Journal of Intelligent & Fuzzy Systems 45 (2023) 8989–9005
DOI:10.3233/JIFS-230766
IOS Press

A dense layer model for cognitive emotion recognition with feature representation

S. Yuvaraj[a,*] and J. Vijay Franklin[b]
[a]*Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India*
[b]*Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamilnadu, India*

Abstract. The predictions of cognitive emotions are complex due to various cognitive emotion modalities. Deep network model has recently been used with huge cognitive emotion determination. The visual and auditory modalities of cognitive emotion recognition system are proposed. The extraction of powerful features helps obtain the content related to cognitive emotions for different speaking styles. Convolutional neural network (CNN) is utilized for feature extraction from the speech. On the other hand, the visual modality uses the 50 layers of a deep residual network for prediction purpose. Also, extracting features is important as the datasets are sensitive to outliers when trying to model the content. Here, a long short-term memory network (LSTM) is considered to manage the issue. Then, the proposed Dense Layer Model (DLM) is trained in an E2E manner based on feature correlation that provides better performance than the conventional techniques. The proposed model gives 99% prediction accuracy which is higher to other approaches.

Keywords: Cognitive emotion recognition, deep learning, prediction, visual modality, handcrafted features

1. Introduction

There is an important feature to finish the communication between the machine and human cognitive emotion recognition, which is the effective date to communicate with humans. The recognition of cognitive emotion application is identified in various domains. Consider an example that the states of cognitive emotions are utilized for predicting and monitoring the state of fatigue [1]. The recognition of cognitive emotion is utilized in call centers in speech recognition to predict the cognitive and emotional state of the caller and give feedback on the quality of service [2]. Due to the lack of human cognitive emotions of the temporal boundaries and various single expressions and accepting the cognitive emotions in various ways, cognitive emotion recognition is difficult [3]. Other modalities like visual information, such as facial gestures, are used even though the present work, like recognition of cognitive emotion, focuses on inferring the subjects' cognitive emotions rather than the speech. In the past years, several ground-breaking enhancements have been captured with the deep network model in various implemented regions of pattern recognition like speech, object, and speaker recognition and the joined issue-resolving techniques such as in recognition of audio and visual and the present paralinguistics field. inherent structure.

Different researches are presented in the needed network property with the variants for modeling the obtained in the signal of speech [4], having more present attempts in research for the end-to-end optimization to use like the less human priority knowledge [5]. Nonetheless, these works have the majority utilize the general handcrafted engineered

*Corresponding author. S. Yuvaraj, Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India. E-mail: yuvaraj.syr@gmail.com.

features like coefficients of supra-segmental features and Mel-Frequency Cepstral Coefficients (MFCC) like these are utilized in the difficulties of AVEC and ComParE series [6] that is established on the knowledge which is gained in auditory study decades, and that is presented to be powerful for multiple domains of speech. Moreover, in recent years, a community of machine learning has emerged to derive the presentation of input signals from unprocessed and raw data. The aim behind the concept is that the network acquires the intermediate presentation of the input signal, which is raw and automatically suited to better and improved performance.

The automatic sensing effect is proposed with the help of both visual and speech data in a point-to-point way. The CNN architecture is utilized for feature extraction from the speech signal and is modeled for the audio channel, and the ResNet-50 architecture is used for the visual data [8]. The network output is combined and given to the LSTM to identify the individuals' affective states. On the contrary, every network is trained separately; the outcomes are provided simply to the consecutive classifier, and the proposed system is trained point-to-point. In the research, this is the primary work that uses these point-to-point systems to recognize cognitive emotion for the audio and visual.

Further, the concordance correlation coefficient (ρ_c) has explicit maximization, which is used in the proposed system, and the performance is improved concerning the prediction of cognitive emotion when compared with the objective of the mean square error that is optimized, and that is utilized conventionally [9]. Lastly, interpretable cells have existed that are found by researching the various cell activations in the recurrent layers that are greatly correlated, having different features of acoustic and prosodic, which have the assumption for conveying the information related to affective in speech like the basic frequency and loudness. The primary version of the work is described in [10], which uses the raw waveform of speech. The modality related to visuals is considered the extension of the proposed system in a point-to-point way.

The database of REmote COLlaborative and Affective (RECOLA) is determined in the proposed system to provide the benefits of the suggested multimodal model. In 2016, the Audio/Visual Cognitive Emotion Challenge and Workshop (AVEC) was used as part of the database. The proposed system is tested and trained with the help of a complete database. The merits of the multimodal model are shown by the outcomes from the two modalities using the better outcomes production for the valence and arousal like the visual and speech networks [11–13]. The multimodal and unimodal models are compared with the help of attained outcomes in AVEC 2016. The modalities of visual, audio, or visual audio are used in the system [14, 15]. However, in this work, the proposed system creates better outcomes for both the modalities of visual and speech that are presented using the proposed experiments. The significance of the anticipated model is provided below:

1) Feature learning for cognitive emotion recognition automatically;
2) The proposed DLM model transforms the audio and video features and provides a suitable CNN model.
3) The experimental outcomes specify the learned features with promising results.

The work is structured as follows: the related works are analyzed in Section 2. The dataset description is given in Section 3. The anticipated DLM model is provided in Section 4. The experimental outcomes are provided in Section 5. The conclusion is provided in Section 6.

2. Related works

The models of pattern recognition have performance that is enhanced by having DNNs. In recent years, the sequence of the new architectures of neural networks has been revitalized like the autoencoder networks [11], models of memory that enhanced neural networks or Deep Belief Networks (DLMs) [13], CNNs [12], and models of LSTM [14]. The models are utilized differently for multimodal recognition, like speech recognition. Multimodal Deep Autoencoder (MDAE) network is suggested by Keren et al. [15] for the feature extraction from video and audio data. Bimodal DLM is primarily trained for initializing the deep autoencoder and fine-tuned with MDAE to minimize both reconstruction error modalities. A temporal multimodal network called Restricted Boltzmann Machine (RBM) is proposed by Mayya et al. [16] for modeling the series of audio-visual data. Also, DNNs are utilized to recognize the gesture. The researchers utilize the skeletal data and the images of RGB-D to recognize the gestures in [17]. DLMs are used for processing the skeleton's features, and CNN 3D is for the data of RGB-D. Hidden Markov Model (HMM) is stacked on top concerning temporal data.

The domain of cognitive emotion recognition benefits from having DNN's advent. Some deep learning techniques are employed to recognize speech cognitive emotion. The handcrafted features are used by Dhall et al. [18] for feeding the DNN which creates probability distribution rather than the states of definite cognitive emotion. The classification is performed using extreme machine learning trained from the probabilities to calculate the statistic from the complete utterance. After the data transformation, with the help of a short time Fourier transform by Simonyan et al. [19] using the CNN for the high-level features extraction. LSTMs are used to obtain the temporal structure. The end-to-end model suggested by Ciregan et al. [20] employs CNN for the feature extraction from the raw signal and LSTM is used for obtaining the contextual data in the information.

Cognitive emotion recognition is solved using the works with the help of facial data having DNNs. For instance, Krishna et al. [21] suggest a framework for transductive learning to recognize image-oriented cognitive emotion using the combination of hypergraphs and DNNs. Every node is concerned with fully connected layer for forming the hyperedge in cognitive emotion classification. RNNs and CNNs are combined to recognize the unconditional cognitive emotion in video. CNN is trained primarily for classifying the images related to static with cognitive emotion. Thus, CNN is utilized for feature extraction by training RNN to create the complete video cognitive emotion [22–25].

The CNNs extract the features, and multiscale Dense SIFT features (MSDF) are used to extract the features from the faces for training the SVR (linear Support Vector Regression). The acoustic Parameter Set (APS) is utilized to haul out the audio features. The features are combined to use for SVR learning. It is learned using combination of features. Multimodal CNN is used by Zeng et al. [26] to classify cognitive emotions using visual and audio modalities. There are two phases to training the model. The two CNNs are pre-trained in the first phase on the large image datasets, and it is tuned finely for performing the recognition of cognitive emotion. Audio CNN is considered with the meal-spectrogram segment as the input for the video and audio signal to take the face from CNN. The DNN is trained in the second phase to comprise the number of fully connected layers. The two CNNs are utilized to concatenate the extracted features as input. BLSTM-RNN is used for capturing the contextual data presented in the multimodal features like video, audio, and physiological data for the extraction by Cootes et al. [27].

The framework for the strength modeling is established by Edwards et al. [28], which has been proposed recently as decision-level and feature-level techniques. It is combined with the feature vector and gives the second regression for prediction. The motivations for cognitive emotion recognition shows significance with AVEC created [29]. The modalities of video, physiological, and audio are concerned with the challenge of 2016. Relevance Vector Machine (RVM) is proposed by Liu et al. [30] to model video, audio, and visual-audio data. Zhong et al. [31] anticipated a model in another work with the features of high-level geometry to identify the dimensional features. Yu et al. [32] uses higher and lower-level features to model cognitive emotions. Some baseline features for video and audio are complemented by Senechal et al. [33] for performing cognitive emotion recognition. The extra features are used by Siddiqi et al. [34] where an audio-only modality is used. The work utilizes the general handcrafted features in visual or audio or, in a few cases, both. However, the temporal information in the data is only sometimes concerned [35]. The trained multimodal model is proposed in this research for the point-to-point, which concerns the information related to contextual temporal.

