International Journal of Communication Networks and Information Security

*ISSN: **2073-607X, 2076-0930***

*Volume **14** Issue **02 Year 2022** Page **155**:**167***

Multimodal Sentiment Sensing and Emotion Recognition Based on Cognitive Computing Using Hidden Markov Model with Extreme Learning Machine

[*1]Ms. Diksha Verma, [2] Sweta Kumari Barnwal, [3] Dr. Amit Barve,
[4] Dr M K Jayanthi Kannan, [5] Mr. Rajesh Gupta, [6] R. Swaminathan

[1]Assistant Professor, Department of Computer Science Engineering, Chandigarh Engineering College, Jhanjeri, India*
[2] Assistant Professor, Department of Computer Science, ARKA JAIN University, Jamshedpur, Jharkhand, India
[3] Associate Professor, Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat, India
[4] Professor, HOD Professor and HOD of Information Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-To-Be University), Bangalore, India
[5] Pro Chancellor, Department of Management, Sanskriti University, Mathura, Uttar Pradesh, India
[6] Professor, Department of Electrical, Electronics and Communications, Galgotias University, Greater Noida, Uttar Pradesh, India
[1]diksha.j1629@cgc.ac.in, [2]sweta.b@arkajainuniversity.ac.in,
[3]amit.barve17535@paruluniversity.ac.in, [4]k.jayanthi@jainuniversity.ac.in,
[5]chancellor@sanskriti.edu.in, [6]r.swaminathan@galgotiasuniversity.edu.in

Article History

Received: 11 April 2022
Revised: 26 July 2022
Accepted: 30 August 2022

Abstract

In today's competitive business environment, exponential increase of multimodal content results in a massive amount of shapeless data. Big data that is unstructured has no specific format or organisation and can take any form, including text, audio, photos, and video. Many assumptions and algorithms are generally required to recognize different emotions as per literature survey, and the main focus for emotion recognition is based on single modality, such as voice, facial expression and bio signals. This paper proposed the novel technique in multimodal sentiment sensing with emotion recognition using artificial intelligence technique. Here the audio and visual data has been collected based on social media review and classified using hidden Markov model based extreme learning machine (HMM_ExLM). The features are trained using this method. Simultaneously, these speech emotional traits are suitably maximised. The strategy of splitting areas is employed in the research for expression photographs and various weights are provided to each area to extract information. Speech as well as facial expression data are then merged using decision level fusion and speech properties of each expression in region of face are utilized to categorize. Findings of experiments show that combining features of speech and expression boosts effect greatly when compared to using either speech or expression alone. In terms of accuracy, recall, precision, and optimization level, a parametric comparison was made.

Keywords: multimodal content, emotion recognition, social media review, HMM_ExLM, decision level fusion

1. Introduction

In a globe of 7.8 billion people, 50.64 percent of the population, regardless of age, uses social media. This population currently uses popular social networking sites such as Facebook, Instagram, YouTube, Facebook Messenger, WhatsApp, Twitter and Reddit. Furthermore, microblogging services such as Twitter, Instagram, and Reddit are frequently utilised social media platforms where people make short, frequent posts [1]. Through text, audio, image and video contributions, OSM (online social media) platforms allow users to express, discuss, and share their opinions, thoughts, views and perspectives. Social media posts are public and full of emotions. Analysing and analysing these social media posts may reveal emotional states as well as the reasons behind those emotions. However, due to the huge amount of data, this analysis is extremely challenging [2]. Artificial Intelligence can assist in the automated detection of emotions, feelings, personal attributes, viewpoints, and their effects on social trends. Emotions play a role in nearly every decision we make and every minute of our life. As a result, detecting emotions piques our curiosity, because understanding how others feel allows us to interact with them more successfully. This capacity allows a system, such as (CS) Conversational Systems and ECAs (Embodied Conversational Agents), to react to these events and alter their behaviours [3]. There are currently systems that can distinguish particular emotions (or impairments) and can aid in diagnosis of specific diseases as well as improve patient treatment. Automobile safety is another important application of facial expression recognition. Recognizing negative emotions like tension, rage, or exhaustion is critical for intelligent vehicles to minimise traffic accidents and promote road safety by permitting them to respond to driver's state. Emotions are important in future 'Next Revolution,' which will necessitate creation of more social robots. To display closer personal relationships between humans and technology, these robots will require to recognise people's emotions and convey and produce their emotional states [4]. Other physiological factors, such as trust, will also be important in defining the personality of these machines. Deep learning has had remarkable success in several fields in recent years, including signal processing, artificial intelligence, and emotion recognition, by depending on the most modern technology. The most widely used deep learning algorithms are DBN (deep belief networks), CNN (convolution neural networks) and RNN (recurrent neural networks). Natural language processing techniques are generally used in text-based emotion and sentiment mining, and researchers have had some success collecting sentiment data from news, forums, consumer reviews, and social media (Twitter) [5]. Affective computing relies heavily on emotion recognition. It is a multidisciplinary field that includes computer science, artificial intelligence, psychology, and cognitive neuroscience. Facial expressions, voice, behaviour, and physiological signals can all be used to identify human moods. The first three methods of emotion recognition, on the other hand, are somewhat subjective. For example, the people under investigation may purposefully hide their genuine feelings, which could affect their performance. Emotion recognition based on physiological cues, on the other hand, is more dependable and objective [6].

The contribution of this paper is as follows:

- To collect the online social media dataset which consist of audio and visual data in recognition of multimodal sentiment and emotion.
- Here the classification of audio and visual data has been carried out using hidden Markov model based extreme learning machine (HMM_ExLM).
- From the classification output the audio features and image features has been classified. Then by fusing both the features using decision level fusion technique, the emotion and sentiment of the features has been analysed and decision is made in recognising whether the sentiment and emotion are same.
- The experimental results show the extracted features of audio data, visual data and the result showing the fused data. Comparative analysis has been made in terms of accuracy, precision, recall and optimization level of fusion.

