escalera, J üri Allik, and Gholamreza Anbarjafari, "Dominant and Complementary Emotion Recognition From Still Images of Faces," in IEEE Access, vol. 6, pp. 26391-26403, 2018.

[15] B. Draper, K. Baek, M. Bartlett, J. Beveridge, “Recognizing faces with PCA and ICA”, Computer Vision Image Understanding, Vol.91, no.1-2, pp.115, 2003.

[16] M. H. Yang, “Kernel Eigenfaces vs. kernel Fisherface: face recognition using kernel methods”, International Conference on Automatic Face and Gesture Recognition, pp.215, 2002.

[17] B. Tenenbaum, V. Silva, J. Langford, “A global geometric framework for nonlinear dimensionality”, Science, Vol.290, No.5500, pp.2319, 2000.

[18] X. He, S. Yan, Y. Hu, P. Niyogi, H. J. Zhang, “Face Recognition Using Laplacianfaces”, IEEE Trans. Pattern Analysis & Machine Intelligence, Vol. 27, No. 3, pp.1, Mar. 2005.

[19] Xiaofei He, Deng Cai and Partha Niyogi, "Tensor Subspace Analysis", Advances in Neural Information Processing Systems 18 (NIPS), Vancouver, Canada, Dec. 2005.

[20] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S, ” Multimodal fusion for multimedia analysis: a survey”, Multimedia Systems, vol.16, no.6, pp.345–379, 2010.

[21] Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J., “Recognition of emotions using multimodal physiological signals and an ensemble deep learning model”, Computer Methods and Programs in Biomedicine, vol.140, pp.93–110, 2016.

[22] Tengfei Song ; Wenming Zheng ; Cheng Lu ; Yuan Zong ; Xilei Zhang ; Zhen Cui, "MPED: A Multi-Modal Physiological Emotion Database for Discrete Emotion Recognition", IEEE Access, vol.7, pp.12177 - 12191, 09 January 2019.

[23] Poh, N.; Bengio, S. How do correlation and variance of base-experts affect fusion in biometric authentication tasks? IEEE Trans. Signal Process. 2005, 53, 4384–4396.

[24] Liu, Y.-J.; Yu, M.; Zhao, G.; Song, J.; Ge, Y.; Shi, Y. Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals. IEEE Trans. Affect. Comput. 2018, 9, 550–562.

[25] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song, " Multi-modal Emotion Recognition using Semi-supervised Learning and Multiple Neural Networks in the Wild", In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp.529–535, November 2017.

[26] Muhammad Adeel Asghar, Muhammad Jamil Khan, Fawad , Yasar Amin, Muhammad Rizwan, MuhibUr Rahman , Salman Badnava, andSeyed Sajad Mirjavadi, " EEG-Based Multi-Modal Emotion Recognition using Bag of Deep Features: An Optimal Feature Selection Approach", Sensors, vol.19, no.23, 2019.

[27] Zheng, W.L., Zhu, J.Y., Peng, Y., Lu, B.L.: EEG-based emotion classification using deep belief networks. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2014).

[28] Hamester D, Barros P, Wermter S, “Face expression recognition with a 2-channel Convolutional Neural Network”, In 2015 International Joint Conference on Neural Networks (IJCNN), pp 1–8, 2015.

[29] He, L., Jiang, D., Yang, L., Pei, E., Wu, P., Sahli, H.: Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: International Workshop on Audio/Visual Emotion Challenge, pp.73–80 (2015)

[30] Jiamin Liu, Yuanqi Su, Yuehu Liu, " Multi-modal Emotion Recognition with Temporal-Band Attention Based on LSTM-RNN", Pacific Rim Conference on Multimedia, vol.10735, pp.194-204, 10 May 2018.

[31] Kim Y, Lee H, Provost EM, “Deep learning for robust feature generation in audiovisual emotion recognition”, In 2013 I.E. International Conference on Acoustics, Speech and Signal Processing, pp 3687–3691, 2013.

[32] A. Sherly Alphonse and Dejey Dharma, " Novel directional patterns and a Generalized Supervised Dimension Reduction System (GSDRS) for facial emotion recognition", Multimedia Tools and Applications, vol.77, no.8, pp.9455–9488, April 2018.