3. Dataset

Here, modalities refer to the analysis with three diverse datasets like eNTERFACE05, RML and BAUM. BAUM dataset consists of 1184 multi-modal facial expressions and speech of 13 mental and emotional states. It is a time-series dataset with 1184 instances. It includes short video clips in MP4 format, and the mental and emotional states include fear, happiness, sadness, anger, surprise, disgust, contempt, boredom, bothered, concentration, neutral, confusion, and so on.

Similarly, RML is an adult emotion dataset that shows slight variations in feelings and human thoughts. It offers social wealth clues to the researchers that concentrate on motivation, intention, attention and emotion. It is considered a productive tool for performing communication. Analysis of these expressions provides better insight into human behavior.

eNTERFACE05: The dataset for the visual and audio by eNTERFACE05 [36] has cognitive emo-

tions, which are six. They are disgust, anger, joy, fear, surprise, and sadness, having 14 various nationalities from 43 subjects. The video samples of 1290 are present. Every audio sample having a rate of sampling of 48,000 Hz is recorded having a mono channel and 16-bit resolution. Every subject must listen to the six consecutive small stories to induce the specific cognitive emotion. Two experts evaluate if the reaction presents the cognitive emotion intended unambiguously. The speech utterances gathered from the video files have speaking subjects. The video files should be, on average, 3–4 seconds. The original video frames have the size of 720×576×3. Few samples of the cropped face image are presented on the dataset of eNTERFACE05.

4. Methodology

Figure 1 presents the proposed model, which is the model of deep hybrid learning, which has two single streams of input, such as the processing of the visual network having the visual data with the model of 3D-CNN and the processing of the audio signal with the audio network having the CNN. The fully connected layers (FCL) have the outputs for combining the two networks in the implemented fused network with the DLN model.

The previous 3D-CNN and CNN models are used in the proposed system having the largescale image, which is pre-trained, and the works of video classification for initializing the 3D-CNN and CNN accordingly because of the restricted volume of labeled data. Thus, the two CNN models fine-tun with the labeled cognitive emotion data. The CNN initialization and the C3D-Sports-1M model are chosen for the initialization of 3D-CNN

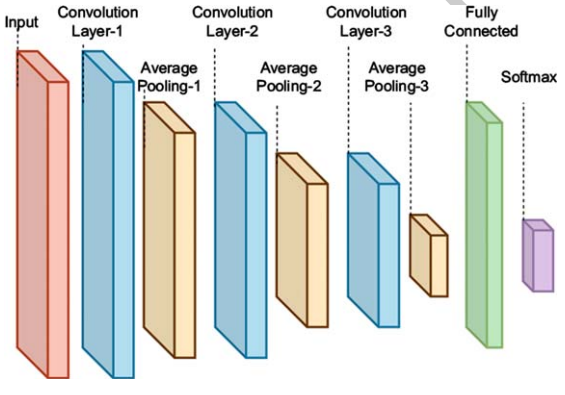


Fig. 1. Feature representation with DLM.

by AlexNet [38] in the proposed system. There are 5 convolution layers like Conv1-to-Conv5 by AlexNet [37] and 3 fully connected (FCFC) layers, and 3 max-pooling layers, such as Pool1-to-Pool5. The FCFC layers, which are the first two, fc6 and fc7, have units of 4096, and the last layer of FCFC, such as fc8, includes 1000 dimensions, which are relevant to the classification of 1000 images. There are 8 convolution layers such as (Conv1a − Conv2a − · · · −Conv5aConv5b, *Pool* 1 − Pool2 − Pool3 − Pool4 − *Pool*5) as 5 max-pooling layers and 3 layers of FC in the C3D-Sports-1M model [39]. The fc6 and fc7 have units of 4096, and the fc8 has the classification of 487 videos in 3D-CNN. Figure 1 shows the initialization of the visual and audio networks in the proposed system to copy the network parameters from the 3D-CNN and pre-trained CNN models presented earlier. It is important to consider that the parameters of fc8 in the pre-trained two models are utilized. This process explains how to create both 3D- and CNN inputs and the training process with the hybrid deep learning system.

4.1. Network generation

Every sample is partitioned into several overlapped segments, and learning the visual and audio features from every segment since the cognitive and emotional video samples have various times. The amount of trained data is enlarged for the proposed deep models. The complete log Mel-spectrogram is obtained from signals of audio extracted primarily. The calculation of log Melspectrogram that are extracted having the Mel-frequency filter output and discriminant power is shown rather than MFCC to recognize the cognitive emotion of audio [40]. Thus, the proposed system uses the context (fixed) to divide the spectrogram into overlapped segments that are transformed into relevant CNN input. The relevant segment of the video is used in the window context as the 3D-CNN input after the pre-processing process. The Mel-spectrogram segment and the video frames are produced in the proposed system for every video segment presented in Fig. 1. A detailed presentation about processing visual and audio cues is done.

4.1.1. Input generation

Audio input: Fig. 1 shows the conversion of 1D audio signals to the relevant CNN input. Three segments of log Mel-spectrogram channels are extracted: delta, delta-delta, and static, which have 64×64×3

sizes. More particularly, the 64 Mel-filter banks are chosen from 20-8000 Hz for obtaining a complete log Mel-spectrogram (LMS) with the help of 10 ms overlapping and 25 ms Hamming window for the provided utterance in the proposed system. Thus, the context window has 64 frames for dividing the complete LMS into audio segments of 64×64 sizes. During segmentation, a size of shift having the frames as 30 is utilized, such as two adjacent segments having 30 frames through overlapping. Every segment is divided into 64 frames in length and 10 ms×(64–1)+25 ms as time duration. Here, the divided segment is 3 times higher than the minimal length segment, such as 260 ms, to find the cognitive emotion. Every divided segment gives adequate temporal cues consecutively to find the cognitive emotions. The segment of 2-D Mel-spectrogram [41] is created having the 64×64 size to consider as the primary static channel in three Mel-spectrogram channels. The first order, such as delta and the second order, such as delta to delta, is calculated in the proposed system for the derivations of the frame-to-frame time after extracting the segment with the static Mel-spectrogram having 64×64 size. Better temporal information related to Melspectrogram is obtained, such as the trajectory of the feature across time that is generally completed in speech recognition works. The below regression formula is required to compute the coefficients of delta having the static segment of the 2-D Melspectrogram.

$$d_t = \frac{\sum_{n=1}^{N} n\,(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (1)$$

Here, delta coefficients for the frame t are d_t specifies static coefficients of the segment of the Mel-spectrogram having c_{t+n} *to* c_{t-n}. The *n* value specifies window of regression having the value 2 typically. Thus, delta coefficients to the delta are computed similarly from the attained coefficients of the delta. Three segments of Mel-spectrogram channels obtained have the 64×64×3 sizes presented in Fig. 1. The LMS is extracted concerning the RGB images with the audio data feature representations. There are two needed properties. Primarily, the proposed system establishes the operation of 2D convolution with the axis and frequency of time over the operation of 1-D convolution. Secondly, this is easy to resize the image to the relevant size like the pre-trained models of CNN as input as the feature representation of the RGB image. More particularly, the audio network is initialized in the proposed system by AlexNet [28], including the input size of 227×227×3. Hence, the original spectrogram is resized from having the 64×64×3 sizes to the new size of 227×227×3 in the proposed system having the bilinear interpolation. Here, a is presented as the audio input in the proposed system.

2) Video input: After dividing the video sample into segments, the video segments are utilized as the 3D-CNN input. The detection of the face is determined, and the eye's distance is determined for every frame in the video segment. Lastly, the image size of 150×110×3 for the RGB face is cropped. The powerful real-time face detector is used in detail to detect faces automatically on every frame. The two eye centers are placed in the up-right face, typically from the outcomes for automatically detecting the face. Thus, the facial images have the eye distance computed, normalized to the fixed 55 pixels of distance. It is noticed that the height is three times lengthier than the distance of the eye for the face image, and the width is twice what is estimated roughly. A resized image of RGB, such as 150×110×3, is cropped from every frame depending on the normalized distance of the eye. The facial image for every frame is cropped to resize for the input 227×227×3 for the 3D-CNN model, which is pre-trained to perform the fine-tuned work. The same resizing operation is utilized in the existing study [42]. Every video segment used has frames of 16, which is the size of the input presented. The videos overlap frames as $\frac{L-16}{2}$ in the proposed system when the video segment includes $L \geq 16$. On the other hand, the first and last $\frac{L-16}{2}$ overlapping frames are repeated for $L \geq 16$ in the proposed system. It is important to consider that the duration of every segment is 655 ms by 20 frames of video in every segment of the video, such as 0.6 s×30 frame/s, as 65 audio frames are used in the context window for dividing the LMS extraction to the segments of audio. The proposed system's implementation does not require dealing with the $L < 16$ frames scenario. On the other hand, the first 5 and last 6 overlapping features are repeated in the proposed system when using 15 frames of audio for the segments of Mel-spectrogram relevant to over frames of 5 video such as $L = 5$ for the experiments. The v is represented as visual input in the proposed system.