2. Related Works

Although state-of-the-art SER also called as speech emotion recognition systems have low accuracy and high computational cost for a long time [7]. The author built a lightweight CNN design with plain rectangular kernels as well as improved pooling layers that obtained state-of-the-art IEMOCAP and EMO-DB datasets performance [8]. To solve this problem, some research, such as [9], employed partial face images with only mouth movements to categorize emotions by utilizing transfer-learning on pre-trained ImageNet models. On the study of four emotions: neutral, joyful, startled, and furious, their examination revealed a loss of accuracy of only 5% when compared to version using entire face image. Textual data is another source of data that is becoming more important in real-time systems. Many publications have appeared since introduction of transformers [10] and BERT models [11], owing to advantages of utilizing natural language, such as smaller size of text files compared to audio or images. As a result of these benefits, this modality are used for a variety of tasks, including sentiment and emotion recognition in various publications [12], demonstrating the modality's versatility. [13] investigated the integration of visual as well as textual data in speech emotion identification via a hybrid fusion technique known as a multimodal attention network (MMAN). They propose cLSTM-MMA, a new multimodal focus mechanism that enhances attention across three modalities and selectively integrates information. cLSTM-MMA is fused with other unimodal subnetworks during late fusion. Findings show that visual and textual clues help tremendously in recognising speech emotions. The proposed cLSTM-MMA alone achieves the same level of precision as existing fusion techniques while having a significantly smaller network configuration. [14] investigated the use of a "BERT-like" design for SSL to represent both languages well as text modalities to understand multimodal language emotions. They show how a simple fusion mechanism can simplify overall structure while also strengthening sophisticated fusion mechanisms. [15] described a deep learning-based method for safeguarding emotion-related codes. The researchers were able to extract acoustic features from raw audio using a SincNet layer, band-pass filtering and NN and output of those band-pass filters was then applied to input to DCNN. On the N-gram level, a collection of representations is first determined in a bidirectional RNN, then in another RNN utilising cross attention, and finally integrated as a final score. [16] proposed a new method for merging a convolutional neural network based on raw waveforms with cross-modal focus. Their prototypes show that suggested methods are capable of achieving cutting-edge emotional classifications. [17] discovered that SVM-based technique to ML is effective for voice customer sentiment analysis through experiments. Proposed a multimodal speech emotion identification system based on SVM. The experimental results show that by applying this SVM method to common database classification issue, SVM algorithm has evolved considerably. Finally, employs method to comprehend emotional expression as well as achieves emotional recognition through effective speaking. [17] created a multimodal deep learning model that incorporates facial photos as well as textual details to understand the situation. To categorise the characters' face expressions in Korean TV show. They created two multimodal models to identify emotions using photos and text. The results of the experiment showed that using text definitions of character behaviour greatly improves recognition performance. In terms of audio, the use of LSTM network for classification is recommended and classification effect is substantially boosted when compared to other machine learning approaches. A novel multimodal music emotion method was created based on music audio quality as well as text for music lyrics [18]. Bert is offered as a way to represent the feelings of lyrics in terms of lyrics, which effectively tackles long-term dependency. In terms of multimodal fusion, LSFM is mentioned in the lyrics. Emotion dictionary is utilized to change how lyrics are classified emotionally. NN is developed using stage fusion with linear weighted decision making, which improves efficiency as well as precision.

3. Proposed Multimodal Sentiment And Emotion Recognition Using Classification And Feature Fusion Techniques

This section discusses about the novel design in sentiment and emotion recognition using hidden Markov model with extreme learning machine-based data classification. From classification outputs, the audio features and visual feature has been fused using decision level feature fusion. The overall proposed architecture is shown in figure-1.

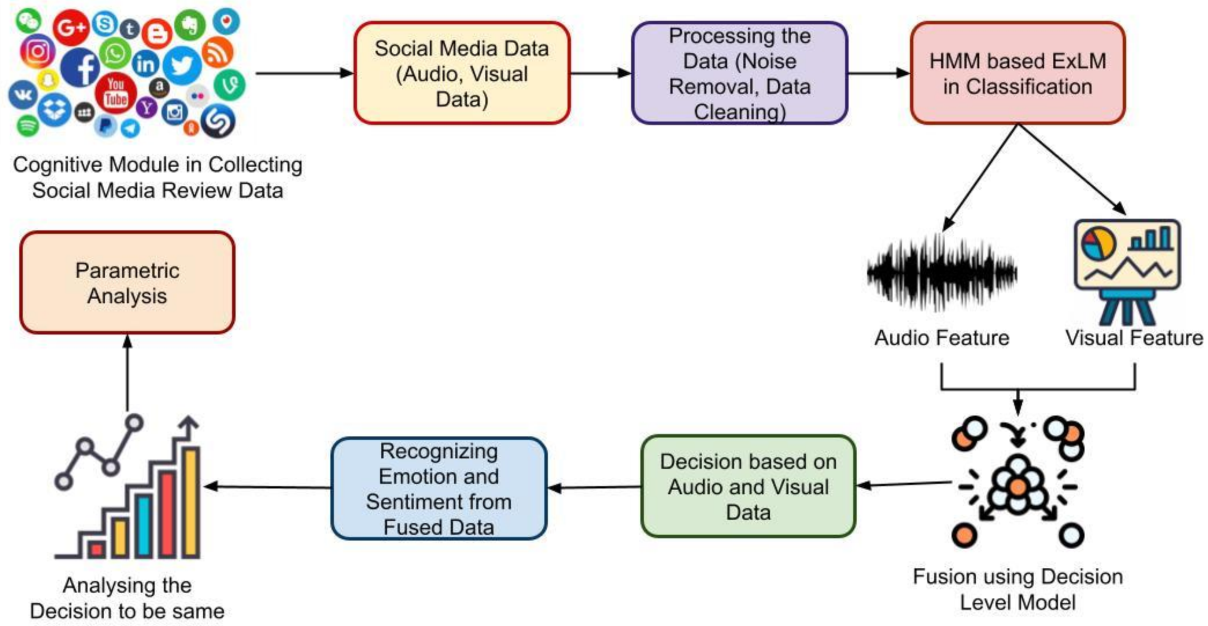

Cognitive Module in Collecting Social Media Review Data
Social Media Data (Audio, Visual Data)
Processing the Data (Noise Removal, Data Cleaning)
HMM based ExLM in Classification
Audio Feature
Visual Feature
Fusion using Decision Level Model
Decision based on Audio and Visual Data
Recognizing Emotion and Sentiment from Fused Data
Analysing the Decision to be same
Parametric Analysis