[33] ChaoLia, Zhongtian Bao, Linhao Li, and ZipingZhao, " Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition", Information Processing & Management, vol.57, no. 3, May 2020.

[34] Kuan Tung ; Po-Kang Liu ; Yu-Chuan Chuang ; Sheng-Hui Wang ; An-Yeu Andy Wu, "Entropy-Assisted Multi-Modal Emotion Recognition Framework Based on Physiological Signals", In proceedings of the IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 3-6 Dec. 2018.

[35] Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. Knowl.-Based Syst. 2019, 167, 26–37.

[36] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, “Multi-layer temporal graphical model for head pose estimation in real-world videos,” in IEEE International Conference on Image Processing, pp. 3392–3396, 2015.

[37] S. Jain, C. Hu, and J. K. Aggarwal, “Facial expression recognition with temporal modeling of shapes,” in IEEE International Conference on Computer Vision Workshops, pp. 1642–1649, 2011.

[38] R. Ali, “Depth camera-based facial expression recognition system using multilayer scheme,” Iete Technical Review, vol. 31, no. 4, pp. 277–286, 2014.

[39] M. Valstar, M. Pantic, and I. Patras, “Motion history for facial action detection in video,” in IEEE International Conference on Systems, Man and Cybernetics, vol.1, pp. 635–640, 2004.

[40] DEAPdataset, "https://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html", accessed on February 2020.

[41] Tapabrata Chakraborti, Brendan McCane, Steven Mills, and Umapada Pal, " LOOP Descriptor: Encoding Repeated Local Patterns for Fine-grained Visual Identification of Lepidoptera", October 2017.

[42] Chahi, A., Ruichek, Y. and Touahni, R., “Local directional ternary pattern: A new texture descriptor for texture classification”, Computer vision and image understanding, vol.169, pp.14-27, 2018.

[43] Lv, Z. and Qiao, L., “Deep belief network and linear perceptron based cognitive computing for collaborative robots”, Applied Soft Computing, vol.92, pp.106300, 2020.

[44] Inoue, M., Inoue, S. and Nishida, T., “Deep recurrent neural network for mobile human activity recognition with high throughput”, Artificial Life and Robotics, vol.23, no.2, pp.173-185, 2018.

[45] D. Binu and B. S Kariyappa, " RideNN: A New Rider Optimization Algorithm-Based Neural Network for Fault Diagnosis in Analog Circuits", IEEE Transactions on Instrumentation and Measurement, vol.68, no.1, pp. 2 – 26, January 2019.

[46] Yang, X.S., “A new metaheuristic bat-inspired algorithm”, In Nature inspired cooperative strategies for optimization (NICSO 2010), Springer, Berlin, Heidelberg, pp. 65-74, 2010.

Authors’ Profiles

Jaykumar M. Vala is currently pursuing Ph.D. in Computer/IT Engineering at Gujarat Technological University, Chandkheda, Gujarat, India. He has completed his B.E. from Birla Vishvakarma Mahavidyalaya Engineering College, Vallabh Vidyanagar, Gujarat, India. He has completed his M.E. from L.D. Engineering College, Ahmedabad Gujarat. He is currently serving as an Assistant Professor in G H Patel College of Engineering and Technology, Vallabh Vidyanagar, Gujarat. His research interests include Computer Vision, Machine Learning and Deep Learning.

Prof. (Dr.) Udesang K. Jaliya, is a leading researcher & academician in Computer Engineering and holds Ph.D. degree. He is currently working as an Assistant Professor in the Department of Computer Engineering at Birla Vishvakarma Mahavidyalaya Engineering College, Vallabh Vidyanagar, Gujarat, India. He is a member of Computer Society of India. He has published 35 international articles. He has guided 28 research projects at master degree. He is currently guiding 4 Ph.D. students. His research interests include Image Processing, Machine Learning, and Deep Learning.

How to cite this paper: Jaykumar M. Vala, Udesang K. Jaliya, "Deep Learning Network and Renyi-entropy Based Fusion Model for Emotion Recognition Using Multimodal Signals", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.4, pp. 67-84, 2022.DOI: 10.5815/ijmecs.2022.04.06