3) Network training: The data of audio and visual is presented as $X = \{(a_i, v_i, y_i)\}_{i=1,2,\cdots,K}$, having index as 'i' for segments for classified visual and audio. The visual data and audio data are denoted by v_i and a_i accordingly. The class label of the segment is represented by y_i. The class label must be used for

the global video sample as the class label segment is y_i. The fc7 has the output of 4096-D, denoted by $\Upsilon^{A}\left(a_i,\ \theta^{A}\right)$ in the network of audio that is represented as A, having the θ^{A} as the parameters of the network. In the same way, the fc7 has the visual feature of 4096-D denoted by $\Upsilon^{V}\ \left(v,\ \theta^{V}\right)$ for the visual network that is presented as V having the θ^{V} as the parameters of the network. The audio/visual models have trained accordingly during the network training, thus training the fused network in the successive phase [43].

The visual and audio networks are primarily trained separately, using fine-tuning techniques. The final FCLs are replaced for the 3D-CNN and CNN, that is, the layer of fc8, the two new layers of FCFC that is relevant to the categories of cognitive emotion on the target dataset to recognize the cognitive emotion related to visual and audio. Consider an example that fc8 needs to create the outputs as 6 for the 6 cognitive emotions [44]. The cognitive and emotional labels are predicted in the proposed system having the networks of visual and audio accordingly to compute the prediction errors, and lastly, the parameters of the network need to be updated to lower the L as a negative log-likelihood across the trained data. The below minimization issue is followed in the proposed system on training the audio data for updating the network of audio A having the back-propagation.

$$\min_{W^{A},\ \theta^{A}} \sum_{i=1}^{K} L\left(soft\max\left(W^{A}.\Upsilon^{A}\left(a_i,\ \theta^{A}\right)\right),\ y_i\right) \tag{2}$$

Here, the softmax layer has the weight values are represented by $W^{A}WA$, a_i specifies minimization issue, θ^{A} represents back-propagation and the calculation for the softmax (log loss) is presented below.

$$L\left(A,\ y\right) = -\sum_{j=1}^{l} y_j \log\left(y_j^{A}\right) \tag{3}$$

Here, the ground truth label has the j^{th} value denoted by y_j, the softmax layer has the j^{th} output value denoted by y_j^{A} for A, and the total class labels are presented as $L(A, y)$. Since a similar minimization issue is present in 3D-CNN like CNNs, the minimization issue is solved in the proposed system on training the visual data, which is the same as the audio network. The prediction error is minimized similarly, V for updating the V as a visual network. The parameters are updated individually in the networks of visual and audio during the first stage of training to generate more discriminate visual and audio features, that is $\Upsilon^{v}\left(v_i;\ \theta_V\right)$ and $\Upsilon^{A}\left(a_i;\ \theta_A\right)$. The training of the fused network is presented in the proposed system to combine visual and audio features [45].

4) Network fusion: The fc8 layers are discarded in the proposed system after the training of visual and audio networks and combine the layers of fc7 into the fused network, presented in Fig. 1. Two features of 4096-D $\Upsilon^{V}\left(v_i;\ \theta_V\right)$ and $\Upsilon^{A}\left(a_i;\ \theta_A\right)$ are combined to constitute the feature of 8192-D as the fused network input as

$$f\left(\left[\Upsilon_i^{A},\ \Upsilon_i^{V}\right];\ \theta^{F}\right)$$

that is presented as F having the θ^{F} as the parameters of the network. Where, $\Upsilon_i^{V} = \Upsilon^{V}\left(v_i;\ \theta^{V}\right)$ and $\Upsilon_i^{A} = \Upsilon^{A}\left(a_i\ ;\ \theta^{A}\right)$. The proposed system has the fused network, which is implemented by having the model of deep DLM that focuses on obtaining a high non-linear relationship over modalities and creating the discriminative representation of features to classify the cognitive emotion. There are two hidden layers, one visible and one output layer, as the softmax layer, presented in Fig. 1. The two RBMs are stacked to construct the model of DLM; that is, the bipartite graph and the hidden nodes can have the high-order input data correlation for the visible nodes.

The proposed system trains the fused network via the two training stages. Primarily, the pre-trained unsupervised model implementation is done bottom-up with the help of a training algorithm for greedy layer-wise. The reconstruction error is lowered by the pre-trained unsupervised model with the total training samples as K, and the loss function of cross-entropy is denoted by $C\left(z_i,\ z_i'\right)$ between the reconstructed data z_i' and input data, where the definition of $C\left(z_i,\ z_i'\right)$ is presented below.

$$C\left(z_i,\ z_i'\right) = \sum_{d=1}^{D}\left(-z_{i,j}\log z_{i,d}' + \left(1 - z_{i,d}\right)\log\left(1 - z_{i,d}'\right)\right) \tag{4}$$

Secondly, every RBM layer is introduced after the pre-trained model. Thus, the fine-tuned supervised model is carried out to optimize the parameters of the network. The last hidden layer has the output performed in detail as the classifier input, and the classification error is calculated. Thus, the network parameters are readjusted by using back-propagation.

The Gaussian-Bernoulli RBM is used, having the hidden nodes as 4096 for the first layer, there are 2048 hidden layers in the Bernoulli-Bernoulli RBM for the second layer, and the 2048-D features have the outputs to classify the cognitive emotion since the DLMs have the input features are the continuous values. The DLMs structure as 8192-4096-2048- C is obtained similarly to find the cognitive emotions as C on the target visual and audio cognitive, emotional datasets. The parameters are fixed in *A* and *V* when training the successive stage and updating the fused network as *F* for accurately creating more cognitive and emotional predictions, giving the best feature fusion results.

4.2. Classification process

The representation of the 2048-D joint feature is calculated after training the fused network completion on every segment of visual and audio (See Fig. 1). Average pooling is used in complete features of the segment from every sample of video for creating the global feature representation of video as the fixed length since every video sample of visual and audio consists of the various segments. The max pooling and average pooling are compared in the proposed system experiments, and identified that the better performance is achieved by average pooling. Hence, average pooling is utilized for processing the extracted features from the segments. The linear SVM classifier is used easily to identify cognitive emotion depending on the representation of global video features [45].

5. Numerical results and discussion

The experiments on cognitive emotion recognition are performed on three public datasets of cognitive emotional visual and audio which are the acted dataset of eNTERFACE05, to test the efficiency of the suggested networks of deep hybrid learning to recognize the visual and audio cognitive emotion. The unimodal video and audio recognition outcomes are presented to determine the performance, and the results of multimodal cognitive emotion recognition are provided to integrate video and audio cues.

5.1. Parameter setup

The mini-batch size is 30, and the stochastic momentum of 0.9 has stochastic gradient descent (SGD) to train the models. The rate of learning for fine-tuning is 0.001. The epoch numbers are 400 for 3D-CNNs, 100 for DLMs, and 300 for CNNs. The parameter of the dropout is 0.3 for the method of FCFC fusion. The CNNs are implemented in the proposed system with the MatConvNet tool, the DLMs with the DeeBNet toolbox and one 3D-CNN with the Caffe toolbox. The package of LIBSVM is used in the proposed system for the classification of cognitive emotion for performing the algorithm of SVM, having the one-versus-one approach and the linear kernel function.

5.2. Result analysis

The recognition performance has two features presented (See Fig. 2): extracted features learned with the fine-tuned AlexNet and extracted features with the models. The created visual and audio data in the proposed system as the models for the extracted features having the models of C3D-Sports-1M and AlexNet to create the features of 4096-D from the fc7 layer output accordingly. Table 1 depicts the performance of feature recognition shown on the eNTERFACE05. Table 1 shows that the learned features are shown with the models of fine-tuned deep, such as C3D-Sports-1M and AlexNet, that are presented to perform well considerably than the extracted features having the original models of pre-trained deep learning. The accuracies are improved using the technique of fine-tuning in detail on the dataset, 53.03% to 68.09% for visual features and 60% to 66% for audio features.