Figure 1 Overall proposed architecture

3.1 Multimodal data classification using hidden Markov model with extreme learning machine (HMM-ExLM):

HMM is a doubly stochastic process that is hidden from observation as well as observable method that is found by it. Model is defined by following elements in terms of a first-order hidden Markov process. Set of hidden states is given by eqn (1)

$$S = \{S_1, S_2, \cdots, S_N\} \tag{1}$$

where N is number of states in model.
State transition probability distribution is given by eqn (2)

$$\mathrm{A} = \{a_{ij}\} \tag{2}$$

where, for $1 \le i, j \le N$,

$$a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i] \\ \begin{cases} 0 \le a_{ij} \\ \sum_{j=1}^{N} a_{ij} = 1 \end{cases} \tag{3}$$

Set of observation symbols is given by eqn (4)

$$V = \{v_1, v_2, \cdots, v_M\} \tag{4}$$

Observation symbol probability distribution is given by eqn (5)

$$B = \{b_j(k)\} \tag{5}$$

where for $1 \le j \le N, 1 \le k \le M$

$$b_j(k) = P[v_k \text{ at } t \mid q_t = S_j] \\ \begin{cases} 0 \le b_j(k) \\ \sum_{k=1}^{M} b_j(k) = 1 \end{cases} \tag{6}$$

Initial state probability distribution is given by eqn (7)

$$\pi = \{\pi_i\} \tag{7}$$

where for $1 \leq i \leq N$,

$$\tilde{\pi}_i = P[q_1 = S_i] \\ \begin{cases} 0 \leq \pi_i \\ \sum_{i=1}^{N} \pi_i = 1. \end{cases} \tag{8}$$

In the following discussion, an HMM will be referred to as a triplet is given by eqn (9).

$$\lambda = (A, B, \pi) \tag{9}$$

- Form:

$$a_t(i) = P(o_2 o_2 \cdots o_4, q_t = S_k \mid \lambda) \tag{10}$$

$\alpha_t(i)$is solved in actively:

1. initialization:

$$\alpha_1(i) = \pi_2 b_1(o_1), \quad 1 \leq i \leq N \tag{11}$$

and
2. induction:

$$a_{t+1}(j) = \mid \sum_{i=1}^{N} a_t(t) a_i j b_j (o_{t+1})_t \tag{12}$$

- back to and variables

$$\beta_t(i) = P(\sigma_1 + 1O_4 + 2 \cdots \sigma_T \mid q_k = S_i, \lambda)_n \tag{13}$$

$\beta_r(t)$is solved inductively by eqn (14)
1 initialization:

$$\beta_r(i) = 1, \quad 1 \leq i \leq N \tag{14}$$

and
2. induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j (o_{2+1}) \beta_{1+1}(j) \tag{15}$$

- observation covariation is given by eqn (16)

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_i(i), \beta_1(i), \forall t \tag{16}$$

especially,

$$P(O \mid \lambda) = \sum_{i=1}^{N} \propto r(i) \tag{17}$$

auxiliary function is given by eqn (18)

$$Q(\lambda, \lambda) = \sum_{Q} P(O, Q \mid \lambda) \log P(O, Q \mid \lambda) \tag{18}$$

Where λ is auxiliary variable with respect toλ.
When value of $Q(\lambda, \lambda)$is high, value of $\dot{P}(O \mid \lambda)$high as well is given by eqn (19)

$$Q(\lambda, \lambda) \geq Q(\lambda, \lambda) \longrightarrow P(O \mid \lambda) \geq P(O \mid \lambda)$$

$$\frac{\partial P(O|\lambda)}{\partial \lambda_i} = \left. \frac{\partial Q(\lambda, \bar{\lambda})}{\partial \lambda_i} \right|_{\lambda = \lambda}, 1 \leq i \leq D \tag{19}$$

First one of these variables is shown in eqn (20):

$$\xi_{\mathrm{t}}(i, j) = p[q_t = i, q_{t+1} = j \mid O, \lambda] \tag{20}$$

Which is given by eqn (21)

$$\xi_e(i,j)=\frac{p[q_t=i,q_{t+1}=j,O\langle\lambda\}}{p\{O(\lambda)}\tag{21}$$

Forward variable αt(i) is given by eqn (22)

$$\alpha_{\mathrm{t}}(i)=p\big(O_1,O_{2,\cdots,}O_t q_z=it\lambda\big)\tag{22}$$

Where O1, O2 …, OT are partial ranking sequences represented by eqn(23):

$$\begin{aligned}&\alpha_{t+1}(j)=c_j(o_{+1})\textstyle\sum_{l=1}^{N}\alpha_i(t)o_y,1\le j\le N_,1\le t\le T-1\\&\alpha_1(j)=\pi_j\varphi(o_2)_4\quad 1\le j\le N\end{aligned}\tag{23}$$

Backward variable βt(i) is given by eqn (24)

$$\beta_i(i)\colon O_{T+1},O_{r+2},\dots,\hat{O}_r\tag{24}$$

If current state is $i,\beta_i(i)$ is probability of partial ranking sequence $\beta_i(i)\colon O_{T+1},O_{r+2},\dots,\hat{O}_r,\beta_i(i)$is evaluated by utilizing following eqn (25):

$$\beta_1(t)=\textstyle\sum_{j=1}^{N}\beta_{7+1}(j)a_0e_j(o_{1+1})_,\quad 1\le i\le N,1\le t\le T-1\tag{25}$$

Where

$$\beta_1(i)=1,\quad 1\le i\le N\tag{26}$$

Evaluate$\xi_t(i,j)$ variable by utilizing forward as well as backward variables:

$$\xi_k(i,j)=\frac{a_i(i)a_{\mathrm{ij}}\beta_{i+1}(j)e_j(o_{1+1})}{\sum_{j=1}^{N}\sum_{j=1}^{N}a_1(i)a_{ij}\beta_{2+1}(j)e_j(o_{u+1})}\tag{27}$$

Posteriori probability is given by eqn (28),

$$\gamma(i)=p[q_{\mathrm{A}}=i\mid O,\lambda]\tag{28}$$

This can be stated in both forward as well as backward variables in eqn (29)

$$\gamma(i)=\left[\frac{a_r(i)\beta_2(i)}{\sum_{k=1}^{N}\alpha_t(i)\beta_1(i)}\right]\tag{29}$$

Relationship between $\gamma_t(i)$ and $\xi_t(i,j)$ is denoted by eqn (30)

$$\gamma_i(i)=\textstyle\sum_{j=1}^{N}\xi_i(i,j),\quad 1\le i\le N,1\le t\le M\tag{30}$$

Let's have a look at a set of pattern class observation sequences.