In the same way, the proposed system creates the enhancement from 49% to 55% for the feature of visuals and 52% to 79% for the features of audio accordingly. The enhancement of 8% for visual features and 6% for audio features are obtained on the dataset accordingly. The efficiency of the proposed feature learning technique is demonstrated by the results from experiments such as the deep model used for learning the features of cognitive emotion. The proposed system has the learned features that have to give the robust ability for the deep learning models potentially for extracting the more discriminant cues over the designed features manually. The results from the experiments determine the validity of the fine-tuning technique. The pre-trained models are allowed by fine-tuning to learn the useful feature representations to recognize the cognitive emotion in other domains (See Fig. 3a to 3c).

The proposed system has the performance of having recorded outcomes of the prior research with the help of handcrafted features to present the merits of

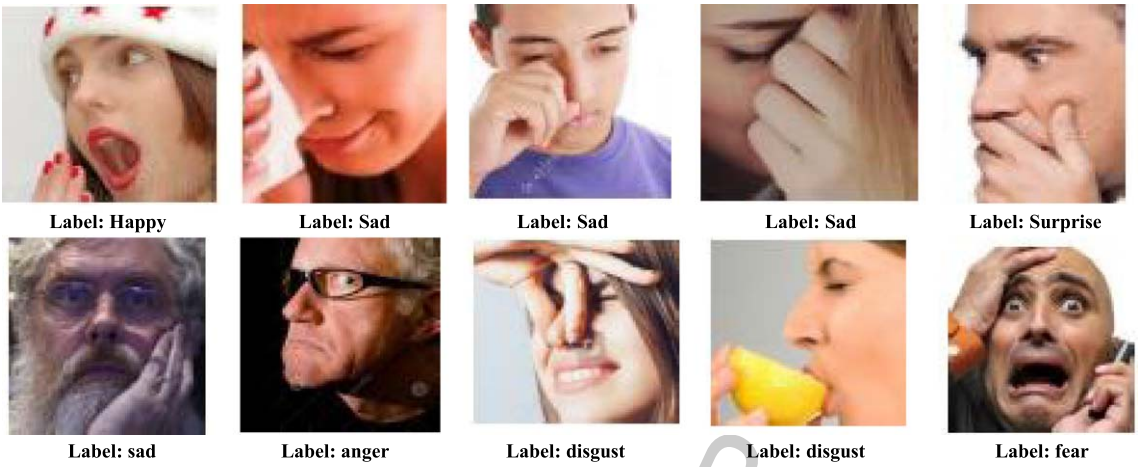


Fig. 2. Cognitive emotions with labels.

Table 1
Modality description

Modality	Features	RML	eINTERFACE	BAUM
Audio	Audio features	60	52	37
Audio net	66	79	42	
Visual	Video features	54	49	42
Visual net	69	55	52	

the learned features on the datasets. The comparison of the reported outcomes is considered due to the works that utilize similar settings of the experiments subject to independent test runs. The performance comparisons for cognitive emotion recognition are individually depicted in Tables 2 and 3 between the relevant handcrafted features and proposed leaned features. The proposed learned features of audio having the CNNs perform well than the handcrafted features of audio from Table 2 that are used broadly to recognize the cognitive emotion of audio like MFCC, features of prosody, acoustic Low-level Descriptors (LLD), Power Normalized Cepstral Coefficients (PNCC) and Relative Spectral Transform - Perceptual Linear Prediction (RASTA PLP). The learned audio features are shown in the proposed system using the fine-tuned model of AlexNet, which is more discriminant than the handcrafted features of audio to classify the cognitive emotion of audio. Also, the performance of the proposed learned audio features indicates the worth of using the three channels with $64 \times 64 \times 3$ with AlexNet as input from the Mel-spectrogram. It happened due to the robust learning ability of features of AlexNet, such as greater level convolutions being inferred semantically in a progressive way from the greater receptive fields. The representation of the RGB image is the same as the Mel-spectrogram, which is extracted. The robust low-level time-frequency features are extracted using the representation with the help of low-level 2-D convolutions. Thus more discriminant features are inferred using the greater levels of convolution. Three Mel-spectrogram channels represent cognitive emotions, such as structures and particular shapes, efficiently received using the trained AlexNet. The methodology is presented to transform the 1-D audio signals to relevant CNN input, which processes the images of 2D or 3D conventionally.

It is obtained using the proposed visual learned features having the 3D-CNNs obtain the good performance from Table 3 over the handcrafted features like LBP, Gabor wavelet, facial points, LPQ, and QIM as Quantized Image Matrix, which is compared. The merits of the proposed system, which is learned visual features, are demonstrated and are created using the fine-tuned model of C3DSports-1M that gives more discriminant power over the handcrafted features of visual to recognize the visual cognitive emotion. The mentioned experiments provide that the deep model is robust to learning the feature, and more discriminant features are produced manually over the designed model for the extracted features. Moreover, the deep model needs a huge volume of trained data. It inspires by the proposed system for transferring the

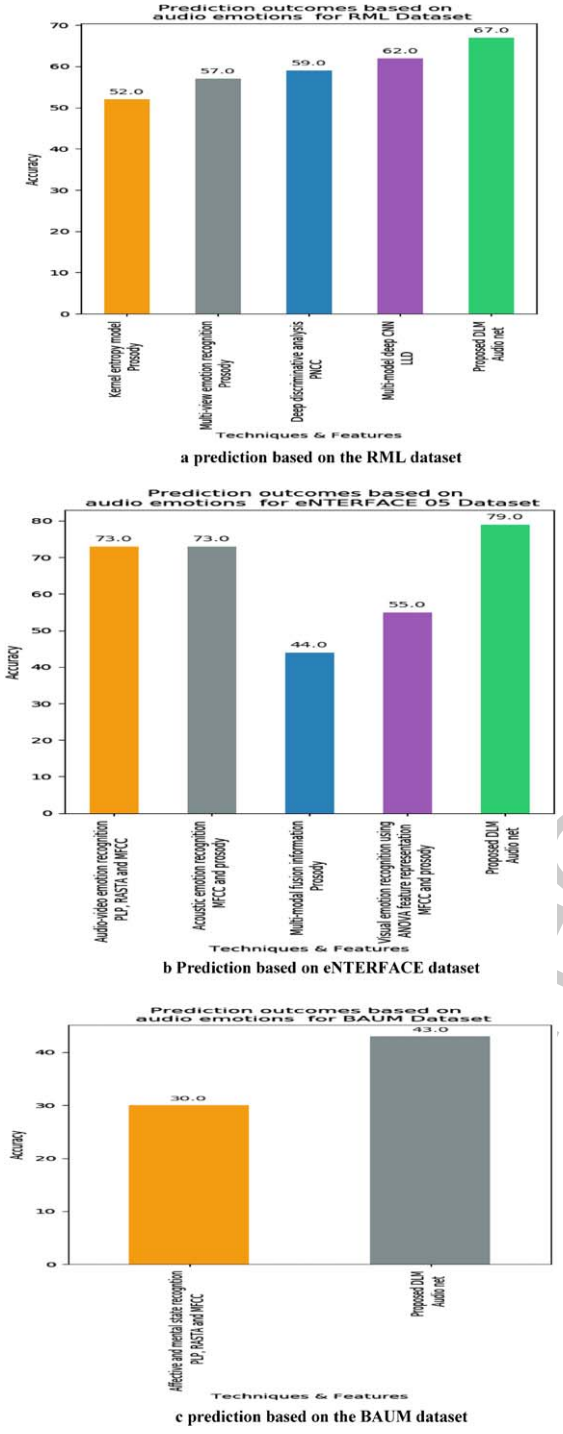


Fig. 3. a) Prediction based on the RML dataset. b) Prediction based on eNTERFACE dataset. c) prediction based on the BAUM dataset.

pre-trained technique for learning the cognitive and emotional features to other domains (See Fig. 4a to 4b).

The proposed system has the learned features presented in Tables 4 and 5, which are more discriminant in recognizing the cognitive emotion of the handcrafted features. Moreover, the proposed feature learning requires a huge training set, which can easily suffer from over-fitting the handcrafted features. However, the features are extracted having deep models, which needs costly calculation because of the huge parameters of the network. The performance of the two kinds of extracted Mel-spectrograms having various lengths is compared, such as $64\times15\times3$ and $64\times64\times3$, to explain the process of extracting the segments of Mel-spectrogram having the length of the frames as 64 over 15 frames that are used broadly in the speech recognition. Table 4 depicts the summarization of the experimental outcomes that the extracted Mel-spectrogram size of $64\times64\times3$ performs well compared with another (See Fig. 5). The length of the segment as 15 frames is not relevant to recognizing the cognitive emotion of audio. The frames of 15 may be very small for conveying adequate data to differentiate the cognitive emotions. The anticipated system is compared with four multimodal fused techniques: feature-level, decision-level, score-level fusion, and the recently presented methodology to check the proposed fusion method's efficiency. It is important to consider that the proposed system uses the deep DLM technique for implementing the fused network.

On the other hand, two layers of FCFC are used. The proposed system aims to predict the global video cognitive emotion sample. Hence, feature fusion approaches are needed for aggregating the extracted features on the visual and audio segments to the feature representation of the global video. Thus, the linear SVM is utilized for classifying cognitive emotion on the created features of the global video. The feature and decision-level fusion structures have the proposed models of two CNN. It is important to consider that the score-level fusion structure is completely the same as the decision-level fusion (See Fig. 6).

There are six ensemble rules like "Max", "Majority vote", "Min", "Sum", "Product", and "Average" that are tested for the decision-level fusion. The ensemble rule performance is investigated primarily on the works of decision-level fusion, and the best one is identified that is needed for creating the recorded performance. The six ensemble rules

Table 2
Prediction outcomes based on audio cognitive emotions

Datasets	Techniques	Features	Accuracy
RML	Kernel entropy model	Prosody	52
	Multi-view cognitive emotion recognition	57	
	Deep discriminative analysis	PNCC	59
	Multi-model deep CNN	LLD	62
	Proposed DLM	Audio net	**67**
eNTERFACE 05	Audio-video cognitive emotion recognition	PLP, RASTA and MFCC	73
	Acoustic cognitive emotion recognition	MFCC and prosody	73
	Multimodal fusion information	Prosody	44
	Visual cognitive emotion recognition using ANOVA feature representation	MFCC and prosody	55
	Proposed DLM	Audio net	79
BAUM	Affective and mental state recognition	PLP, RASTA and MFCC	30
	Proposed DLM	Audio net	**43**

Table 3
Prediction outcomes based on visual cognitive emotions

Datasets	Techniques	Features	Accuracy
	Kernel entropy model	Gabor wavelet	65
	Multi-view cognitive emotion recognition	LBP	57
	Deep discriminative analysis	Visual net	69
	Proposed DLM	LQP	**43**
eNTERFACE 05	Audio-video cognitive emotion recognition	Facial points	38
	Acoustic cognitive emotion recognition	QIM	40
	Proposed DLM	Visual net	55
BAUM	Affective and mental state recognition	LPQ	46
	Proposed DLM	Visual net	**51**

have the performance to compare with the proposed learned features on the works of decision-level fusion. The proposed learned features have six ensemble rules to compare the performance presented in Table 5. The better performance is obtained by the rule "Product", presented in Table 6. Hence, the decision-level fusion has the performance of the rule "Product," which is reported in the proposed system in the below experiments. The score-level fusion is implemented to refer to the techniques. More particularly, the equal-weighted summation is chosen concerning the attained score values of the class presented. The modalities of visual and audio on the recognition performance are presented in Table 6 with the assistance of various fusion techniques. The fusion approach of FC-FC is observed in Table 6 to perform better than the product's score-level, feature-level and decision-level fusion. The fusion network merits are implemented by having the two layers of FC-FC. The fusion network of FC-FC is implied for learning the joint feature representation of visual and audio from the two fine-tuned deep model outputs to identify the cognitive emotion via the back-propagation learning algorithms.

Table 6 shows the proposed fusion method of DLM performs well than other methods. The fusion of DLM is considered the deep fusion model when compared with the methods of score-level, feature-level and decision-level fusion (See Fig. 7). It shows the deep fusion efficiency to provide a better ability of

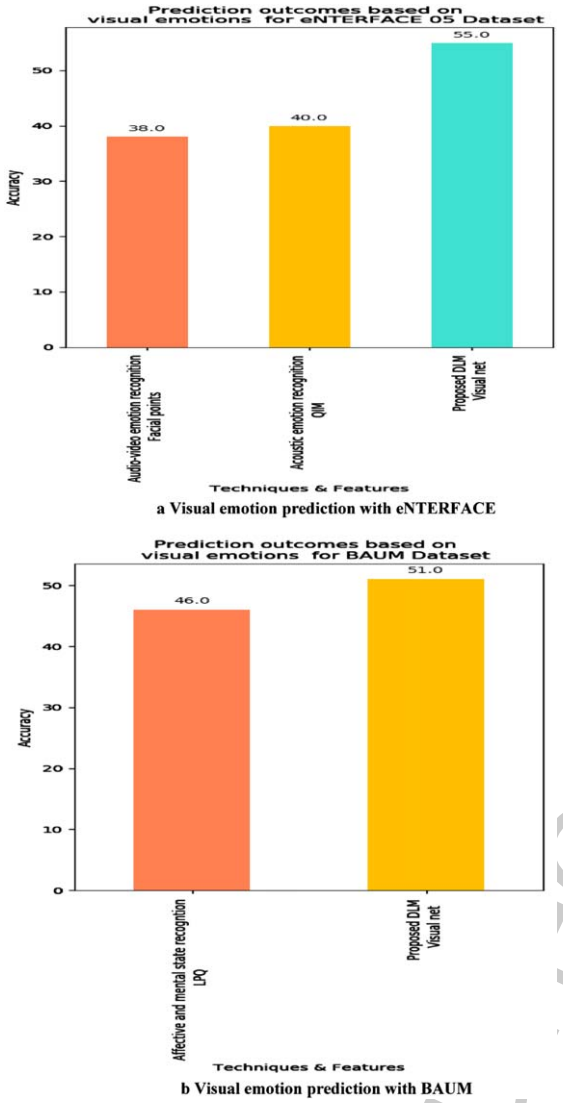


Fig. 4. a) Visual emotion prediction with eNTERFACE. b) Visual emotion prediction with BAUM.

feature learning to capture the greater non-linear relationships over modality. It is possible due to using the DLMs multilayer structure with many RBNs to form the various hidden layers stacked. A generative model is RBM to represent the probability distribution assigned with input data. Every RBM can learn the joint probability distribution in the DLM for video and audio input data. DLMs learn the non-linear dependencies efficiently over modalities using RBMs and training algorithms layer-wise, and the outcomes are better for the audio and visual fusion features. The consistent findings are there in the prior research. It is needed for visualizing the learned weights of the proposed model of DLM. Moreover, the proposed DLMs input is visual and audio features over the semantic image. The learned weights made the DLMs hard to interpret. Rather than the method of FCFC fusion, the DLM performed well, such as 84% vs. 86% on the eNTERFACE05 dataset, 53% vs. 55% on the dataset of BAUM-1 s, and 79% vs. 81% on the RML dataset accordingly. Because of the pre-trained unsupervised model in DLMs, the merits are presented as the local optimum weights to initialize the network. On the other hand, the method of FCFC fusion has the initial weights created randomly.

Table 4
Segments-based recognition

Spectrogram	BAUM	eNTERFACE	RML
64*15*3	34	53	51
64*64*3	43	79	79

Table 5
Multi-modal-based recognition

Decision making	BAUM	eNTERFACE	RML
Voting	46	70	64
Maximal	49	81	73
Summation	51	81.5	73
Minimal	51	79	73
Average	51	81	73
Product	52	82	75

When the DLM is performed well on the dataset of BAUM-1 s, then the multimodal cognitive emotion recognition results in a confusion matrix. The multimodality performance is presented in Table 8 in terms of % to measure for every cognitive emotion during the DLM is provided with the average accuracy of 80.36% on the dataset of RML, Cognitive emotion Precision Recall F-score Anger 85.33 88.70 86.98 Disgust 89.71 95.50 92.51 Fear 80.71 71.33 75.73 Joy 65.53 68.93 67.19 Sadness 91.03 83.33 87.01 Surprise 83.21 87.07 85.10 or results to recognize the LOSGO. It is attractive to identify the dataset of RML as “fear” and “joy” is more challenging to find than other cognitive emotions. It is possible due to the cues of audio and visual as fear” and “joy” is not enough in a distinct way. The cognitive emotions of “surprise”, “fear”, and “sadness” are identified as having less accuracy relatively on the dataset of eNTERFACE05 that is over 80%.

On the other hand, identifying other cognitive emotions is done with 90% accuracy. The average accuracy in classification on the dataset of BAUM-1 s is lesser than the other two datasets. Spontaneous cognitive emotions are shown, which are more chal-

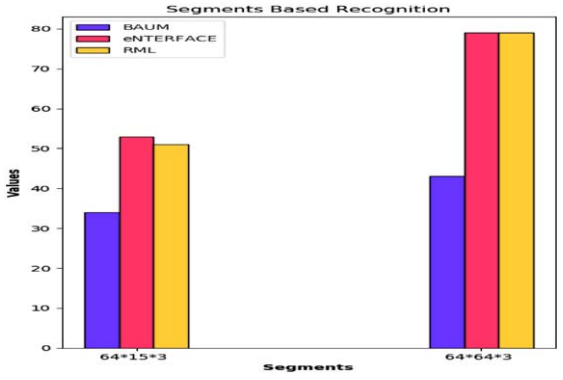


Fig. 5. Segments-based prediction.