$$\begin{aligned}&O=\{O^{(1)},O^{(2)},\cdots,O^{(K)}\}\\&O^{(\mathrm{k})}=o_1^{(k)}\omega_2^{(k)}\cdots o_{T_1}^{(k)},1\le k\le K\end{aligned}\tag{31}$$

Normally, there is no way of knowing if these observation sequences are independent of one another. Furthermore, assuming independence property when these observation sequences are statistically connected can lead to a contradiction. Without losing generality, we have the following phrases in any scenario.

$$\begin{cases}P(O\mid\lambda)= & P\big(O^{(1)}\mid\lambda\big)P\big(O^{(2)}\mid O^{(1)},\lambda\big)\dots\\ & P\big(O^{(K)}\mid O^{(K-1)}\dots O^{(1)},\lambda\big)\\ P(O\mid\lambda)= & P\big(O^{(2)}\mid\lambda\big)P\big(O^{(1)}\mid O^{(2)},\lambda\big)\cdots\\ & P\big(O^{(1)}\mid O^{(K)}\cdots O^{(2)},\lambda\big)\\ \qquad\vdots & \\ P(O\mid\lambda)= & P\big(O^{(K)}\mid\lambda\big)P\big(O^{(1)}\mid O^{(K)},\lambda\big)\cdots\\ & P\big(O^{(K-1)}\mid O^{(K)}O^{(K-2)}\cdots O^{(1)},\lambda\big).\end{cases}\tag{32}$$

Multiple observation probability is given by eqn (33).

$$P(O \mid \lambda)=\sum_{i=1}^{N} \alpha_i(i), \beta_1(i), \forall t \tag{33}$$

Where

$$\begin{cases} w_1 = & \frac{1}{K} P\left(O^{(2)} \mid O^{(1)}, \lambda\right) \ldots P\left(O^{(K)} \mid O^{(K-1)} \ldots O^{(1)}, \lambda\right) \\ w_2 = & \frac{1}{K} P\left(O^{(3)} \mid O^{(2)}, \lambda\right) \cdots P\left(O^{(1)} \mid O^{(K)} \ldots O^{(2)}, \lambda\right) \\ \quad \vdots & \\ w_K = & \frac{1}{h} P\left(O^{(1)} \mid O^{(K)}, \lambda\right) \cdots P\left(O^{(K-1)} \mid O^{(K)} O^{(K-2)} \ldots \right. \\ & \left. O^{(1)}, \lambda\right) \end{cases} \tag{34}$$

Create an auxiliary function for model training based on the given expression is represented by eqn (35).

$$Q(\lambda, \bar{\lambda})=\sum_{k-1}^{K} w_k Q_k(\lambda, \bar{\lambda}) \tag{35}$$

where$\bar{\lambda}$ is auxiliary variable with respect to λshown in eqn (36)

$$Q_{\perp}(\lambda, \bar{\lambda})=\sum_Q P\left(O^{(k)}, Q \mid \lambda\right) \log P\left(O^{(k)}, Q \mid \bar{\lambda}\right), \quad 1 \leq k \leq K \tag{36}$$

Let's have a look at the auxiliary function in the context of boundary conditions.

$$\begin{gathered} \mathrm{Q}(\lambda, \bar{\lambda})=\sum_{k-1}^{K} w_k Q_k(\lambda, \bar{\lambda}) \\ 1-\sum_{j-1}^{N} \bar{a}_{ij}=0, \quad 1 \leq i \leq N \\ 1-\sum_{d-1}^{M} \bar{b}_j(k)=0, \quad 1 \leq j \leq N \\ 1-\sum_{n-1}^{-\mathrm{N}} \pi_i=0 \end{gathered} \tag{37}$$

The Lagrange multiplier approach can be used to create an objective function.

$$\begin{aligned} F(\hat{\lambda})= \quad & Q(\lambda, \hat{\lambda})+\sum_{i=1}^{N} c_u\left[1-\sum_{j=1}^{N} \bar{a}_{ij}\right] \\ & +\sum_{j=1}^{N} \alpha_j\left[1-\sum_{i=1}^{M} \bar{b}_j(k)\right]+c_r\left[1-\sum_{i=1}^{N} \pi_i\right]. \end{aligned} \tag{38}$$

where$c_{\Delta i}, \omega_{j,}$ and c_t are Lagrange multipliers.

For N distinct samples $x_i \in R_N \times R_j, y_i \in R_N \times R_m (i=1,2, . . .,)$, Numerical link between hidden layer's output and output of output layer may be described as (1) and hidden layer's outputs is calculated as (39).

$$\begin{gathered} h=g(a x+b) \\ h(x_i) V=y_i, \quad i=1,2, \ldots, N \\ h_L(\mathbf{x}_j)=\sum_{i=1}^{L} \beta_t G(\mathbf{w}_i, b_i, \mathbf{x}_j)=\mathbf{t}_j, \quad j=1,2, \ldots, P \end{gathered} \tag{39}$$

$G(w_i, b_i, x_j)$is an activation function that can take many different shapes, including the sigmoid function is given by eqn (40).

$$G(\mathbf{w}, b, \mathbf{x})=\frac{1}{1+\exp\left(-(\mathbf{w}^T \mathbf{x}^T+b)\right)} \tag{40}$$

HV=Y

$$H=\begin{bmatrix} g(\vec{a}_1, b_1, \vec{x}_1) & g(\vec{a}_1, b_1, \vec{x}_2) & \cdots & g(\vec{a}_n, b_n, \vec{x}_N) \\ g(\vec{a}_2, b_2, \vec{x}_1) & g(\vec{a}_2, b_2, \vec{x}_2) & \cdots & g(\vec{a}_n, b_n, \vec{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ g(\vec{a}_n, b_n, \vec{x}_1) & g(\vec{a}_n, b_n, \vec{x}_2) & \cdots & g(\vec{a}_n, b_n, \vec{x}_N) \end{bmatrix}^T$$

$$V=\begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}_{n \times m}, \quad Y=\begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times m}$$

$$\mathbf{H}(\mathbf{w}_1, \ldots, \mathbf{w}_2, b_1, \ldots, b_L, \mathbf{x}_1, \ldots, \mathbf{x}_p)$$
$$= \begin{bmatrix} G(\mathbf{w}_1, b_1, \mathbf{x}_1) & \cdots & G(\mathbf{w}_L, b_L, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ G(\mathbf{w}_1, b_1, \mathbf{x}_p) & \cdots & G(\mathbf{w}_L, b_L, \mathbf{x}_p) \end{bmatrix}_{p \times L_L}$$
$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}$$
$$\mathbf{O} = \begin{bmatrix} \mathbf{o}_1^T \\ \vdots \\ \mathbf{o}_P^T \end{bmatrix}_{P \times m} \qquad (41)$$

where H signifies hidden layer output matrix and denotes final output matrix. In actual training, number of nodes L in hidden layer is typically less than number of training samples P. In case of a differentiable activation function, input weights and thresholds determined at random prior to training should remain static during training. Calculate least squares of following linear system to determine this.

$$\min_{\beta} \|\mathbf{H}\beta - \mathbf{O}l \quad \| \qquad (42)$$

By minimising the regularised cost function of least squares estimate regularisation, the output weights may be calculated, leading to following formulation (43).

$$\| min \quad L_{\mathrm{RELM}} = \frac{1}{2} \| V \| \quad ^2 + \frac{C}{2} \| Y - HV \|^2 \qquad (43)$$

where C is a scale parameter that regulates structural as well as experiential risk. By lowering LExLM's gradient corresponding to V to zero by eqn (44).

$$V + CH^T(Y - HV) = 0 \qquad (44)$$

Output weight matrix V in ExLMis represented by eqn (45).

$$V = \left(\frac{I}{C} + H^T H\right)^{-1} H^T Y \qquad (45)$$

Use kernel function instead of HHT when the mapping is uncertain by eqn (46).

$$HH^T = \Omega_{\mathrm{ELM}} = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{bmatrix}$$
$$h(x)H^T = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}^T \qquad (46)$$

Most popular kernel of ExLMin use is Gaussian kernel $K(x_i, x_j) = \exp\left(-\|x_i - x_j\|/\gamma\right)$ where γ is kernel parameter. Output weight matrix V in ExLMis given by (47)

$$V = \left(\frac{I}{C} + \Omega_{\mathrm{ELM}}\right)^{-1} Y$$
$$f(x) = h(x)H^T V = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}^T \left(\frac{I}{C} + \Omega_{\mathrm{ELM}}\right)^{-1} Y \qquad (47)$$

To begin, ELM is a supervised NN with a label as its output, whereas ELM-AE is an unsupervised NN with same output as input. Second, ExLM's input weights and hidden layer bias are both orthogonal, but ELMs are not is given by eqn (48).

$$h = g(ax + b), where\ a^T a = I, b^T b = 1$$
$$h(x_i)V = x_i^T, \text{ i=1,2……N} \qquad (48)$$

3.2 Decision-Level Fusion for multimodal sentiment and emotion recognition:

In speech and face recognition, the same dataset typically yields distinct predictions. The optimization method utilized to fuse two modalities to increase recognition accuracy as well as balance of distinct expressions after fusion, so that recognition results of their various modalities can compensate for each other. Employ two coefficients to linearly merge two basic emotion identification methods in this paper. Also, utilise an optimization approach to improve both the accuracy (precision) and uniformity of emotion detection evaluation model at same time represented by eqn (50).

$$R = w_1 \times R_1 + w_2 \times R_2 \tag{50}$$

where R1 and R2 are final prediction results of SER using a DCNN and facial expression recognition using DCNN, respectively. Furthermore, coefficients w1 and w2 must be optimised. The restrictions of the two coefficients are as follows, according to the model's actual meaning represented by eqn (51).

$$w_1 + w_2 = 1 \tag{51}$$

The optimization algorithm's goal is to optimise two coefficients such that two recognition methods can work together effectively. Accuracy of optimization coefficient will have a direct impact on the model's final recognition result, which will therefore have an impact on facial expression recognition accuracy. Following the acquisition of the face and voice results, decision level fusion approach is used to combine two to provide final recognition solution. To attain best results, an optimization technique is utilised for the first time in this paper to optimise emotion recognition model. Results of emotion recognition utilising the two recognition techniques are shown in R1 and R2. w1 and w2 are coefficients of combining two recognition procedures at same time.

4 Performance Analysis

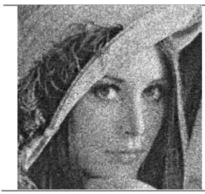
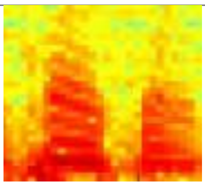
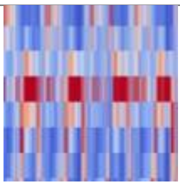
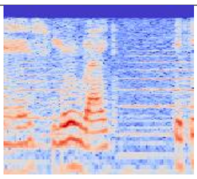
4.1 Dataset description

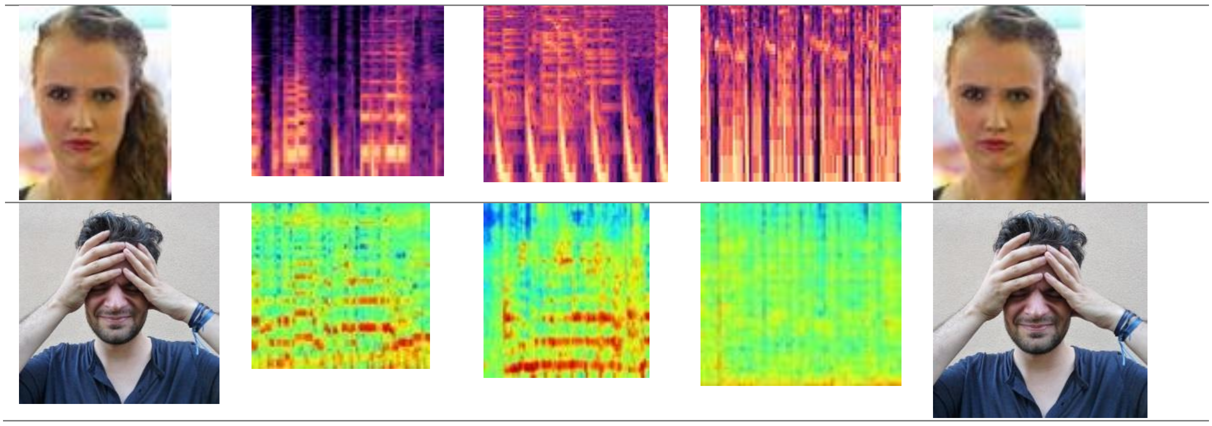
RAVDESS dataset: There are three modalities and two vocal channels in these files. There are 7356 recordings with acted-emotional content in RAVDESS denoted as Ryerson Audio-Visual Database of Emotional Speech and Song. Each file has a single actor who represents one of the eight emotions listed below: calm, happy, sad, furious, neutral, afraid, astonished, and disgusted. These emotions are formed at 2 levels of emotional intensity, with exception of neutral emotion, which only consists of regular intensity.