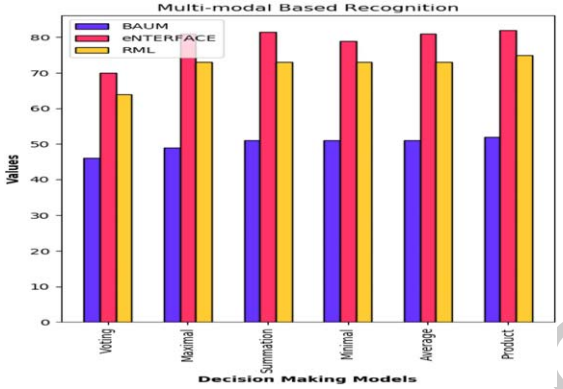


Fig. 6. Decision-making model.

Table 6
Subject independence-based recognition

Method	BAUM	eNTERFACE	RML
Feature	52	82	75
Product	52	82	75
Score	52	81	74
FCFC	52	83	79
DBN	55	85	80
DLM	56	86	81

lenging to recognize than the acted cognitive emotion. The precision, F-score, and recall are computed in the proposed system for measuring the cognitive emotion of multimodality for performance recognition on the three datasets added to the classification of the confusion matrix. Tables 7 to 9 present the experimental outcomes accordingly. These results indicate that the three datasets provide various challenges in predicting particular cognitive emotions. Consider an instance that the identification is easy for the dataset of RML as "disgust" over the other two datasets. The identification of "joy" is easier on the BAUM-1 s and eNTERFACE05 over the dataset of RML.

Table 7
Performance evaluation with multi-modal dataset

Cognitive emotion	Precision	Recall	F1-score
Performance evaluation - RML			
Anger	86	89	87
Disgust	90	96	93
Fear	81	72	76
Joy	66	69	68
Sadness	92	84	88
Surprise	84	89	86
Performance evaluation – eNTERFACE			
Anger	90	91	91
Disgust	92	90	91
Fear	76	80	78
Joy	93	93	93
Sadness	87	84	86
Surprise	80	80	80
Performance evaluation – BAUM			
Anger	28	28	27
Disgust	83	66	73
Fear	53	54	53
Joy	21	26	23
Sadness	26	26	26
Surprise	42	65	51

Table 8
Performance evaluation – BAUM

Cognitive emotion	RML	eNFERENCE	BAUM
FC (layer 1)	79	84	53
FCFC (layer 2)	78	83	52
FCFC (layer 3)	76	82	51
DBN (layer 1)	79	85	52
DBN (layer 2)	81	86	55
DBN (layer 3)	81	86	54
DLM	82	87	56

Table 9
Multi-modal-based cognitive emotion recognition

Datasets	Techniques	Accuracy
RML	Kernel entropy model	73
	Multi-view cognitive emotion recognition	76
	Deep discriminative analysis	75
	Proposed DLM	**81**
eNTERFACE 05	Audio-video cognitive emotion recognition	71
	Acoustic cognitive emotion recognition	72
	Proposed DLM	**86**
BAUM	Affective and mental state recognition	52
	Proposed DLM	**55**

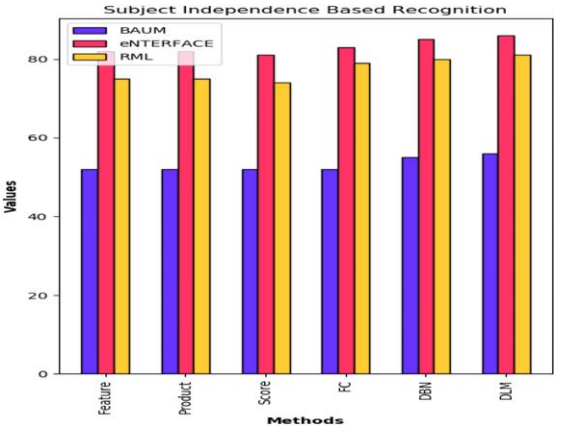


Fig. 7. Subject independence-based recognition.

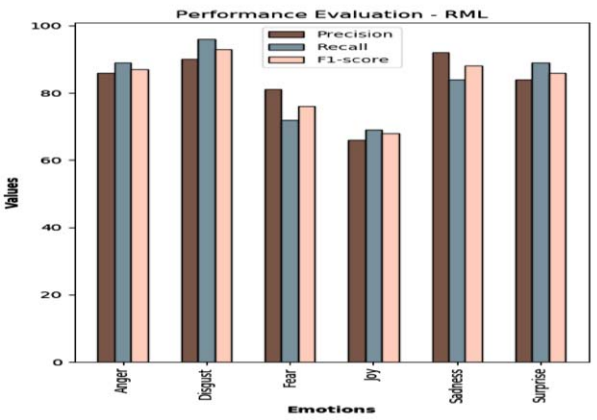
Table 10
Comparison of various performance metrics

Methods	Accuracy (%)	Error rate	MCC
Hybrid CNN and NADE	95%	0.089	0.4656
CNN	96.13%	1.02	0.4534
CNN and KELM	93.6%	1.56	0.4317
Deep CNN with data augmentation	94.58%	1.63	0.4658
CNN and Genetic Algorithm	94.2%	0.077	0.4205
Hybrid VGG16-NADE	97.31%	0.075	0.3564
ResNet+SE	98.8%	0.065	0.2645
Proposed	99%	0.055	0.1253

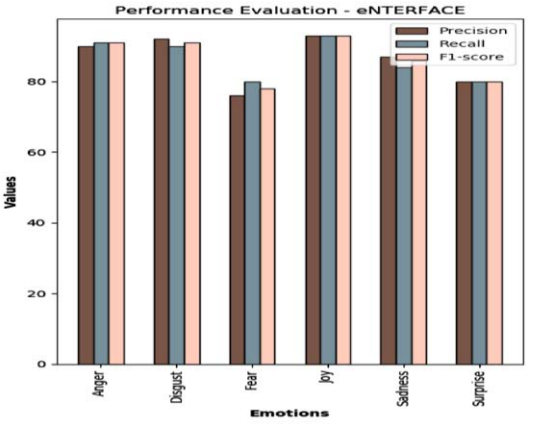
The DLMs structure affects the fusion performance for visual and audio modalities. The performance is presented for the three fusion networks of DLM to determine the efficiency of various deep structures. They are (i) DLM-1 (8192-4096-6), (ii) DLM2 (8192-4096-2048-6), and (iii) DLM-3 (8192-4096-2048-1024-6). In the same way, the three fusion networks of FCFC have a performance that is relevant to DLMs in the proposed system. They are (i) FC-1 (8192-4096-6), (ii) FC2 (8192-4096-2048-6), and (iii) FC-3 (8192-4096-2048-1024-6). The dropout layer is included before the last layer of softmax that is relevant to the classification of cognitive emotion for the fusion networks completed. The overfitting is reduced by setting the parameter of dropout is 0.3. The comparison of various structures shows the performance in the fusion network. The better performance is obtained by FC-1 among the three fusion networks of FCFC from Fig. 8a to 8c in the proposed system. The FC-1 is extremely efficient than FC-2 and 3 for combining visual and audio. It is possible due to the more layers in the network of FCFC and the huge increase in the parameters of the network, which creates the network of FCFC is prone to the problem of overfitting. The DLM-2 performs better than DLM3 to obtain the best performance over DLM-1 for the fusion network of DLM. The deeper models of DLM, such as DLM-1 and DLM-3, can feature fusion over 1-layer DLM-1 because of utilizing the many RBIs and the training algorithm for the efficient layer-wise. DLM-3 degrades the DLM-2 performance due to the depth of DLM-3 over DLM-2, which has more parameters of the network that is more challenging for optimizing the training dataset, which is small scale.

Table 11 depicts that the anticipated method is evaluated with existing research on the three datasets. It is important to consider that these researches are performed based on the experiments of the subject, which are independent and consistent with the setting of experiments. Table 11 provides the outcomes which present the proposed model as competition to the modern outcomes. More particularly, the proposed methods performed better than existing studies on the datasets of eNTERFACE05 and acted RML using over 5%. The performance is improved in the proposed system from 51.29% to 54.57% on the spontaneous dataset. The comparison to the use of the handcrafted features and the fusion methodologies shallows for the integration of the visual and audio modalities. Then, the merits of the proposed learned features and techniques for fusion are provided (See Fig. 9a to 9c). Also, the proposed system enhances the existing study on the dataset of RML from 74.32% to 80.36%. There are two enhancements to obtain (See Figs. 10 and 11). They are (i) the proposed model has the models of 3D-CNN for extracting the cues of spatial and temporal from video to compared with the models of CNN, and (ii) The fusion method of DLM provides better fusion ability of multimodal feature over the fusion method of FCFC is presented in the proposed systems' experiments.