IEMOCAP multimodal emotion database: The IEMOCAP dataset contains 151 recorded discussion videos, each with two speakers, for a total of 302 videos. The existence of 9 emotions, as well as valence, arousal, and dominance, is noted in each segment. The data was collected over course of five sessions with five different speaker pairings.

MOUD dataset: They used YouTube to collect 80 product review and recommendation videos. Each video was broken down into its individual utterances (498 in total), each of which was assigned a sentiment label (positive, negative and neutral). Each video features an average of 6 statements, each lasting 5 seconds.

Table 1 processed multimodal data using proposed technique

Input data	Processed data	Classified data	Fused data	Recognition of emotion
				

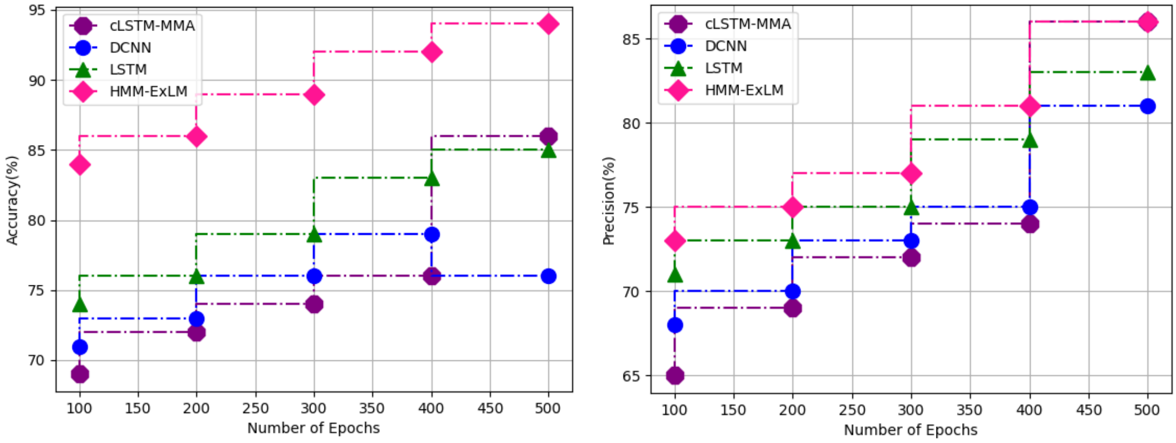


The above table-1 shows processed stages of multimodal data using proposed technique. Here the input data has been shown and its pre-processed data and classified data of audio and visual data have been shown. The fused data of both audio and visual data has been shown and finally recognized emotion is identified.

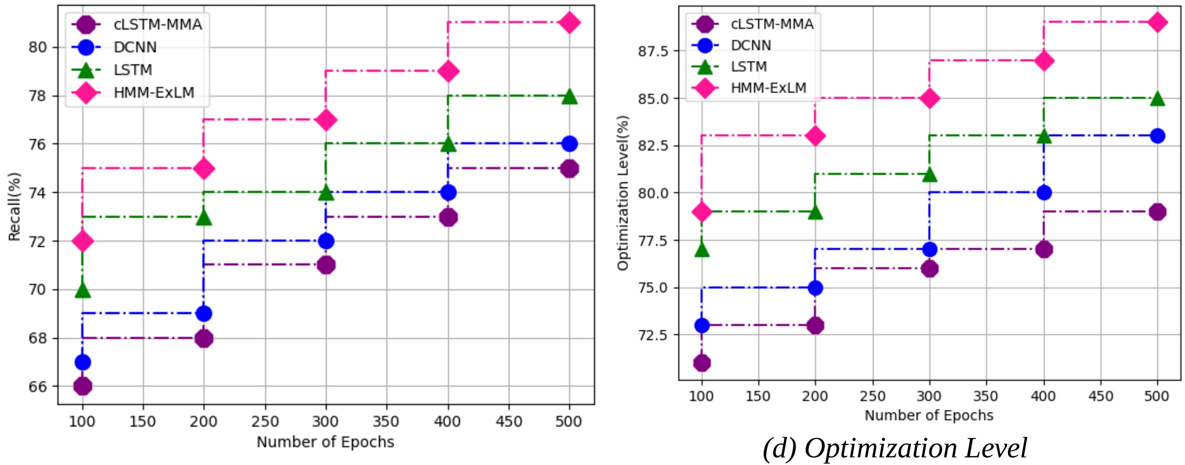
Table 2 comparative analysis between proposed and existing techniques

Dataset	Techniques	Accuracy	Precision	Recall	Optimization Level
RAVDESS Dataset	**cLSTM-MMA**	86	80.3	75	79
	DCNN	76	80.5	76	83
	LSTM	85	83	78	85
	HMM-ExLM	94	87	81	89
IEMOCAP dataset	**cLSTM-MMA**	88	87	73	83
	DCNN	91	88	77	85
	LSTM	93	91	81	87
	HMM-ExLM	95	95	83	95
MOUD Dataset	**cLSTM-MMA**	94.3	84	81	91
	DCNN	95.5	85	83	92
	LSTM	97.3	87	85	93
	HMM-ExLM	97.5	87.5	89	96

The above table-2 shows parametric analysis for various dataset in which the multimodal sentiment and emotion data has been collected. Here the parameters analyzed are accuracy, precision, recall and Optimization Level comparison with existing techniques namely cLSTM-MMA, DCNN, LSTM with proposed HMM-ExLM



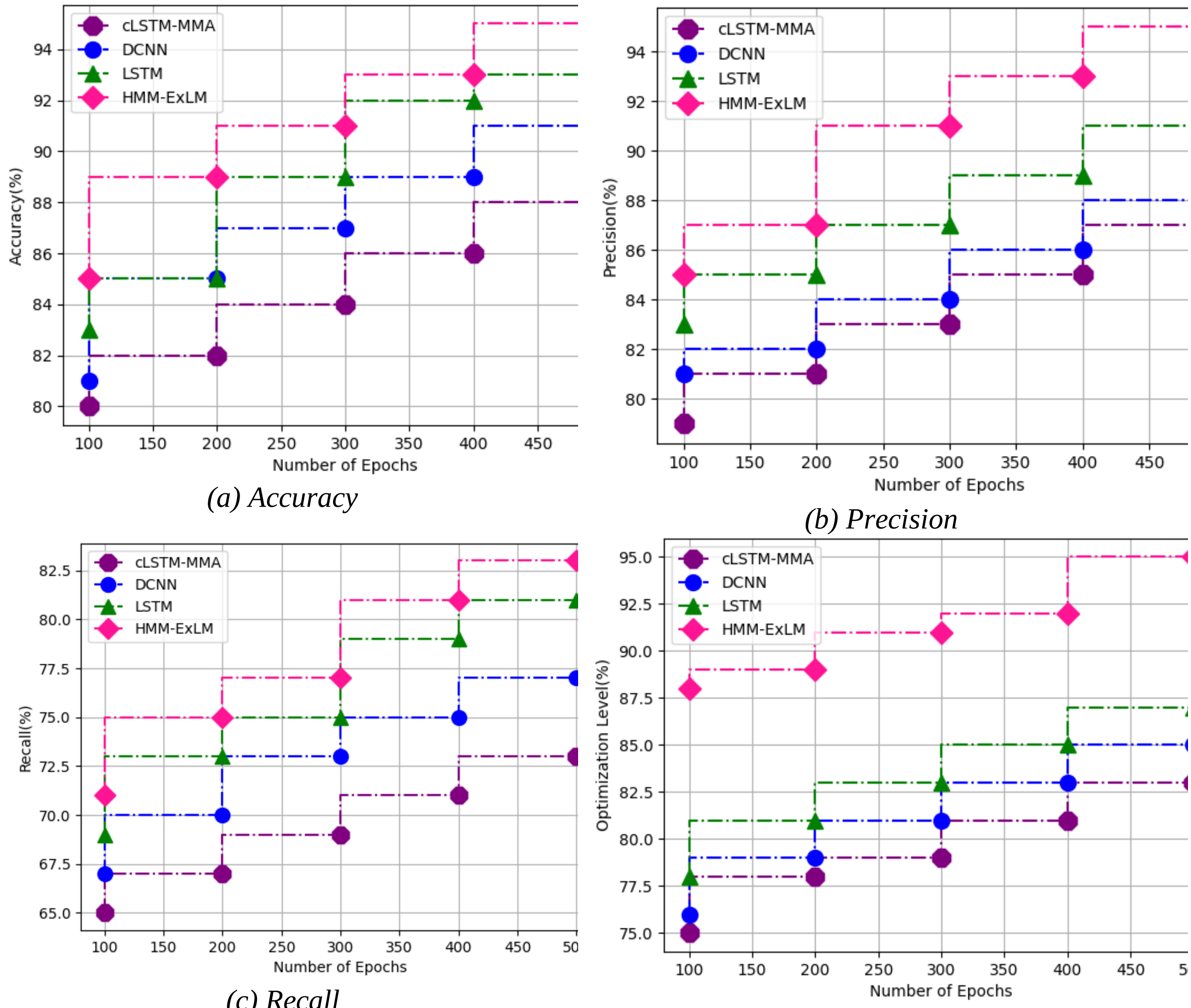
(a) Accuracy *(b) Precision*



(d) Optimization Level

(c) Recall

Figure 2 Comparative analysis for RAVDESS dataset in terms of (a) Accuracy, (b) Precison, (c) Recall, (d) Optimization Level