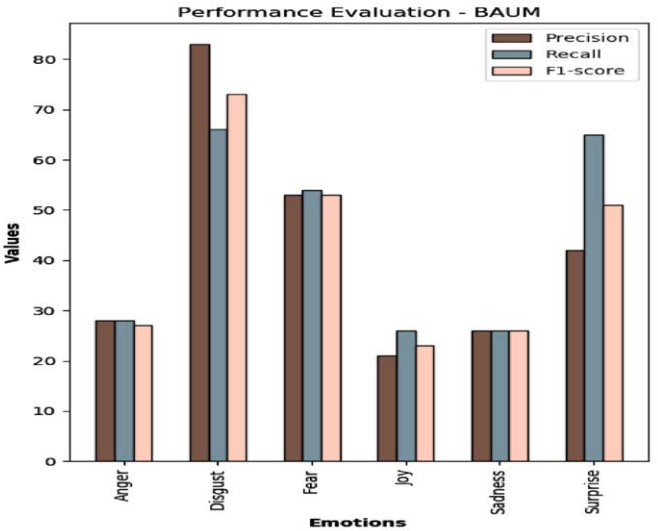
Table 10 and Fig. 12 compare metrics like prediction accuracy, error rate and MCC. The prediction accuracy of the layered network model is 99% which is 4%, 2.8%,5%, 4%, 4%, 4%, 1.6% and 0.2% higher than hybrid CNN and NADE, CNN, CNN and KELM, deep CNN with data augmentation, CNN-GA, hybrid VGG16-NADE and ResNet+SE. The error rate of the proposed model is 0.055, which is lesser than other approaches. Generally, the MCC value should range from -1 to 1; however, the proposed model gives a better MCC value of 0.1253 while the others are 0.46, 0.45, 0.43, 0.46, 0.42, 0.35 and 0.26, respectively.



a Performance evaluation RML



b Performance evaluation eNFERFACE



c Performance evaluation BAUM

Fig. 8. a) Performance evaluation RML. b) Performance evaluation eNFERFACE. c) Performance evaluation BAUM.

6. Conclusion

The multimodal system is proposed to operate on the raw signal, which carries out the point-to-point spontaneous prediction work of cognitive emotion from the data of visual and speech. The LSTM (recurrent network) is concerned with contextual data. The visual and speech networks are pre-trained in the proposed system to fasten the model's training individually. Also, the recurrent layers have gate activations in the modality of speech, and the cells are found which are greatly correlated, having the features prosodic, which are assumed to cause arousal in the proposed system. The proposed system obtains a better performance considerably on the test set by

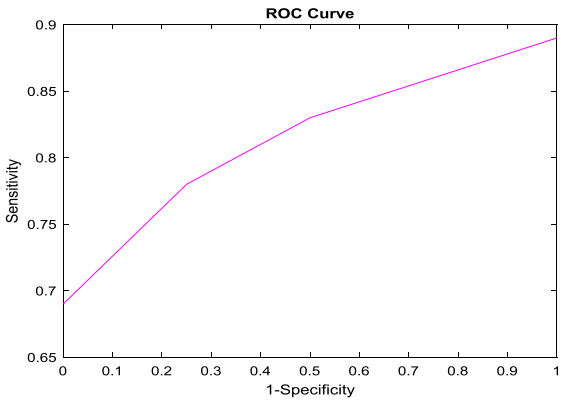
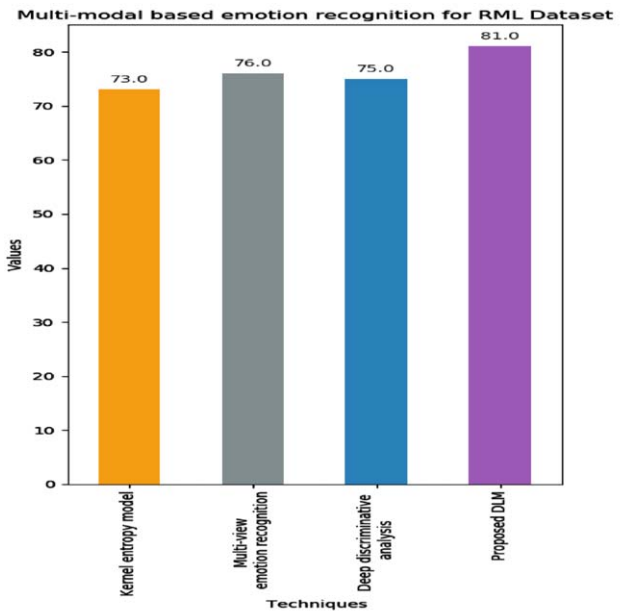
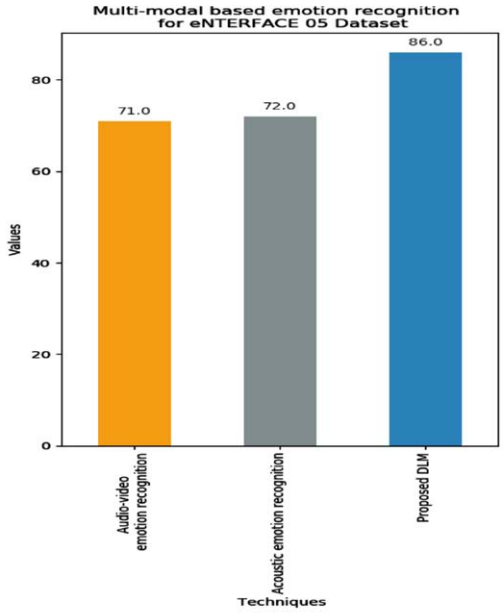


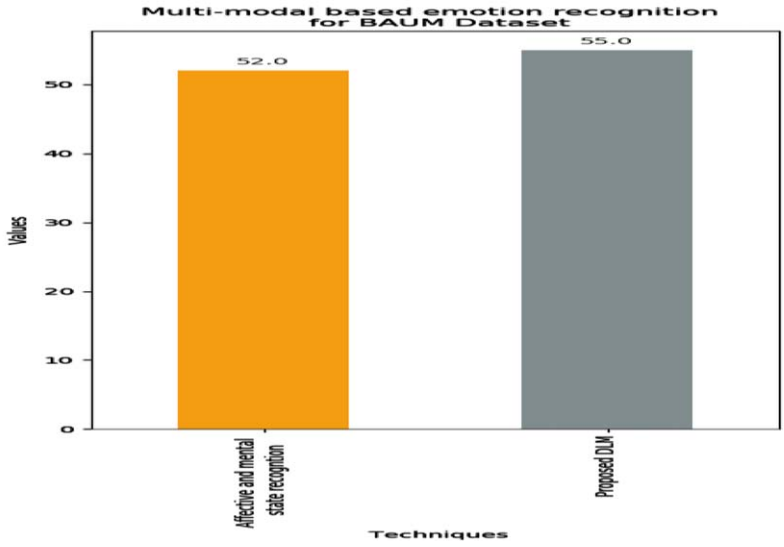
Fig. 11. ROC curve.



a Multi-modal analysis with RML dataset



b Multi-modal analysis with eNFERENCE dataset



c Multi-modal analysis with BAUM dataset

Fig. 9. a) Multi-modal analysis with RML dataset. b) Multi-modal analysis with eNFERENCE dataset. c) Multi-modal analysis with BAUM dataset.

using the experiments carried out on the uni-modal modality when compared with the other models with the help of the available dataset, which has included that is given to the challenge of establishing the learning features efficiency, which suits better for hand work. The proposed model give 99%, 0.055 error rate and 0.12 MCC which is substantially higher to other approaches. Also, the proposed multimodal model performs superiorly in both arousal and valence dimensions than other models. The major research constraint is the computational complexity where the proposed model consumes time during execution as it deals with three different multi-modal datasets. The future study relies on analyzing behavior in the wild nature of human expression. Incorporating more modalities is done in the proposed system as physio

Fig. 10. Prediction outcomes.

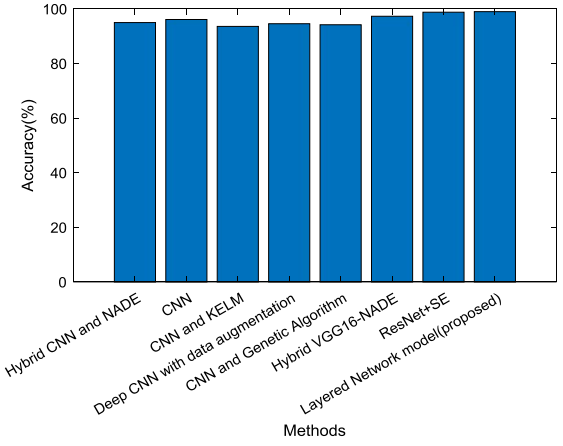


Fig. 12. Accuracy comparison.