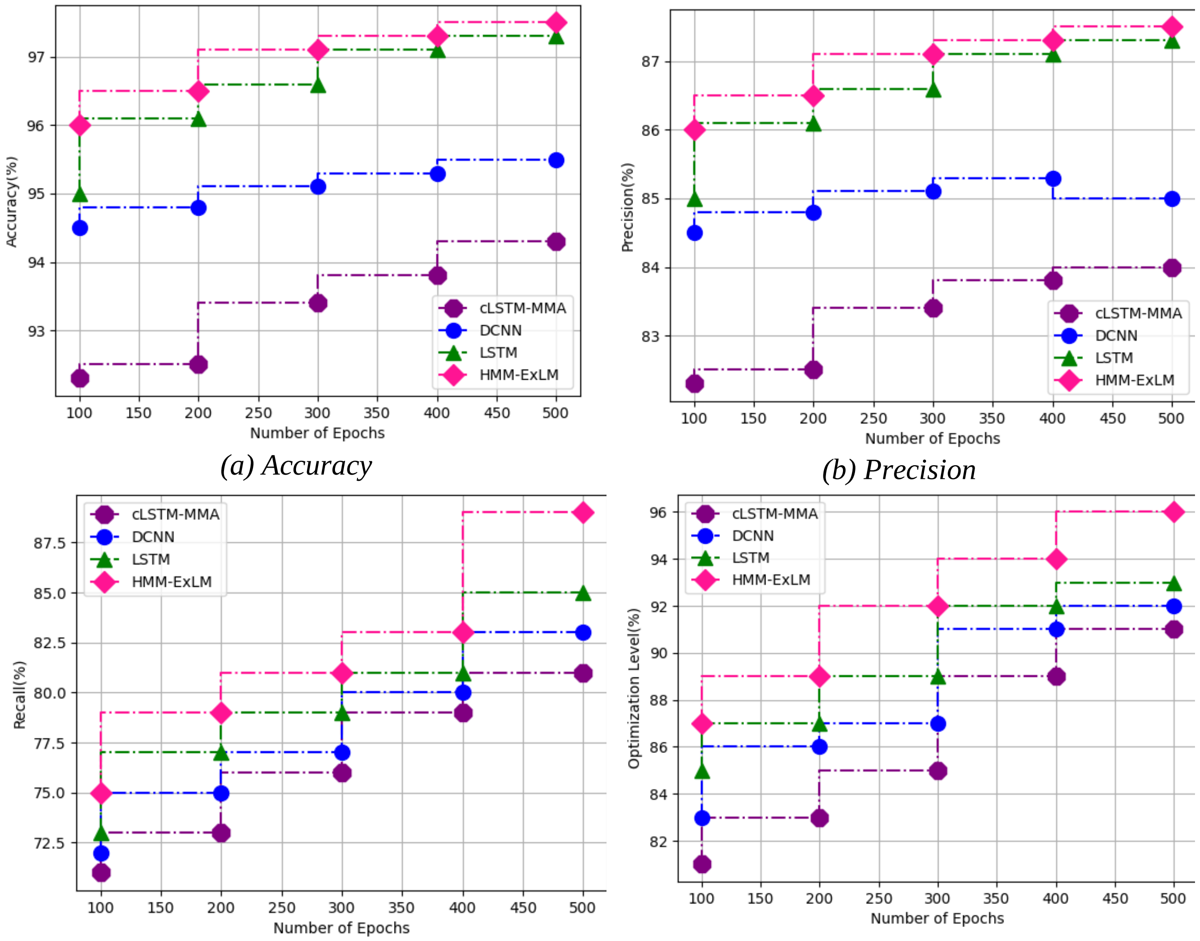
(a) Accuracy

(b) Precision

(c) Recall

(d) Optimization Level

Figure-3 Comparative analysis for IEMOCAP dataset in terms of (a) Accuracy, (b) Precison, (c) Recall, (d) Optimization Level



(a) Accuracy

(b) Precision

(c) Recall

(d) Optimization Level

Figure 4 Comparative analysis for MOUD dataset in terms of (a) Accuracy, (b) Precision, (c) Recall, (d) Optimization Level

The above figure 2-4 shows comparative analysis for RAVDESS dataset, IEMOCAP dataset and MOUD dataset in terms of accuracy, precision, recall and Optimization Level. Comparatively the proposed technique obtained enhanced result in classifying the ontology based dataset. For RAVDESS dataset HMM-ExLM obtained accuracy by 94%, precision attained is 87%, recall obtained is 81% and Optimization Levelis 89%. The accuracy for IEMOCAP dataset is 95%, precision obtained is 95%, recall is 83% and Optimization Levelof 95% is obtained by proposed technique. Finally, for MOUD dataset, HMM-ExLM obtained accuracy by 97.5%, precision attained is 87.5%, recall obtained is 89% and Optimization Levelis 96%. From this analysis the proposed technique has obtained optimal results in multimodal data classification and fusion with recognition of emotions.

5 Conclusion

This paper proposed novel technique in multimodal emotion recognition and classification. The aim of this proposed technique is to collect the online social media dataset which consist of audio and visual data in recognition of multimodal sentiment and emotion. Here the classification of audio and visual data has been carried out using hidden Markov model based extreme learning machine (HMM ExLM). From the classification output the audio features and image features has been classified. Then by fusing both the features using decision level fusion technique, the emotion and sentiment of the features has been analysed and decision is made in recognising whether the sentiment and emotion are same. The experimental results show the extracted features of audio data, visual data and the result showing the fused data. Comparative analysis has been made in terms of accuracy, precision, recall and optimization level of fusion.

References

[1] S. Li and W. Deng, "Deep facial expression recognition: a survey," IEEE Transactions on Affective Computing, 2020.

[2] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," IEEE Transactions on Image Processing, vol. 28, pp. 2439–2450, 2018.

[3] R. Tanabe and H. Ishibuchi, "An easy-to-use real-world multi objective optimization problem suite," Applied Soft Computing, vol. 89, article 106078, 2020.

[4] M. J. Mayer, A. Szilágyi, and G. Gróf, "Environmental and economic multi-objective optimization of a household level hybrid renewable energy system by genetic algorithm," Applied Energy, vol. 269, article 115058, 2020.

[5] J. Zhang, Y. Huang, Y. Wang, and G. Ma, "Multi-objective optimization of concrete mixture proportions using machine learning and metaheuristic algorithms," Construction and Building Materials, vol. 253, article 119208, 2020.

[6] F. Wang, Y. Li, F. Liao, and H. Yan, "An ensemble learning based prediction strategy for dynamic multi-objective optimization," Applied Soft Computing, vol. 96, article 106592, 2020.

[7] M. Abdel-Basset, R. Mohamed, and S. Mirjalili, "A novel whale optimization algorithm integrated with Nelder-Mead simplex for multi-objective optimization problems," Knowledge-Based Systems, vol. 212, p. 106619, 2021.

[8] M. D. Ding and L. Li, "CNN and HOG dual-path feature fusion for face expression recognition," Information and Control, vol. 49, no. 1, pp. 47–54, 2020.

[9] Q. Lan and X. Zhang, "Facial expression recognition method based on a joint normalization strategy," Journal of Beijing University of Aeronautics and Astronautics, vol. 46, no. 9, pp. 1797–1806, 2020.

[10] A. A. Muhammad and J. K. Muhammad, "EEG-based multi-modal emotion recognition using bag of deep features: an optimal feature selection approach," Sensors, vol. 19, no. 23, 2019.

[11] A. Zadeh, P. P. Liang, and S. Poria, "Multi-attention recurrent network for human communication comprehension," in Proceedings of the 32th AAAI Conference on Artificial Intelligence, pp. 5642–5649, New Orleans, USA, 2018.

[12] Li, X., Parizeau, M., &Plamondon, R. (2000). Training hidden markov models with multiple observations-a combinatorial method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(4), 371-377.

[13] Momenzadeh, M., Sehhati, M., & Rabbani, H. (2019). A novel feature selection method for microarray data classification based on hidden Markov model. *Journal of biomedical informatics*, *95*, 103213.

[14] Ding, S., Zhang, N., Xu, X., Guo, L., & Zhang, J. (2015). Deep extreme learning machine and its application in EEG classification. *Mathematical Problems in Engineering*, *2015*.

[15] Fan, Q., & Fan, T. (2021). A Hybrid Model of Extreme Learning Machine Based on Bat and Cuckoo Search Algorithm for Regression and Multiclass Classification. *Journal of Mathematics*, *2021*.

[16] Q. W. Fan and T. Liu, "Smoothing L0regularization for extreme learning machine," Mathematical Problems in Engineering, vol. 2020, Article ID 9175106, 10 pages, 2020.

[17] G. Xu-Sheng, Q. Hong, M. Xiang-Wei, W. Chun-Lan, and Z. Jie, "Research on ELM soft fault diagnosis of analog circuit based on KSLPP feature extraction," IEEE Access, vol. 7, pp. 92517–92527, 2019.

[18] N. Nabipour, A. Mosavi, A. Baghban, S. Shamshirband, and I. Felde, "Extreme learning machine-based model for solubility estimation of hydrocarbon gases in electrolyte solutions," Processes, vol. 8, no. 1, p. 92, 2020