Notations

AVEC	Audio/Visual Cognitive Emotion Challenge and Workshop
BAUM	Bahcesehir University Multi-modal face database
CNN	Convolutional Neural Networks
E2E	End-to-End network
DLM	Dense Layer model
DNN	Deep Neural Networks
FCL	Fully Connected Layer
HMM	Hidden Markov Model
MDAE	Multimodal Deep Autoencoder
MFCC	Mel-Frequency Cepstral Coefficients
MSDF	Multi-scale Dense SIFT features
LLD	Low-level Descriptors
LSTM	Long Short Term Memory
LIBSVM	Library of Support Vector Machine
MCC	Mathews Correlation coefficient
NADE	Normalized Auto-encoder and decoder
PNCC	Power Normalized Cepstral Coefficients
RASTA PLP	Relative Spectral Transform - Perceptual Linear Prediction
ResNEt	Residual Network
RBM	Restricted Boltzmann Machine
RGB	Red Green Blue
RML	Ryerson Emotion dataset
RNN	Recurrent Neural Networks
RVM	Relevance Vector Machine

as the goal in the future to enhance the performance to recognize the cognitive emotion works. Also, the future system is intended to experiment with huge cognitive emotion databases of discrete labels. It is very attractive for experimenting with works over recognizing cognitive emotion.

References

[1] Atmaja and M. Akagi, Speech cognitive emotion recognition based on speech segment using LSTM with attention model, in Proc. IEEE Int. Conf. Signals Syst. (ICSigSys), Jul. 2019, pp. 40–44.

[2] Schuller, Speech cognitive emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends, *Commun. ACM* **61**(5) (2018), 90–99.

[3] Song, Transfer linear subspace learning for cross-corpus speech cognitive emotion recognition, *IEEE Trans. Affect. Comput.* **10**(2) (2019), 265–275.

[4] Pan P. Luo, J. Shi and X. Tang, Two at once: Enhancing learning and generalization capacities via IBN-Net, in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 464–479.

[5] Wei and Y. Zhao, A novel speech cognitive emotion recognition algorithm based on wavelet kernel sparse classifier in the stacked deep auto-encoder model, *Pers. Ubiquitous Comput.* **23**(3–4) (2019), 521–529.

[6] Takaki H. Kameoka and J. Yamagishi, Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis, in Proc. Interspeech, Aug. 2017, pp. 1128–1132.

[7] Park W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk and Q.V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, 2019, arXiv:1904.08779. [Online]. Available: http://arxiv.org/abs/1904.08779

[8] Neumann and N.T. Vu, Improving speech cognitive emotion recognition with unsupervised representation learning on unlabeled speech, in Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 7390–7394.

[9] Zeng, L. Dong, G. Chen and Q. Dong, Multi-feature fusion speech cognitive emotion recognition based on SVM, Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC), Jul. 2020, pp. 77–80.

[10] Assuncao and P. Menezes, Intermediary fuzzification in speech cognitive emotion recognition, in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), Jul. 2020, pp. 1–6.

[11] W. Liu, W.L. Zheng and B.L. Lu, Cognitive emotion recognition using multimodal deep learning, in Proc. Int. Conf. Neural Inf. Process., Kyoto, Japan, 2016, pp. 521–529.

[12] Sariyanidi, H. Gunes and A. Cavallaro, Learning bases of activity for facial expression recognition, *IEEE Trans. Image Process* **26**(4) (2017), 1965–1978.
[13] Liu, S. Shan, R. Wang and X. Chen, Learning expressionless on the spatiotemporal manifold for dynamic facial expression recognition, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, 2014, pp. 1749–1756.
[14] S. Dinesh, K. Maheshwari, B. Arthi, P. Sherubha, A. Vijay et al., Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms, *of Healthcare Engineering* **2** (2022), 1–9.
[15] Keren and B. Schuller, Convolutional RNN: An enhanced model for extracting features from sequential data, in Proc. IEEE Int. Joint Conf. Neural Netw., Vancouver, BC, Canada, 2016, pp. 3412–3419.
[16] May, R.M. Pai and M.M. Pai, Automatic facial expression recognition using CNN, *Procedia Computer Science* **93** (2016), 453–461.
[17] Kim, H. Lee, J. Roh and S.-Y. Lee, Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition, in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015.
[18] Dhall, O. Ramana Murthy, R. Goecke, J. Joshi and T. Gedeon, Video and image-based emotion recognition challenges in the wild: Emotiw 2015, in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015.
[19] Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR, vol. abs/1409.1556, 2014.
[20] Ciregan, U. Meier and J. Schmidhuber, Multi-column deep neural networks for image classification, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
[21] Krishna, V.K. Deepak, K. Manikantan and S. Ramachandran, Face recognition using transform domain feature extraction and PSO-based feature selection, *Appl. Soft Comput.* **22** (2014), 141–161.
[22] Liu, S. Li, S. Shan and X. Chen, AU-inspired deep networks for facial expression feature learning, *Neurocomputing* **159** (2015), 126–136.
[23] Zavaschi, A.S. Britto, Jr., L.E.S. Oliveira and A.L. Koerich, Fusion of feature sets and classifiers for facial expression recognition, *Expert Syst. Appl*. **40**(2) (2013), 646–655.
[24] Diao, F. Chao, T. Peng, N. Snooker and Q. Shen, Feature selection inspired classifier ensemble reduction, *IEEE Trans. Cybern.* **44**(8) (2014), 1259–1268.
[25] Zeng et al., One-class classification for spontaneous facial expression analysis, in Proc. 7th Int. Conf. Autom. Face Gesture Recognition., Southampton, UKUK, 2006, pp. 281–286.
[26] Zeng et al., Audio-visual spontaneous emotion recognition, in Artificial Intelligence for Human Computing (LNCS 4451). Heidelberg, Germany: Springer, 2007.
[27] Cootes, C.J. Taylor, D.H. Cooper and J. Graham, Active shape models-their training and application, *Comput. Vis. Image Understand.* **61**(1) (1995), 38–59.
[28] Edwards and C.J. Taylor, Active appearance models, in Computer Vision—ECCV98. Heidelberg, Germany: Springer, 1998, pp. 484–498.
[29] Ojala, M. Pietikäinen and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7) (2002), 971–987.
[30] Liu et al., Non-manual grammatical marker recognition based on the multiscale, spatiotemporal analysis of head pose and facial expressions, *Image Vis. Comput.* **32**(10) (2014), 671–681.
[31] Zhong, Q. Liu, P. Yang, J. Huang and D.N. Metaxas, Learning multiscale active facial patches for expression analysis, *IEEE Trans. Cybern.* **45**(8) (2015), 1499–1510.
[32] Yu et al., Is interactional dyssynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues, *IEEE Trans. Cybern.* **45**(3) (2015), 492–506.
[33] Senechal et al., Facial action recognition combining heterogeneous features via multi-kernel learning, *IEEE Trans. Syst., Man, Cybern. B, Cybern.* **42**(4) (2012), 993–1005.
[34] Siddiqi, R. Ali, A.M. Khan, Y.-T. Park and S. Lee, Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields, *IEEE Trans. Image Process* **24**(4) (2015), 1386–1398.
[35] Pang, H. Yan, Y. Yuan and K. Wang, Robust CoHOG feature extraction in human-centered image/video management system, *IEEE Trans.Syst., Man, Cybern. B, Cybern* **42**(2) (2012), 458–468.
[36] Liu and Z. Wang, Facial expression recognition based on the fusion of multiple Gabor features, in Proc. 18th Int. Conf. PatternRecognit. (ICPR), vol. 3. Hong Kong, 2006.
[37] Larochelle, Y. Bengio, J. Louradour and P. Lamblin, Exploring strategies for training deep neural networks, *J. Mach. Learn. Res.* **10** (2009), 1–40.
[38] Liu, S. Han, Z. Meng and Y. Tong, Facial expression recognition via a boosted deep belief network, in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., Columbus, OH, USA, 2014, pp. 1805–1812.
[39] Zhi, M. Flierl, Q. Ruan and B.W. Kleijn, Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition, *IEEE Trans. Syst., Man, Cybern. B, Cybern.* **41**(1) (2011), 38–52.
[40] Huang and H. Yin, Visom for dimensionality reduction in face recognition, in Advances in Self-Organizing Maps. Heidelberg, Germany: Springer, 2009.
[41] Yu, Z. Wang, M. Hagenbuchner and D.D. Feng, Spectral embedding based facial expression recognition with multiple features, *Neurocomputing* **129** (2014), 136–145.
[42] Zavaschi, A.S. Britto, Jr., L.E.S. Oliveira and A.L. Koerich, Fusion of feature sets and classifiers for facial expression recognition, *Expert Syst. Appl.* **40**(2) (2013), 646–655.
[43] Boucenna, P. Gaussier, P. Andry and L. Hafemeister, A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game, *Int. J. Soc. Robot.* **6**(4) (2014), 633–652.
[44] Seshadri and M. Savvides, Towards a unified framework for pose, expression and occlusion tolerant automatic facial alignment, *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(10) (2016), 2110–2122.
[45] Mollahosseini, D. Chan and M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in Proc. IEEEWinter Conf. Appl. Comput. Vis. (WACV), Lake Placid, NY, USA, 2016.